Revealing Hidden Patterns in Political News and Social Media with Machine Learning

Thesis by
Konstantina Lazaridou

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

HASSO PLATTNER INSTITUTE, DIGITAL ENGINEERING FACULTY,
UNIVERSITY OF POTSDAM
Potsdam, Germany

2020
Submitted [August 20th 2020]

Principal supervisor

Prof. Dr. Felix Naumann
Information Systems, Hasso Plattner Institute
Digital Engineering Faculty, University of Potsdam, Germany

Doctoral thesis reviewers

Prof. Dr. Felix Naumann
Information Systems, Hasso Plattner Institute
Digital Engineering Faculty, University of Potsdam, Germany

Prof. Dr. Alexander Löser
Database Systems and Text-based Information Systems
Data Science Research Center, Beuth University of Applied Sciences in Berlin, Germany

Prof. Dr. Robert Jäschke
Information Processing and Analytics
Institute for Library and Information Science, Humboldt University in Berlin, Germany

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-50273
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-502734

Konstantina Lazaridou
ORCID: https://orcid.org/0000-0002-3713-7612

Dedicated to Dr. Novieku.

I would like to acknowledge my supervisor Prof. Felix Naumann for his support throughout all the steps of my doctoral studies and thank him for my being able to always count on him for thorough feedback on my papers. I thank Prof. Alexander Loeser for supporting me during my last and very challenging year of my PhD, and for introducing me to the state-of-the-art research I wanted to study and acquaint myself with. I would like to mention all my colleagues from Potsdam and Berlin and thank them for their insights and for constant support. Factmata and my former colleagues during my internship have also been essential to my journey, giving me the opportunity to work on my research topic in an industrial, fun and creative environment. Thanks to our collaboration, I had the opportunity to bring research on media bias detection a small step further and publish my last and favorite research paper in the context of my thesis. A special thank you goes to my friends, who as always, stood by me as a family and never stopped believing in me and my abilities. I would not have made it without them.

ABSTRACT

As part of our everyday life we consume breaking news and interpret it based on our own viewpoints and beliefs. We have easy access to online social networking platforms and news media websites, where we inform ourselves about current affairs and often post about our own views, such as in news comments or social media posts. The media ecosystem enables opinions and facts to travel from news sources to news readers, from news article commenters to other readers, from social network users to their followers, etc. The views of the world many of us have depend on the information we receive via online news and social media. Hence, it is essential to maintain accurate, reliable and objective online content to ensure democracy and verity on the Web. To this end, we contribute to a trustworthy media ecosystem by analyzing news and social media in the context of politics to ensure that media serves the public interest. In this thesis, we use text mining, natural language processing and machine learning techniques to reveal underlying patterns in political news articles and political discourse in social networks.

Mainstream news sources typically cover a great amount of the same news stories every day, but they often place them in a different context or report them from different perspectives. In this thesis, we are interested in how distinct and predictable newspaper journalists are, in the way they report the news, as a means to understand and identify their different political beliefs. To this end, we propose two models that classify text from news articles to their respective original news source, i.e., reported speech and also news comments. Our goal is to capture systematic quoting and commenting patterns by journalists and news commenters respectively, which can lead us to the newspaper where the quotes and comments are originally published. Predicting news sources can help us understand the potential subjective nature behind news storytelling and the magnitude of this phenomenon. Revealing this hidden knowledge can restore our trust in media by advancing transparency and diversity in the news.

Media bias can be expressed in various subtle ways in the text and it is often challenging to identify these bias manifestations correctly, even for humans. However, media experts, e.g., journalists, are a powerful resource that can help us overcome the vague definition of political media bias and they can also assist automatic learners to find the hidden bias in the text. Due to the enormous technological advances in artificial intelligence, we hypothesize that identifying political bias in the news could be achieved through the combination of sophisticated deep learning models

and domain expertise. Therefore, our second contribution is a high-quality and reliable news dataset annotated by journalists for political bias and a state-of-the-art solution for this task based on curriculum learning. Our aim is to discover whether domain expertise is necessary for this task and to provide an automatic solution for this traditionally manually-solved problem.

User generated content is fundamentally different from news articles, e.g., messages are shorter, they are often personal and opinionated, they refer to specific topics and persons, etc. Regarding political and socio-economic news, individuals in online communities make use of social networks to keep their peers up-to-date and to share their own views on ongoing affairs. We believe that social media is also an as powerful instrument for information flow as the news sources are, and we use its unique characteristic of rapid news coverage for two applications. We analyze Twitter messages and debate transcripts during live political presidential debates to automatically predict the topics that Twitter users discuss. Our goal is to discover the favoured topics in online communities on the dates of political events as a way to understand the political subjects of public interest. With the up-to-dateness of microblogs, an additional opportunity emerges, namely to use social media posts and leverage the real-time verity about discussed individuals to find their locations. That is, given a person of interest that is mentioned in online discussions, we use the wisdom of the crowd to automatically track her physical locations over time. We evaluate our approach in the context of politics, i.e., we predict the locations of US politicians as a proof of concept for important use cases, such as to track people that are national risks, e.g., warlords and wanted criminals.

PUBLICATIONS

[1] Christian Godde, Konstantina Lazaridou, and Ralf Krestel. Classification of german newspaper comments. In *Proceedings of the Lernen, Wissen, Daten, Analysen Conference*, volume 1670. Online Proceedings for Scientific Conferences and Workshops, 2016.

[2] Toni Gruetze, Ralf Krestel, Konstantina Lazaridou, and Felix Naumann. What was Hillary Clinton doing in Katy, Texas? In *Proceedings of the International Conference on World Wide Web Companion*. Association for Computing Machinery, 2017.

[3] Konstantina Lazaridou and Ralf Krestel. Identifying political bias in news articles. *Special Issue of the Bulletin of the IEEE Technical Committee on Digital Libraries*, 12, 2015.

[4] Konstantina Lazaridou, Ralf Krestel, and Felix Naumann. Identifying media bias by analyzing reported speech. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE Computer Society, 2017.

[5] Konstantina Lazaridou, Toni Gruetze, and Felix Naumann. Where in the world is Carmen Sandiego?: Detecting person locations via social media discussions. In *Proceedings of the International Conference on Web Science*. Association for Computing Machinery, 2018.

[6] Konstantina Lazaridou, Alexander Loeser, Maria Mestre, and Felix Naumann. Discovering biased news articles leveraging multiple human annotations. In *International Language Resources and Evaluation Conference*. European Language Resources Association, 2020.

[7] Konstantina Lazaridou, Alexander Loeser, Felix Naumann, and Ralf Krestel. Reported speech in political news articles: A media bias perspective. *Natural Language Engineering (in revision)*, 2020.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

The *World Wide Wide* (Web) has revolutionized our everyday lives from the way we work and interact to the way we consume products and services. Founded by British Scientist Tim Berners-Lee in the late 1980s [14], the Web is an information system that has made immediate access to magnitudes of information possible for everyone. The rise both of the Internet and the Web have changed and improved health care, education, business, communication, transport and entertainment in numerous ways, becoming the leading factor in social evolution. There has been several milestones that facilitated the Web's entrance in our lives, from the development of web browsers, such as the Mosaic browser in 1993, Microsoft's Internet Explorer in 1996, to the emergence of search engines, e.g., Google in 1998. In 2004 Gmail is released as a modern email communication service and the first smart phones are also launched near the late 2000s. The World Wide Web has now evolved through the last four decades to Web 4.0 [28], incorporating multimedia information, social networking platforms, mobile, semantic and also cloud technologies.

The common core in all web-related technological advances is the generation and the exchange of information. Web pages, services and applications create various kinds of information depending on the respective users and domains. For instance, search engines maintain indices of websites and relevance scores to user queries. Social networks contain personal user information, e.g., person locations. Businesses often own employee documents and machine data, and financial institutions maintain transaction logs. In the health care industry, commonly used information sources are documents with medical records and potential drug interactions.

All above-mentioned information pieces are being manipulated and stored using various data processing systems and database technologies. They are also analyzed in order to discover insights and hidden knowledge that can improve the state of the respective field. Such data pipelines are meant to serve both the provider of the information and also the end users that rely on the products. For instance, online newspapers need reliable infrastructure to provide breaking news without delay, and the readers need both access to user-friendly news websites and accurate information. Banks also need to have robust and scalable systems to serve millions of users at the same time, and on the other side, the public trusts these institutions

with their data privacy and security. Hence, it is essential to consider the amount and also the type of data that is being created in every domain in order to understand how it should be handled.

1.1 Importance of Data

Data is everywhere. Professional or user-generated content lies in every aspect of today's products. Data can help solve problems in organizations, e.g., if a hospital observes high medication errors, then there might be high staff turnover, or if an online shop detects its popular products, it can increase its storage capacity to deliver them without delay. Data can serve as evidence and can help leaders to be more strategic, confident and efficient in their decisions. For instance, educational institutes can analyze their student grades to find which courses are more demanding and improve the respective material, or find out which students are failing and offer them support. Real-time *Internet of Things* (IoT) data that are produced by various devices without requiring human-to-human or human-to-computer interaction, has the potential to improve the citizen's commute by navigating traffic. Data is also extremely important for advancing automation with *artificial intelligence*, i.e., machines trained to think and make decisions as humans would. Such programs are fundamentally dependent on data and specifically its amount, in order to learn common sense and follow it to solve problems. This is similar to our human brain that learns progressively, e.g., when children start to learn by example and eventually can imitate others and solve tasks by themselves.

Overall, data is important both for humans and machines, and it can improve people's lives by helping them make informed decisions. There is certainly not a lack of available data. However, its quality and validity is not always given. Data can be often biased, irrelevant, imbalanced or incomplete. Automatic pre-processing and cleansing is an integral initial part of data pipelines, as well as adapting data analysis algorithms to compensate for existing noise [183]. Manual data examination before any analysis to eliminate errors in the data as well as careful result interpretation, are additional ways to ensure the correct conclusions are drawn. Another timely issue when dealing with data is the privacy and security that it should but is not always accompanied by [61]. For instance, if we were to make our datasets and algorithms for news comment analysis [50] or person location detection in social media [85] publicly available, we would need to carefully anonymize personal information. Lastly, there are also ethical concerns in the way data and algorithms are used in science and industry. The moral behavior both of the humans and the data-driven

software that we construct is essential in order to make sure that technology is used for the greater-good. Potential risks include artificial intelligence software replacing jobs that should not be replaced, lack of transparency and accountability by the developers or even weaponizing machines for military attacks. Hence, with data comes great responsibility. Securing that data is used as appropriate is key to maintain the society's trust to the Web and its technologies.

1.2 Data Mining

Data can contain powerful insights for problem solving, which are often hidden within its distributions and structures. The field of *data mining* refers to algorithms built-in computer programs that are meant to discover knowledge in data. Data mining uses database technologies, machine learning and statistical models to uncover hidden patterns in (typically large) datasets [90]. It is also recently referred to as the unified field of *data science* [29]. Methods and use cases include *anomaly detection*, e.g., detecting abnormal or fraud activity in a bank institutions, and designing *association rules*, e.g., discovering product correlations in online shops for marketing purposes. Statistical methods, such as *regression*, are also used to determine relationships between variables, e.g., the relation between symptoms and health conditions. *Clustering* and *classification* are also two main tasks of data mining, e.g., grouping people with similar interests in social media for advertisement purposes, and classifying hateful or fake online content for protecting the Web users.

Information retrieval (IR) is a related task to data mining and it refers to algorithms that make it feasible to search in stored content, e.g., Google keyword search is a typical example of a modern search engine. The difference between information retrieval and data mining is that in information retrieval systems there is always a specific user query that should be answered by the computer, but in data mining there is not always a predefined goal (e.g., in learning how to meaningfully represent images in a semantic way for recognizing bar coded tags). The quality of a search result can be evaluated by computing the achieved *precision*, which focuses on the number of returned documents relevant to the user's information need, and *recall*, which is the fraction of all available relevant documents that have been retrieved by the query. Information retrieval is also more focused on textual data. In this thesis, we use concepts from IR to evaluate our proposed algorithms for text classification, but also data mining and machine learning techniques to represent our data and discover hidden knowledge in it.

Moreover, *machine learning* (ML), an application of artificial intelligence, can be

seen as a subset of data mining. It involves *supervised learning*, e.g., regression and classification, where the models are trained with given data to be able to perform a specific task on unseen data. It also refers to *unsupervised learning*, namely, input data given to the model, but not necessarily an output target (e.g., clustering similar product descriptions in online shops or user profiles in social networks). In general, machine learning techniques and especially supervised classification models learn, reproduce and predict known knowledge, whereas typical data mining methods discover unknown patterns in the data, often without assistance. When the data that is analyzed consists of textual documents, then the evaluation of the models' performance is often done with the above-mentioned IR metrics, i.e., precision, recall and other metrics based on them. In this thesis, we mainly apply and improve supervised text classification approaches, and we also use unsupervised techniques to either represent numerically our input text data or to explore and analyze them prior to our tasks.

A big part of machine learning is the use of data, and specifically *labelled* or *annotated* data. A dataset that is labelled is essentially augmented with additional knowledge. For instance, news articles with labels for their discussed topics can be used by a topic detection algorithm that learns to categorize articles topic-wise. A dataset with product reviews annotated for their expressed sentiment could be used to predict which review is satisfied by a product and which not, and eventually learn user preferences. Labelled data can serve both as input of a model that learns a specific task, and also as test data, in order to evaluate the performance of the technique. In this thesis, we will analyze text corpora that is sometimes labelled automatically with metadata (e.g., news articles classified to their news source and tweets classified to events they mention) and also annotated by crowd-workers and domain experts. *Crowd-sourcing*[1] can be a very helpful tool to obtain training data for machine learning algorithms in a cheap and efficient manner. Obtaining domain-dependent expert annotations, for instance when doctors assign the name of a health condition to a patient's symptom description, can be powerful as well. Namely, domain knowledge is often the only reliable and correct expertise needed to solve a task, e.g., whether a faulty machine is operable or not. In Chapter 3 we use both crowd-sourced and expert data annotations in news articles and we compare their potential for media bias detection.

Moreover, supervised learning contains different kinds of learning techniques, e.g., *active learning*. This is a specific type of iterative supervised learning, where

[1] https://www.mturk.com/

the algorithm interacts with its user and asks her to annotate additional data that will improve its performance. *Reinforcement learning* is another area of machine learning, where a software agent takes actions in a given environment to solve a problem and after each action the agent receives feedback and potential rewards. In this way, the agent learns what the next best possible action is.

1.3 Supervised Learning

Annotated data that is used by a machine learning algorithm for a given task need to be converted to numbers before they are given to the algorithm. Videos are essentially sets of images, and images are transformed into numbers that correspond to colors and pixels, text can be transformed into frequencies of words, social network users can be considered as numerical ids that form connected pairs, triplets, etc. Learning and using meaningful representations of data is an essential and integral part of supervised learning as it can enhance or worsen the performance of a model. Note that the representation learning process itself is essentially a unsupervised learning task. Data is not only represented by its content, but also by its characteristics, namely its *features*. For instance, a feature of a research paper that is categorized into a set of disciplines can be the venue that it is published. A feature of a social media user profile that is categorized for being fake or not could be its username or profile bio. In this thesis, we are focusing on mining textual documents, thus in the following sections we will focus on data representation and feature extraction methods for text data.

Traditional supervised learning algorithms include the *Naive Bayes classifier*, a probabilistic model which is based on the Naive Bayes theorem that assumes independence between all features of a data instance (e.g., the words of an email to be categorized as spam or not). This classifier uses probability distributions, e.g., the Bernoulli or multinomial distributions to estimate the distributions of the features. Naive Bayes is often used as a baseline in text classification [91] and its main limitation is the very strong assumption of feature independence, which is not always valid in real-word problems.

Other algorithms include *decision trees*, where all observations about an entity (e.g., a news article's title, publication time, source agency etc.) are used in a tree structured model that draws conclusions about a target value of the entity (e.g., the popularity that the article will reach). Decision trees, such as Random Forests [18] can be used for both classification and regression problems. They are very intuitive and easy to interpret, and can be very effective in drawing conclusions without

much data preprocessing and preparation. However, as all rule-based approaches, they might *overfit* the data and a small change in it will result to a big change of the model. Overfitting occurs when a machine learning model is trained to perform very well on a given dataset, but fails to generalize and solve the desired task on unseen data. *Underfitting* is the opposite phenomenon, where the model is unable to learn the patterns of the training data.

Another classification method is called *support vector machines* [31], where a discriminative model is build to detect the category that a new item belongs to (e.g., the topic of a discussion in social media). Support vector machines (SVMs) are often effective with high dimensional data, but it can become cumbersome to train them with large datasets. They are also not easy to explain, because of the absence of probabilities in their decisions. We use Random Forests and SVMs for text classification in Chapter 2 to assign news snippets to news sources.

Linear regression and *logistic regression* are two statistical models that are also popular for supervised learning and are often used as baselines. Even though logistic regression can be seen as a special case of linear regression, the assumptions the models are based on are very different. Linear regression is an algorithm that solves regression problems where the target variable is continuous, e.g., predicting a person's weight. Similar to other statistical models, linear regression assumes independence in the data, which is not often correct, and is sensitive to outliers. It also draws "simple", i.e., only *linear* conclusions between the independent variables, and the mean of the dependent variable.

Logistic regression models predict probabilities of the outcomes, rather than the outcomes themselves. They could also solve classification tasks using thresholds on the derived probabilities. Logistic regression, often thought of as a one layer neural network, is used when the prediction refers to a binary variable, e.g., predicting whether a person appears in a picture or not. Similarly to Random Forests, both approaches are simple, interpretable and do not require a lot of data preprocessing to work well. However, they are also prone to overfit and are in general outperformed by more sophisticated approaches, such as deep neural networks with sufficient training data.

A major disadvantage of the above-mentioned traditional supervised learning algorithms is the need for *feature engineering*. That is, developers need to extract additional information from the data in order to help the classifier learn patterns in them. Such information for a topic detection classifier could be the title of a news

article, the hyperlinks to other articles, the part of speech tags, the names of persons and companies mentioned in the text etc. The model can then learn patterns in the data by observing the dependencies between these features and make conclusions about the target problem, e.g., a recommender system would decide for the relevance of a news article for a given social network user. It is also common that in traditional machine learning algorithms data often needs to be thoroughly preprocessed (e.g., *stemming* or *lemmatization* to simplify word search in text), cleansed (e.g., removing punctuation in text), because models often suffer without these steps. Hence, although such methods can be useful, are mostly used as baselines and compared to more independent and generalizable models that do not required much data exploration to perform a task. *Feature selection* and *dimensionality reduction* algorithms could be a way to eliminate candidate features that are irrelevant. However, deep neural networks discover features in the data while training themselves and they often do not require human intervention, even if the data is noisy or incomplete. In this thesis, we rely on well-designed textual features for text classification in Chapter 2 and 4, whereas in Chapter 3 we do not engineer our own features, but rely on our model to discover them.

1.4 Deep Learning

Deep learning is a family of algorithms that are based on artificial neural networks and their learning process can be supervised, unsupervised or semi-supervised. Neural networks, such as simple *perceptrons*, simulate the way a human would make decisions. Artificial neurons are mathematical functions that are based on a model of biological neurons, where each neuron takes inputs, weighs them separately, sums them up and passes this sum through a nonlinear function to produce output. Artificial neurons are the fundamental unit of an artificial neural network and their connections carry information about the input that is given in the network.

In computer science, there are several neural network architectures, e.g., *recursive* neural networks with tree-based structures [151] and *recurrent* neural networks that unfold over time and are essentially constructed by stacking multiple recursive layers [48]. Recurrent neural networks (RNNs), which make use of sequential input, contain memory cells, with a widely-known cell being the *Long Short-Term Memory* (LSTM) [155], which we utilized in Chapter 3 to classify news articles. At each step of the information processing, an LSTM considers the current data (e.g., the current word of a document), its state and the previous data it has already seen. LSTM networks can be particularly powerful, because of their memory and their

ability to cope with challenges that other RNNs cannot (e.g., the vanishing gradient problem). However, LSTMs tend to be very complex and thus time-consuming to train (cannot be parallelized) and difficult to obtain an optimal solution to the problem they solve. In general, in text mining, simple feed forward neural networks are not able to use enough context for their decisions when relevant words are far from each other. LSTMs solve this issue with selectively remembering or forgetting information. This is one of the reasons we use LSTMs in Chapter 3, hypothesizing that bias is subtle and implicit in the text, and thus more context would help a machine learning model to perform better in discovering the hidden bias in the text. Despite this attribute, LSTMs sometimes suffer from very long dependencies and that they only consider linear distances. To overcome these challenges, attention mechanisms help [163] and convolutional neural networks (CNNs) [65] which can be parallelized and exploit local dependencies. The latest state-of-the-art neural network that is more able to tackle existing challenges than previously introduced models is the Transformer architecture. This type of networks that are based on Transformers [37] contain attention mechanisms and read the sequence of input data all at once, which enables them to consider all the surrounding context and make better decisions for their task.

Deep learning has become a major part of data analysis in the last decade due to its unique ability to identify patterns in data without prior knowledge of the data characteristics and features. It has gained a lot of attention in *natural language processing* [37, 126] (NLP), *machine translation* [77], image, video and audio analysis [115], bioinformatics [57], etc. Applications of deep learning in information retrieval have also emerged with neural language models (that map words to their meaning as a vector of features), which can answer related document search [116]. However, it is worth noting that there are still difficult problems even for humans, e.g., detecting irony or sarcasm in text, and machines suffer to succeed in these tasks as well. There are also challenging tasks that machines compete humans to, but do not outperform them yet, e.g., reading comprehension with complicated questions [70, 170]. Deep neural networks require significantly much more data than a traditional machine learning algorithm, but they often do not need data labels to solve a task. In Chapter 3, in order to overcome the small size of our training data for media bias detection, we explore *transfer learning*, and particularly *curriculum learning* [11] in artificial neural networks, to facilitate our classifier learn the difficult patterns in our news data.

1.5 Text Mining

Document classification problems, such as manuscripts to research areas for organizing digital libraries, tweets to topics for identifying trends, books to subjects for recommendation purposes, etc. is a very common and timely task, given the amount of data that need to be organized in almost every domain. This thesis uses various text classification techniques as the main tool to solve research problems in the area of political news and social media analysis. We utilize both traditional rule-based algorithms, and deep neural models as well.

Machine learning algorithms cannot work with raw text data directly, so the input must be converted into numbers. Specifically, each data point is converted into a vector of numbers that are supposed to represent its meaning and characteristics. There are various ways to numerically represent textual data that is input to a machine learning algorithm. The easiest way is to ignore the order and syntax of the words in a document and regard them as a *bag of words* [60]. Given a collection of documents with `n` words in total, a simple implementation of the bag-of-words model (Bow) would be to represent each document with a binary vector of size `n`, where each position of the vector contains 1 if a word appears in this particular document and 0 otherwise. Another option would be that each word is represented by its document frequency in the respective document. In this case, very frequent words might dominate the feature space (e.g., “the") even though they do not reveal any semantics of the document. An approach to compensate for this problem is called *Term Frequency – Inverse Document Frequency* (TF-IDF), where the scores of each word are normalized based on their frequency in the whole corpus. The above-mentioned techniques are easy to implement and to interpret. However, they entail several limitations, e.g., they ignore word order and they don’t scale with large vocabulary sizes and document collections. An approach that tackles the issue of simplicity in the Bow model is the *n-gram* model, where words are grouped in units based on their order. This model compensates for *out-of-vocabulary* words by assuming that the probability of a word to appear in a document only depends on its previous k words.

More recent *language models* include neural networks that represent text in continuous spaces or embeddings. Even though such methods represent text in a more sophisticated and meaningful way, they sometimes suffer from the *curse of dimensionality*, as word sequences increase exponentially. An example neural representation algorithm is the *skip-gram* [104], which is essentially a generalization of the n-gram model. This model allows the units of `n` words to have gaps between each

other and in this way compensates for the problem of high dimensions and sparsity in the data. A recent approach to represent text is called *word embeddings*, where the goal is that words with similar meaning will appear in nearby positions in the embedding space and thus have similar numerical representations. Each word vector contains real values in a predefined vector space. Such text representation techniques produce vectors are that are significantly shorter than more naive representations (e.g., binary representations with *one-hot encodings*) and thus they compensate the above-mentioned sparsity and high dimensionality in the data. The skip-gram model belongs to this category of language representation models, along with other widely-used approaches such as, *Word2vec* [105], *GloVe* and *ELMO* [126]. The difference between Word2vec and Glove is that the first is repeatedly iterating over the training data, while the second is trying to fit vectors to model a word-word matrix that is built from the given corpus. Due to the fact that they both do not take into account the word order and the context around the words, they both suffer from not being able to handle out-of-vocabulary (OOV) words. More recent approaches such as ELMO and BERT [37] are context dependent, which means that they produce different word embeddings for the same word depending on its position in the text. In this thesis, we make use of text representation techniques. For instance, regarding media distinctiveness in Chapter 2, we initially use TF-IDF to represent news texts [84] and later on, we apply Standford's Glove word embeddings [86] to improve our previous approach. In the future, we plan to use context-dependent approaches to further enhance our performance on classifying news.

The field of *text mining* involves several tasks apart from document classification, e.g., document summarization, where an application could be to generate a summary of a news article and show it next to the article in a search engine. Another research problem is topic detection, i.e., to find the topics that are discussed in a document, e.g., news articles, blog posts and tweets. This is often achieved by *topic modelling* techniques, such as Latent Dirichlet Allocation [15] (LDA) models and their extensions. A topic model is a probabilistic graphical model that assigns "abstract" topics (a topic is a set of keywords) to documents of a collection, using the assumption that the writer of the corpus generated the documents based on a certain topic distribution in mind. Each document is considered to be a random mixture of various latent topics and each topic is represented by a distribution over all the words. We apply LDA on textual messages in social media in Chapter 4 to discover the main public interests in the context of politics. As LDA is a rather outdated and also hard to interpret without human intervention, we also create our

own data annotations for political topics in social media and compare our machine learning approach with the topics discovered by LDA.

Other tasks relevant to this thesis include *named entity recognition*, with entities being persons, locations and organizations, and *named entity disambiguation* and *linking*, which decrease the errors and ambiguity of named entity recognition techniques. We apply these techniques in Chapter 2 to find important words in news comments and also in Chapter 4 to discover mentioned individuals in social media discussions. *Sentiment analysis* and *opinion mining* [96] are also two research problems related to this thesis. Their goal is to detect the expressed sentiment or opinion in a textual document as a whole or in different segments of it (e.g., by analyzing product reviews, online shops could find out how much customers like a product and what do they think about its features). In Chapter 2, we train our own supervised sentiment detection model to discover the journalist's expressed sentiment in political news articles.

1.6 Mining Political Texts

Political documents, such as news articles, politicians' speeches, tweets, blogs, etc. can contain powerful information to study social behaviors. With the right tools, scientists could extract insights on the opinions of the general public and the journalists about ongoing affairs. The motivation for obtaining such knowledge is that it can give us a say in shaping our collective future. It can offer multiple perspectives on society, modern and transparent democracy, and it supports us in holding our politicians and our media to account.

The main advances of political media analysis have been observed in political and social sciences, especially for challenging tasks for humans, such as news media bias [56]. In summary, research in computer science studies political text in social networks, especially regarding topic, event and sentiment detection [78, 138, 141], which we will elaborate on further in Chapter 4. It also includes works on news analysis that discover topics [43], memes [89] and sentiments [5], oftentimes with specific focus on blogs [49] and financial news [145]. Political news articles often attract studies for challenging tasks even for humans, e.g., ideological perspective analysis [95] and fact checking [165]. These studies mainly originate in political sciences and recently appear in computer science as well [56], which we will examine closer in Chapters 2 and 3. Works on political social media posts include analysis of debates [30] and detection of hateful language [9].

One great challenge when analyzing political news articles is that the patterns that

one is trying to discover (e.g., media bias) are not concentrated in one part of the document (e.g., as oftentimes in question answering). The evidence of media bias, such as discriminating against a political party, could be subtle and located in various articles of a given topic. It can also be relative to the reporting behavior of several newspapers, which requires a more complex analysis. Another example is fake news detection, where sometimes, similarly to media bias, it is challenging even for humans to find its evidence in the text.

Moreover, studying political text in social media entails known challenges related to user generated content, e.g., the informal language of the users and the constant evolution of the writing style. One example is the introduction of new hashtags and the appearance of new trending topics online. Another difficulty is the existence of sarcasm, irony and complex language (e.g., negation), which can easily confuse a model with irrelevant information. Social network posts often contain fake political information, e.g., generated by bots, as well. Thus, both in news and social media analysis, especially in the context of politics, obtaining ground truth annotations is a very challenging task. This is due to the ambiguous and often ill-defined phenomena we are looking for in natural language, which are even more apparent in the media domain and not more restricted ones, such as in legal documents, financial reports, research books, etc. In addition, the volume of text data in media can be extremely large, especially when considering real-time systems that process documents constantly. In 2018, it was reported that 456,000 tweets are sent on Twitter every minute of the day and 1.5 billion people are active on Facebook daily. In 2019, over 4 million blog posts were published every day.

1.7 Contributions

In this thesis, our objective is to bring science a step forward towards a trustworthy and transparent media ecosystem. We focus on political text analysis in both news and social media. We contribute to the first research area with two lines of work, i.e., on the article level [87] and the paragraph level [50, 83, 84, 86], namely analyzing quotes from news articles and comments by news commenters in the latter. We contribute to the area of social networks with our work on short documents, i.e., analyzing political debates in Twitter messages [54, 85]. Our goal is to provide more context and insights in the way political content is written on the Web by both social network users and news reporters. We also introduce novel annotated textual corpora, i.e., for news bias and also tweet topics, in order to achieve high quality results in our analyses, but also to contribute to future research.

In news analysis, we tackle the general research problem of the quality assessment of online news. Our first goal is to understand how distinct media outlets are from each other and how predictable they can be in the way they describe the news. On the one side, we study how media cite politicians and parties and whether they discriminate for or against certain entities in a unique way [83, 84, 86]. On the other, we study the commenters of newspapers and observe whether each outlet attracts different kinds of users [50]. We model both problems as classification tasks. Namely, we classify quotations and comments to their originating news source as a way to show how predictable and distinguishable media outlets can be by observing their or their followers' language. Our second research question is to find out how biased a news article is as a whole. We model this problem as a classification task as well, i.e., we introduce novel news datasets and categorize them as biased and unbiased [87].

In social networks, we analyze millions of tweets in order to find insights of the ongoing topics and events. We aim to discover event information in social media about given individuals and their location. That is, we study whether a person's location can be obtained only by looking into what others say about him/her online. We also model this as a classification task and we showcase the performance of our approach in a tweet dataset with political discussions. Namely, after we prepare our corpus and filter the existing noise, we classify each tweet to whether it contains a valid location of a target person of not [54, 85].

In all three chapters of this thesis, we introduce solutions to data science problems that are based on document classification. We utilize various of the above-mentioned supervised learning algorithms and also multiple ways to represent our text data. The individual learning and evaluation approaches are described individually in each section as appropriate.

1.8 Thesis Outline

This doctoral thesis is organized as follows: Chapter 2 discusses the differences in the way newspapers report political news and how distinct they can be when looking at some of their characteristics. It contains our work on the quoting patterns of British newspapers [83, 84, 86] and on the news commenters of German news outlets [50]. Chapter 3 presents the problem of political media bias detection in the news and describes our novel classification approach and dataset [87]. Chapter 4 is examining social networks and political discussions in them. It contains our research on political debates on Twitter in the context of the US election in 2016 [54, 85]. Each individual chapter contains its own related work and future work section.

Chapter 5 outlines our conclusions based on the research conducted in this thesis and presents ideas for future work.

Chapter 2

MEDIA DISTINCTIVENESS IN POLITICAL NEWS

News media shape the public's opinions, perceptions and reactions to current affairs. Studying media outlets, such as newspapers and news blogs, is the task of identifying systematic patterns in the way they report the news and, in turn, understanding the lenses through which we view various topics and issues. These patterns can be found in word choices (*terrorists* versus *freedom fighters*, *death tax* versus *inheritance tax*) or topic preferences[1] of a source. They can also be observed in the quoting choices of a source, e.g., the peoples' voice the newspaper believes are worth given representation. Another example are the positive/negative reactions of the newspapers' readers in comment threads under news articles, e.g., commenters may bring further facts about the current affairs that were neglected by the reporters.

In general, a media outlet is considered biased when it expresses (subtly or obviously, accidentally or deliberately) its ideological beliefs and agendas – thus it becomes less objective. Unfortunately, the distinction of the newspapers according to potential party endorsements into left-wing and right-wing, or liberal and conservative, is not always given or obvious to infer, and it might also change over time [161]. For instance, new journalists or editorial board members might join a given news outlet and the publishing policies could change. That is why, in this chapter, we focus on the reporting behavior of newspapers by identifying relative differences between them instead. This task is independent of the outlet categorization into political orientations and the discovered patterns serve as bias indicators. We refer to our goal as discovering media *distinctiveness*, namely, we identify how unique and distinguishable media outlets can be from one another, when discussing political news events and expressing different political opinions about them.

We classify political reported speech extracted from news articles [83, 84, 86] and also news comments from comment threads under news articles [50] to their respective original news source. We apply machine learning techniques for binary and multi-class text classification, leveraging various features of the quotations and the comments, and we match these input documents to the newspaper they belong. Our goal is to capture systematic quoting and commenting patterns by journalists and

[1] https://www.washingtonpost.com/blogs/erik-wemple/wp/2015/10/23/why-fox-news-ditched-the-benghazi-hearing-and-msnbc-didnt/

news commenters respectively, which can lead us to the newspaper where the quotes and comments are originally published. We hypothesize that the more predictable a newspaper is, the more constant (and potentially biased) language patterns it uses to describe political news. By observing the differences in predictability of each source, we can distinguish between sources with more diverse content (not easily predictable, more balanced, potentially unbiased) and ones that are predisposed to some reporting patterns (more predictable, mostly inclined to one perspective, potentially biased).

We argue that the political orientations of news media can be reflected in their quoting patterns, but also in their respective audience and the kind of comments the users leave. For instance, a certain newspaper might quote members of a political party more often than another one, or criticize it more intensively than others by framing its statements with loaded context (“He embarrassed himself by saying that ...", “She shamelessly warned the parliament that ..."). Furthermore, comments under news articles or tweets that share and discuss news articles can also contain opionionated content and bring new insights about the newspapers’ political views. Such user generated content could provide further clues about the given newspaper, e.g., a conservative commenter might leave a disapproving comment under a liberal article, or an insightful commenter could report new facts that the journalist neglected to report. We are interested in these novel and indirect ways to discover hidden patterns in the news storytelling in pursuit of eliminating the readers’ misinformation and assisting the journalists’ to reflect on their work.

This chapter presents our three contributions in the area of news analysis and distinctiveness. We first refer briefly to our vision paper [83] that shows some preliminary insights into the distribution of mentions and quotations of politicians in British media (articles from the Guardian and the Telegraph in 2000–2015). Our conference paper [84] introduces our novel aspect of media bias detection via news source classification and our journal article [86] (under submission) contains an extended and more comprehensive version of our study with additional datasets (Brexit related articles in 2016–2017 from the Guardian, Telegraph, Independent and Daily Mail). These three works are described in Section 2.1. Moreover, our conference paper [50] based on the master thesis of our student Christian Godde – co-supervised by Dr. Ralf Krestel and the author of this thesis, analyzes news articles and comments by six German newspapers. That is, we consider Bild, Focus, Welt, Spiegel, Zeir, and Faz, in the context of media bias in 2016. We present this contribution in Section 2.2 and all reported results are obtained by Christian Godde,

with the research paper itself written by the author of this thesis.

The remainder of this chapter is organized as follows: In Section 2.1, we initially explain how media bias is expressed in reported speech and outline various related works. We then present our news datasets, introduce our speaker detection algorithm in quotations and our method for detecting bias in reported speech based on various textual feature sets. Furthermore, we depict our classification results and summarize this line of work. The approaches we follow of Section 2.1 to detect bias in the news originate from our most recent and advanced work [86], while we also reference our preliminary study [84] when appropriate, in order to show specific differences and improvements. In addition, Section 2.2 presents our contribution to news comment analysis [50]. Initially, we introduce the problem and cite related work. Moreover, we describe our news datasets and our classification features. Lastly, we display our results for comment classification and summarize our work.

2.1 Classifying Reported Speech

Reported speech analysis can reveal interesting patterns in the way newspapers discuss politicians and their statements. Media bias can be present both in the choice of reporting an event, e.g., a senator's announcement, but also in the words used to describe the event. For instance, framing a utterance with positive or negative context, citing only parts of it, etc., are essential reporting choices that can predispose the reader. They have the power to affect the readers' viewpoints on the discussed topics, and by extension they could also influence their voting behavior.

In this work, we aim to discover political media bias by demonstrating systematic patterns in the quotations of major British newspapers [83, 84, 86]. We classify each extracted quotation automatically to its respective news source, as a means to show how unique and predictable media can be in the way they cite political statements. Thus, we define and model the problem of bias detection in political news reports as a supervised classification task, i.e., we train a classifier that receives quotations and their features to assign them to the news source they originate from. We use deep learning and different kinds of bias indicators in reported speech, and we show that the context different media outlets use to present political news stories varies. We discover interesting insights when considering widely-published quotes or prominent politicians, and that the newspapers can be more predictable for certain parties than for others.

2.1.1 Media Bias and Political Journalism

A media outlet is considered biased when it expresses (subtly or openly) its opinions and ideological beliefs – thus it is less fair and objective. There are also news articles that are opinion pieces, which contain bias by design and intentionally, but also in a transparent manner. These do not belong to the scope of our work. Moreover, since the news pieces are generated by humans, it is possible that either accidental or deliberate bias is introduced in the text. Due to the above-mentioned difficulty in separating newspapers into their political perspectives [161], we analyze the reporting behavior of newspapers by identifying relative differences between them instead. We aim to find possible differences in the way news media report political utterances and bring these patterns into light, because they can influence the public opinion heavily. Analyzing media can assist us in understanding the extent of this influence and bias. In addition to the readers' assistance, quantifying bias in the news can also support journalists to reflect on their work. In general, a system that identifies opinionated text or context in the news would serve two goals: it would advance our understanding of how media communicate information under subjective frames and it would offer guidance for journalists who aim to report news fairly and balanced.

Moreover, we categorize media bias types according to the literature as follows: neglecting to report specific news stories, covering and framing others in an non-objective manner, and finally adding opinions to them [136]. The first decision that a reporter faces with the emergence of an event is whether it should be reported or not. This type of media bias is called **selection bias** [17]. It is based on the importance and interestingness of a topic (both for writers and readers), such as the physical location of the news source and the news story, the preferences of the outlet's target audience, the owner's views, the publishing guidelines of the outlet, etc. For instance, according to a recent study about terrorism, Muslims commit fewer terrorist attacks than non-Muslims, but when attacks by Muslims do happen, they are reported almost 4.5 times more often than other attacks [74].

Another kind of bias is called **coverage bias** and refers to the completeness of an article, in terms of the reported facts and aspects about the discussed event. For instance, prior to general elections or a referendum, it would be expected that media cover a wide range of party statements, since both the opinion and the vote of the readers could be shaped by them [38]. However, this is not always true and media tend to cover more stories about the current governing party in the U.K. [83] and criticize disproportionately different political parties in the U.S. [19] – an expression

of **partisan** and **gatekeeping bias** [45]. Namely, they are more critical with the parties that they do not share the same beliefs with, but at the same time they are not more supportive to the parties that they endorse. It has also been shown that media coverage in Germany affects voting behaviors and party affiliations [38].

Furthermore, the language a fact is described with, either positive, neural or negative, is called **framing bias** [108]. News framing depends on the journalists, but also interest groups, policy makers, and others who have influence the agenda of a news outlet. For instance, "Team Trump *embarrassing gaffe* by spelling PM's name "Teresa" THREE TIMES in press release"[2] by DailyMail, frames the news with an opinionated context. In contrast, the author's explicit remarks on a news topic constitute a different bias type, **statement bias**, which is clearer in the text than framing bias. For instance, "Despite her assurances, Theresa May *doesn't care* about EU-based expats"[3] by the Guardian, explicitly gives the opinion of the journalist on a given political figure and as expected, it belongs to the *Opinion* section of the newspaper.

Reported speech is an integral part of news storytelling. It is used by the media as an element of argumentative discourse to inform and persuade readers [149]. Our hypothesis is that one representative example of media bias is the choice of reported speech, where journalists are responsible of deciding whether and how they will present a person's utterance. Considering quotes from politicians, the way that media report and frame them could reveal the source's beliefs and thus introduce political bias in the news articles.

In Figure 2.1, we show the news production process and which steps we believe might contain media bias. From left to right, a new event happens and several politicians by different political parties make statements about this event. At the next step, news producers decide which statements they will publish, and by extension to which extent they will cover this story and its aspects. Thus, there could be selection and also coverage bias at this stage. The utterances might be filtered even further due to lack of space or political preferences of the news outlet and the target audience. In this way, the event is covered in less depth, with coverage and partisan bias being introduced in the news. Additionally, some of the selected quotes might be shortened (coverage bias) or even presented with opionionated context (framing bias). At the

[2] http://www.dailymail.co.uk/news/article-4163184/Team-Trump-gaffes-spelling-PM-s-Teresa.html

[3] https://www.theguardian.com/commentisfree/2017/mar/06/thesea-may-doesnt-care-eu-expats-brexit

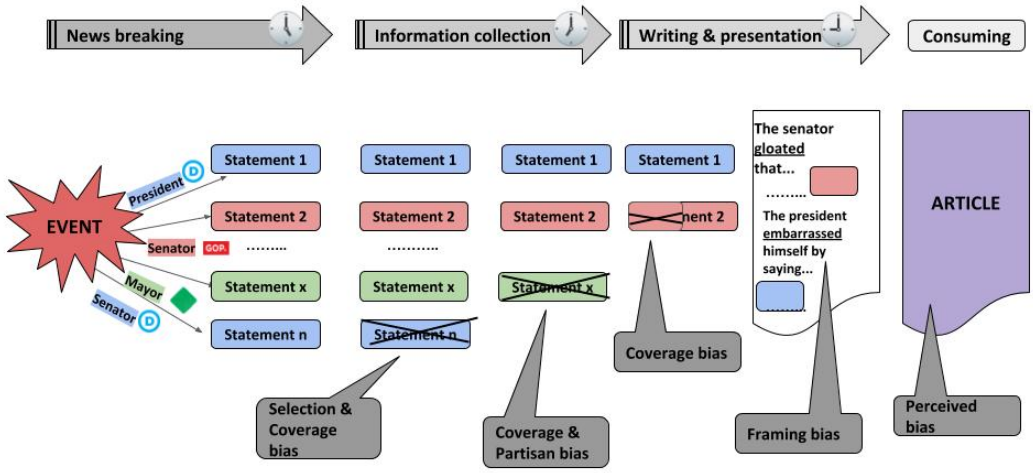


Figure 2.1: Media bias manifestations in reported speech during new production.

end, the reader is also sensitive to her own bias (perceived bias) based on her social background and beliefs. This example shows how the overall news production and consumption process is susceptible to bias and that all above-mentioned bias kinds can appear in many phases of the process, not limited to the ones we demonstrate.

In Figures 2.2 and 2.3, we can see our first insights into how two major British newspapers discuss politicians and their announcements [83]. It is apparent that both newspapers discuss more about the current governments than the remaining parties and the curves of the two popular parties cross at the general election year of 2010 in both news outlets. That is, both parties are mentioned more during their term — 1997–2010 in the case of *labour* and 2010–2015 for the *conservatives*. In addition, it is interesting that until 2010 *labour* is discussed almost two times more than the *conservatives* in the Guardian. The latter is in accordance with Guardian's Wikipedia page stating that its politician alignment is centre-left. However, while *labour* is not exceptionally discussed during Tony Blair's term (1997-2007), the references grow rapidly during Gordon Brown's tenure (2007-2010). The *conservatives*' mentions are increasing during Gordon Brown's term as well.

We also compute preliminary experiments on the politicians' quotations in the news and discover that *labour*'s quotes in 2004 are three times more than the ones from the *conservatives* and twelve times higher compared to the *liberals* [83]. Similarly to the mentions, 2010 is the first year that the *conservatives* outperform *labour* in terms of the media coverage of their quotations and this phenomenon remains until 2015.

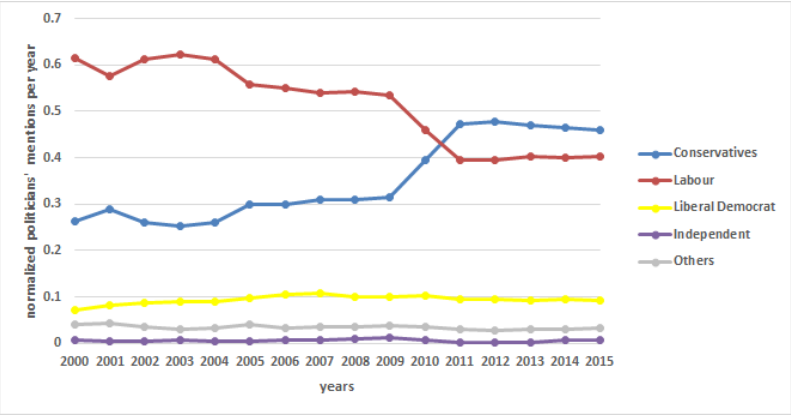


Figure 2.2: Annual politicians' mentions in the Guardian from January in 2000–2015, normalized and aggregated for each political party.

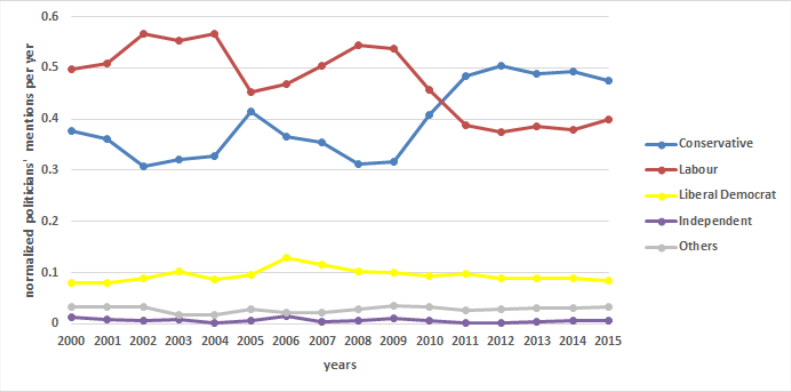


Figure 2.3: Annual politicians' mentions in the Telegraph from January in 2000–2015, normalized and aggregated for each political party.

All above-mentioned insights motivate us to analyze politicians and their references in news media in more detail, and uncover potential media bias manifestations that favor or oppose certain perspectives.

In this chapter, we model bias detection in reported speech as a classification problem. We use direct and indirect speech identified in news articles [142], extract various characteristics of the quotes and perform our analysis for major UK news outlets. We group our features based on the different bias types that each feature set can reveal. In this way, we can also show which bias types are present in every news outlet. Our approach classifies each quotation based on its characteristics to its respective news source and by solving this task we were able to show how specific and unique language each media outlet uses when reporting utterances. Therefore, our contributions include:

- Analyzing media bias in reported speech for news datasets, initially for

two [84] and furthermore for four newspapers [86]

- Detecting different kinds of bias in reported speech with different features and deep learning models
- Defining our context-aware approach for detecting quotation speakers in the news domain
- Interpreting our results to find connections between parties, newspapers and their reporting behaviors

2.1.2 Related Work

In political news, a writer's view often depends on his/her political affiliation, hence it can be identified as conservative, liberal, progressive, environmentally friendly, etc. Thus, identifying **disputed topics** that journalists have diverse opinions about is closely related to political media bias detection. Existing research in this field includes a new version of the link analysis algorithm HITS, which identifies the main disputants of a topic in Korean media and classifies the news articles into different viewpoints of a story [120]. In addition, related research is performed on topic-based bias analysis in news articles. De Clercq et al. extract the discussed topics in UK and US newspapers by making use of DBpedia links [35]. Sentiment and subjectivity are measured for each topic, in order to discover conflicting topics in the news. The presented results are promising and they motivate us to investigate whether English-speaking media systematically differ not only in the way they discuss topics, but also in the way they report quotations.

An alternative way of detecting media bias in the news is to take into account the **reader's political affiliation**. Among others, user comments [120] and user reactions [182] are exploited in order to predict the political position of the media. The authors assume that a liberal reader will express a negative sentiment on a conservative article and a positive one to an article that favors the liberals [120]. Additionally, a reader's beliefs affect not only the comments they leave, but the way they interpret the articles. Thus, perceived bias not only depends on the writer's language use, but also the reader's viewpoints. As shown in an economical study [52], one is more likely to perceive bias the further the slant of the news is from one's own political position. Furthermore, Saez et al. [136] analyze the characteristics of 100 English-speaking social and news media sites in terms of different **bias metrics** in a two-week period. The authors show that bias is more frequently observed in social than in news media. They also illustrate that selection

bias metrics are not as indicative as coverage metrics, which provide more interesting evidence for bias, especially in the political context. This finding is incorporated in our analysis, by using coverage bias as a feature of classifier. Leban et al. [88] compute several bias measures as well, e.g., quantifying the readability of news articles, comparing the speed of reporting in terms of the publication time of the same news story across different media outlets, etc. As expected, geographical coverage is discovered to be the most dissimilar among 30 newspapers worldwide. Differences in the word choices are also observed, with the analysis of sentiment and opinions in the news planned to be future work.

A novel line of research performs a linguistic analysis of **hyperpartisan** (very biased) and **fake news** and shows that these reports of artificial events are often politically biased [128]. Other studies operate on a sentence level, e.g., analyzing the choices news outlets make for their headlines [171]. An additional relevant work generates titles for each article with the opposite ideology of its own, based on the ideologies provided by the website Allsides.com[4] [25]. Related work on discourse and communication explains the importance of reported speech as an element of **argumentative discourse** in newspaper articles in the UK [149]. It is shown that the syntactic and linguistic features of reported speech depend on the political position of the article author and can affect how the reader interprets the news. Inspired by these findings, we use machine learning techniques with several textual features of reported speech to predict the newspaper that a quote originates. We hypothesize that the more predictable media are, the more distinct and consistent patterns they have when reporting the news.

The way that events, such as politicians' public statements, are described in the news is shown to influence the readers' perception of these issues [10]. For instance, Schuldt et al. discover that belief in "global warming" is significantly lower than in "climate change", specifically among Republicans [143]. In general, there are different **language constructs** that can influence the reader's viewpoints on current affairs, such as framing, subjectivity, sentiment, and bias [10]. In this work, we address the problem of political media bias detection and we utilize the sentiment expressed around politicians' statements: we claim that media outlets do not report all statements with objective context. **Sentiment analysis** is mostly studied for short documents, such as product reviews, news comments and tweets, where both the opinion and usually the opinion target are explicitly mentioned in the text [68]. Related work on mainstream news media is a dictionary-based sentiment analysis

[4] https://www.allsides.com/unbiased-balanced-news

approach by Balahur et al. [5]. The authors annotate and analyze only sentences from the news articles that correspond to quotations, because it is likely that these sentences are more subjective and they express the speaker's opinion. Another advantage is that the source entity (quoted speaker) and target entity (a person mentioned in the quote) are also given. It is shown that the sentiment of the immediate context of an entity can be detected easier than the sentiment of longer text segments where entities are mentioned. More recent related work on **perspective analysis** in the news relies on the Stanford Sentiment Treebank [152] to predict sentiment in the article text [59]. Following the literature, we also utilize the above-mentioned dataset to detect the sentiment in the immediate context of political quotes.

Unlike prior research, we focus specifically on bias in political newspaper articles and we are interested in how it is expressed through reported speech in the text. We perform an analysis of four UK newspapers and our task does not require any external knowledge that classifies media into liberal, conservative, etc. [113]. It also does not depend on manual labeling of existing media slant in news articles, [19, 149], which could be expensive and time-consuming to obtain. We leverage solely the utterances that media cite and the way they report them. Manual labels can also be subject to *annotator bias*. Given a text corpus and a labeling task, the annotator bias refers to the differences between the individual preferences of the various (expert or non-expert) annotators. These differences could prevent them from producing the same annotations and result to disagreement [131].

Bias in reported speech. The only work that is closely related to our analysis is the selection bias framework *QUOTUS* [113], which observes how often political blogs and newspapers quote segments from Barack Obama's White House speeches. The results show that after projecting these quotations into a latent space, some of the outlets cluster together by their political affiliation — for instance, Fox News is unexpectedly close to New York Times. The authors manually classify the media into four categories, that is declared liberal (DL), declared conservative (DC), suspected liberal (SL) and suspected conservative (SC). Initially, *QUOTUS* matches segments of the presidential speech transcripts to the news. Furthermore, it estimates the likelihood that a quoted segment q will be cited by an outlet A of a certain category, given that q is already mentioned by another outlet B of a different category. An interesting result is that *DC* outlets are less likely to quote a statement that *DL* media reported compared to a random quote. This outcome motivates our work, as it brings evidence that quote selection choices among outlets can differ.

An additional task of *QUOTUS* that is more relevant to our work is to predict for a given outlet and quote q whether the outlet will report q or not. Given a bipartite quote-to-outlet graph G, this problem is tackled by performing a matrix-completion approach on the adjacency matrix of G and it yields precision of 0.25, while recall is 0.33 [113]. Although our classification task is similar, we aim at predicting the news outlet in which a politician's quotation is published and by extension we show that the text and context of reported speech is presented differently among newspapers. In contrast, *QUOTUS* focuses only on Barack Obama's speech segments and disregards indirect quotations both from Barack Obama and other politicians.

Table 2.1: Statistics for all news corpora from January 2016 until December 2017.

Newspaper	#Articles	#Quotes	Avg Articles/Day	Avg Quotes/Day
Guardian	15,577	27,797	21.63	38.74
Telegraph	4,956	6,881	6.88	9.55
Independent	10,712	23,163	14.87	31.67
Daily Mail	1,415	4,458	1.96	6.19

In our work, we use a state-of-the-art semi-Markov model by Scheible et al. (SEMIMARKOV) [142], which detects direct and indirect quotes in the text and additionally provides us with the introductory verb of the quote, denoted as *cue verb*. SEMIMARKOV improves the previously proposed linear-chain conditional random field (CRF) [118] for quotation extraction. We also further enrich our dataset by determining the author of a quotation based on the context around the quote and the preceding text of the news article.

2.1.3 Datasets and Statistics

This section describes our datasets and the data preparation steps we perform before we classify the quotations.

2.1.3.1 Political News Articles

We detect media bias in four major UK newspaper, namely we crawled all available political news articles from the Guardian[5], the Telegraph[6], the Independent[7] and the Daily Mail[8] in 2016–2017, which mention the word "Brexit" in the text, so they cover political news about the withdrawal of the United Kingdom from the European

[5] https://www.theguardian.com/international
[6] http://www.telegraph.co.uk
[7] https://www.independent.co.uk/
[8] http://www.dailymail.co.uk/ushome/index.html

Union (such as the EU referendum announcement in February 2016 and polling day in June 2016, the general elections in June 2017, etc.). We choose these sources, so that we cover a variety of mainstream news from widely used newspapers in the UK. Moreover, knowingly opinionated articles, e.g., editorials and blog-style posts are excluded from our collection, as well as live reportages that do not constitute traditional news articles. As illustrated in Table 2.1, the first three newspapers are the most active ones. Note that all utterances are included in these statistics, regardless of the speaker being a politician or not, in order to provide a global overview of our data. In general, large amounts of quotations per newspaper are anticipated and can be justified, because reported speech is an integral and widely used part of journalism and it can appear in up to 90% of the sentences in a news article [12]. The Daily Mail corpus consists of significantly fewer political news articles, due to the tabloid character of the newspaper. In our earlier work, we analyze only the Guardian and the Telegraph in a time evolving manner, from 2000 until 2015 [84]. It should be noted that the number of articles in the Telegraph is much lower in comparison to our previous study due to newly introduced access restrictions to the website in 2016. Namely, a significant number of the reported stories is reserved for Premium subscribers.

2.1.3.2 Political Statements

We apply a state-of-the-art model (SEMIMARKOV) for quotation extraction by Scheible et al. [142]. It detects direct and indirect quotes in news articles and additionally provides us with the introductory verb of the quote, denoted as *cue verb*. Research in quotation extraction in other domains includes approaches for quotation attribution in novels [111] and dialogues [21]. We prefer to use SEMIMARKOV due to its proven good performance in news datasets, i.e., 75-85% f1-score. Furthermore, SEMIMARKOV improves the previously proposed linear-chain conditional random field (CRF) [118] for quotation extraction. Its advantage is that it takes into account the full quote span and makes a joint decision about the start and end points of a quotation. Hence, by analyzing the context of the quotes and considering global information in the text, SEMIMARKOV exhibits higher F-1 score. We prefer to apply SEMIMARKOV instead of the more recent model Quootstrap [122], because SEMIMRKOV provides results on a ground truth dataset of thousands of news articles and it is reports better results on this test set in comparison to related work [117, 118]. Even though Quootstrap's performance is reported to be better, it is solely tested on a very small crowd-sourced annotated dataset and there is no

Table 2.2: Political party names and abbreviations (in parentheses) in the UK.

Alliance Party of Northern Ireland (**APNI**)	Plaid Cymru - Party of Wales (**Plaid**)
Conservative and Unionist Party (**Conservative**)	Scottish National Party (**SNP**)
Democratic Unionist Party (**DUP**)	Sinn Fein (**Sinn Fein**)
Green Party (**Green**)	Social Democratic Party (**SDP**)
Labour Party (**Labour**)	UK Independence Party (**UKIP**)
Liberal Democrats (**Lib Dems**)	Ulster Unionist Party (**UUP**)

comparison with other models in related work.

Prior to our experiments, we remove all quotations with duplicates (either in the same or a different newspaper) from our training and test set, which prunes 2-3% of data instances There are 45,489 news articles in the training set (44,385 unique) and 11,375 in the test set (11,170 unique). Our similarity metric is a fuzzy string matcher based on the Levenstein distance[9] and we consider two quotations as duplicates when they are at least 95% the same (e.g., some symbols or spaces at the beginning or the end of two identical quotes might differ). The duplicate quotes might also have the same characteristics (e.g., same speaker, introductory verb), but at this step we only take into account the quote text, in order to ensure that our learner is not misguided by the same data instances appearing multiple times in one or more classes. In addition, it is possible that one quotation is actually a part of another one. These partial matches are not in the scope of our work and we also do not consider them duplicates. We are more interested in the news context that each individual quotation is surrounded with and we leave identifying (partial) quotes of the same event or topic for future work.

2.1.3.3 Parliamentary Members and Parties

In order to detect reported speech that originates from parties in the United Kingdom, we extracted the politicians' names and affiliations from a publicly available parliament dataset provided by *mySociety*[10]. This corpus contains information about all officially recorded general elections in the United Kingdom, all political parties and their members. During the time period that our analysis covers, the number of party members was 932. The parties along with their acronyms are shown in Table 2.2.

Moreover, we discover 30,845 quotations in the news of 2016 and 31,096 in 2017.

[9] https://en.wikipedia.org/wiki/Levenshtein_distance
[10] https://github.com/mysociety/parlparse

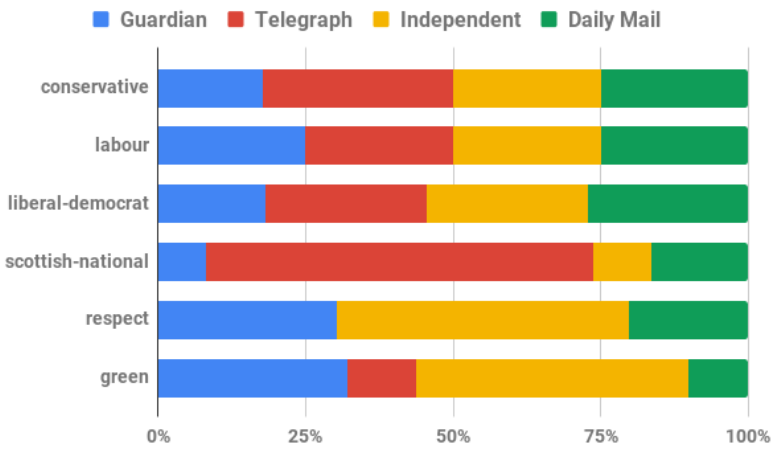


Figure 2.4: For each political party in alphabetical order, the percentage of reported quotes is shown in each newspaper. The numbers are normalized by the number of sentences in each newspaper.

Figure 2.4 shows how well represented each political party is in the media with respect of the citations – a basic measure for political selection and coverage bias. We consider the parties that are mentioned at least once in each newspaper and we depict them ordered according to their overall mentions. One interesting insight is that the *conservative* party is covered more by the Telegraph than the other outlets and that the *labour*'s coverage is considerably balanced among the newspapers. Surprisingly, the quotes by the *Scottish national* party are mainly cited by the Telegraph. In general, the reported speech coverage varies more among small parties, with the *green* party being represented mostly by the Independent. Although it is understandable that the governing and the opposition parties will gain more attention in the media than other parties, it is still noteworthy to see which newspapers are mindful to represent all parties to a certain extent and which are not.

2.1.4 Feature Extraction and Initial Insights

In this section we present our quotation metadata (cue verbs, quote speakers and quote contextual sentiments) and some preliminary insights about them.

2.1.4.1 Introductory Verbs

As discussed in existing literature [149], the quote introductory verbs are essential, in a way that they set the tone for the reader. They describe the relation between the reported speech of the speaker and the author of the article. We use them as a feature in our classification model, because we intuit that they are a representative example of

Table 2.3: The top-10 most frequent cue verbs that the Guardian, the Telegraph, the Independent and the Daily Mail use to frame reported statements from the *conservative* (c) and *labour* (l) party respectively, ranked by frequency in each newspaper

Guardian		**Telegraph**		**Independent**		**Daily Mail**	
c	l	c	l	c	l	c	l
say	say	say	say	say	say	say	say
tell	tell	warn	tell	tell	add	insist	insist
suggest	argue	tell	claim	add	tell	claim	claim
insist	add	add	add	warn	claim	tell	think
claim	believe	suggest	insist	claim	suggest	think	suggest
add	think	insist	warn	suggest	warn	warn	tell
announce	suggest	announce	suggest	insist	accuse	announce	warn
argue	claim	think	ask	announce	insist	add	add
think	call	claim	believe	accuse	call	suggest	believe
warn	ask	accuse	think	admit	believe	accuse	accuse

media's preferred writing style and potential media bias. These verbs can predispose the reader positively or negatively to the reported statement and its speaker: there exist neutral verbs (writes, says, announces, etc.) and more specific ones (intones, hints, fears, admits etc.). We use the cue verbs as provided by the reported speech detector we deploy (SEMIMARKOV – described in Section 2.1.3.2).

In order to shed more light in the quoting selections of the news sources, we perform a comparative analysis of the introductory verbs that each outlet adopts. After applying the Lucene's Snowball stemmer[11] to the complete cue verb list of all news corpora, we rank the verbs by their usage frequency individually in every outlet. For the purpose of introducing political context in the current statistics, we rank the introductory verbs found in each source by their usage frequency separately for *conservative* quotations and *labour* ones. We illustrate the 10 most popular verbs per newspaper in Table 2.3.

The top-2 cue verbs are understandably the verbs "say" and "tell", which are framing approximately 40% of the quotes in all newspapers. One can also easily observe that there is a very high overlap among all columns. In general, in every newspaper we see many common cue verbs in the quotes by both parties. Hence, the usage of these verbs might depend more on the news source than the discussed politicians. On the other side, when considering a certain party the ranking of the cue verbs is different for each newspaper.

[11] http://snowballstem.org/

Table 2.4: Example quotes with different context structure

Burley told the BBC on Thursday: T*hey are launching a preliminary investigation and I ...*	Guardian
Mr **Duncan Smith** said: *"I'm very happy to be guided in that direction"*	Telegraph
And I thought I was having a bad day, Mr **Cameron** added to renewed laughter.	Independent
Mr **Corbyn**, whose efforts so far in the campaign have been criticised as lukewarm, said *Labour was making "the strongest case we can" for a Remain win on June 23.*	Daily Mail

For instance, when Telegraph is quoting the *conservative* party, it is framing the reported speech more often with the verb *warn* compared to Guardian. This cue verb is strongly a subjective and negative word [173]. There are also cases where the reported statements of different parties are framed differently by the same news outlet. For instance, the Telegraph uses the verb *accuse* (another opinionated word) frequently for the *conservative* party, but not the *labour* party. In addition, the word *insist* appears in all newspapers-party combinations in Table 2.3, but it is very rarely used by Guardian when citing the *labour* party. Finally, a surprising finding is that the verb *announce* appears in all four news sources as a popular introductory verb for the *conservative* quotes, but not for the quotes by the *labour* party. This might be justified, because in the time period of our analysis the governing party is the *conservative* party, and by extension its politicians deliver public statements and speeches more frequently than in other parties.

2.1.4.2 Speaker Detection

We further enrich our dataset by determining the speaker of a quotation based on the context around the quote and the preceding text of the news article. We introduce an unsupervised context-based approach to discover the speakers of reported speech in news articles – this approach is briefly discussed in our first work [54] and formulated in the following one [86]. The method is presented in pseudo-code in Algorithm 1. Namely, our technique matches the list of politicians names to the article text. Each quotation belongs to a longer sentence in the article. Initially, for each quote, we determine the location of its cue verb in the sentence (given by SEMIMARKOV), either in the context before or after the quotation (Lines 3-4). We further identify the closest politician's full name to this verb (Line 5 in main and Function `search_in_sentence`).

It is possible that the full name of the speaker cannot be identified with the first try. For instance, such a challenging example is shown in the first row of Table 2.4, where the mention "Burley" has to be linked to the correct politician with this last name. In these cases, we first detect the closest last name to the introductory verb of the quote (Line 7). We then disambiguate the speaker via using the article text prior to the quotation (Line 8 in main and Function `search_in_article`). In the previous example, we discover that the person corresponding to "Burley" is the politician Aidan Burley. The speaker assignment occurs when we find the full name in a preceding sentence of the article, namely in : "Labour calls for whip to be withdrawn from Aidan Burley as prosecutor . . . ".

Since the journalist is responsible of defining the persons that an article is about, we assume that the full name of each discussed politician is included at least once before an abbreviation is used[12]. In addition, given that we are interested specifically in the political domain, a dictionary approach seems more appropriate than a universal named entity recognition tool[13]. Our approach can be generalized and is applicable to news pieces mentioning politicians from other countries as well, by compiling the names of the respective parliament members.

Moreover, we are also able to cope with cases where the journalists use the middle instead of the first name of a politician to introduce their quote. As depicted in the second example, the politician Iain Duncan Smith is abbreviated by "Mr. Duncan Smith". Our technique detects the full name of this speaker, by leveraging a previous sentence of the article, that is "Iain Duncan Smith has mocked Sir John Major as the . . . ". Thus, we can also prune other UK politicians, e.g., Angela Smith and Julian Smith, and successfully select Iain Duncan Smith. When our method is not able to determine a politician's full name as a speaker, it discards this quotation from our dataset.

We evaluate our approach by manually annotating 100 randomly selected quotations for their detected speaker, and we discover that in 70% of the quotes the speaker is correctly identified. The most common error occurs when there is already a preceding error by SEMIMARKOV, namely when the quote extraction mistakes a passive voice construction for reported speech. For instance, the sentence "Ms Mordaunt was *accused* of "plain and simple" lying over the possibility of Turkey joining the EU" is written in passive voice and it is identified wrongly as a quote (as if the politician Penny Mordaunt were accusing someone for lying). Another

[12] https://www.theguardian.com/guardian-observer-style-guide-a
[13] https://nlp.stanford.edu/software/CRF-NER.shtml

```
N: set politicians' full names
Q: list quotations
A: list one article per quote
S: list one sentence per quote
for q in Q with cue verb v in sentence s and article a do
    speaker = string()
    if article_idx(v) < article_idx(q) then
        c = prefix(s) # preceding context
    else
        c = suffix(s) # context following the quote
    speaker = search_in_sentence(v, c, 1)
    if speaker.empty() then
        speaker_last_name = search_in_sentence(v, c, 2)
        if ! speaker_last_namer.empty() then
            speaker = search_in_article(a, speaker_last_name)# disambiguate
function SEARCH_IN_SENTENCE(verb, text, pass)
    mentions = list()
    for w1, w2 ∈ sliding_window(text) do
        if pass = 1 then
            speaker = string(w1 + " " +w2) # search full name
        else if pass = 2 then
            speaker = w1 # search last name
        if speaker in N then
            mentions.add(speaker)
        if pass = 1 and w1 = v and ! mentions.empty() then
            return(mentions.last()) # found full name in context before verb
        else if pass = 2 and w2 = v and ! mentions.empty() then
            return(mentions.last()) # found last name in context before verb
function SEARCH_IN_ARTICLE(article, last_name)
    for w ∈ article do
        speaker = string(w + " " + last_name)
        if speaker ∈ N then
            return(speaker)
```

Figure 2.5: Context-aware quotation speaker detection algorithm in news articles

less frequent error occurs when there is an additional named entity in the context of the quote, specifically between the cue verb and the speaker. For instance, in the sentence “Philip Hammond, who until Wednesday was Britain ’s foreign secretary, a position now held by prominent leave campaigner Boris Johnson, has confirmed that discussions on giving up the presidency are already under way", our approach detects Boris Johnson instead of Philip Hammond as the speaker of the quote with cue verb *confirmed*.

Politicians’ Mentions . Having discovered the speakers of the quotations in our news corpus, we perform a basic experiment for selection and coverage bias. Two recent news bias analyses for German [32] and British [83] media respectively demonstrate that an initial indicator of media bias is the frequency that an outlet refers to a certain political party and its members. Following the literature, we calculate the distribution of politicians’ mentions in our four newspapers in 2016–2017 for the most popular parties in Figures 2.6 and 2.7. With this basic experiment, we aim to show how popular each political party is in the media. The mentions are calculated on a monthly basis and they are grouped by political party.

Based on the differences in the politicians’ mentions, we hypothesize that the speaker or his/her party affiliation can be indicative of a newspaper, especially for small parties, such as the *UKIP* and the *SNP*. We also observe that all news outlets discuss the current governing party more than the other parties, as shown in related work as well [83]. In general, all four newspapers discuss approximately the same 15 parties, with some of them having seats in the House of Commons and some not. We observed a rise of mentions during the last general election days (June 2017) and the referendum (June 2016) in the UK. Two other peaks appear in October 2016 and March 2017, corresponding to the announcement of Theresa May that the UK will start the Brexit negotiation process by the end of March 2017 and the actual triggering of Article 50 respectively.

Finally, the politicians’ mentions are increased at the end of 2017 as well, potentially due to the intensive meetings and negotiations between the UK and the EU. In conclusion, the coverage of stories about politicians in the four examined UK newspapers seems to correlate with how popular political parties are in the elections[14], with the top discussed parties having almost the same rank in each outlet. There are indeed differences in the mentioning patterns of the newspapers, although they are not always indicative of the newspapers or their ideology.

[14] https://en.wikipedia.org/wiki/Elections_in_the_United_Kingdom

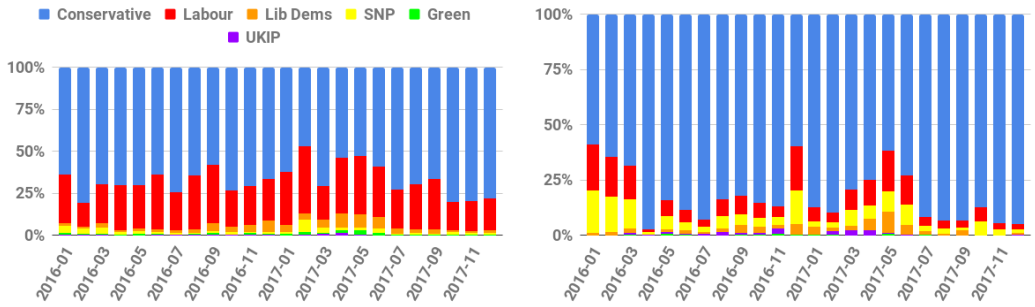


Figure 2.6: Monthly politicians' mentions in the Guardian (left) and Telegraph (right) from January 2016 until December 2017 aggregated for each political party. The party abbreviations are based on Table 2.2.

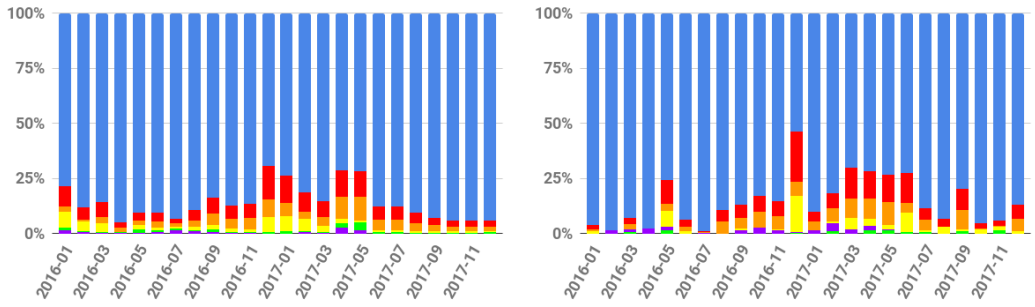


Figure 2.7: Monthly politicians' mentions in the Independent (left) and Daily Mail (right) from January 2016 until December 2017 aggregated for each political party. The party abbreviations are based on Table 2.2.

2.1.4.3 Contextual Sentiment

We hypothesize that each writer might frame reported speech differently, thus we compute the sentiment of the text that frames a quotation and we expect that it will lead us to the respective newspapers. This attribute is used in our machine learning model later on to classify reported statements to the original news sources. The sentiment of a quotation itself could also be high[15,16], but also neutral in more factual statements[17,18]. However, we are not interested in the biased personal opinions or announcements of facts by politicians. We are focusing on the way that media cite them and predispose the reader in favour or against them.

Each quotation is contained in a longer sentence in the article, with the remaining

[15]"George Eustice believes that Cameron and Osborne are mistaken if they believe the Scottish referendum playbook"

[16]"Cameron said the UK must give France *no excuse* to tear up the treaty."

[17]"Jeremy Corbyn says he will be on ballot"

[18]"David Davis has said the Government will introduce a Bill to Parliament to begin the legal process of Brexit within days, with MPs likely to vote on the legislation as soon as next week."

words (including the cue verb and the speaker) outside the quote being located either before or after the quotation. We consider these words as the immediate context/frame of a reported statement and we compute the sentiment for this text. We build a fastText text classification model [72] to identify sentiment with the Stanford Sentiment Treebank dataset [152], following the literature that shows promising performance of fasttext on sentiment analysis [64, 72, 112]. Fasttext is a fast scalable shallow neural network for text classification that is based on n-grams, and can work well with small document collections. The model is trained with the labeled phrases of the dataset instead of sentences [59], since quotations as well as their context can include more than one sentence. The original number of phrases is 239,232 with labels from 0.0 to 1.0. There are five classes that can be inferred: very negative [0, 0.2], negative (0.2, 0.4], neutral (0.4, 0.6], positive (0.6, 0.8] and very positive (0.8, 1.0]. We filter out the neutral texts (50% of the data) and use the rest to train a binary classifier that detects negative (class=0) and positive (class=1) sentiment of an input phrase. We apply the model and annotate the context of quotes, resulting to approximately 55% of the quotes in our training set having negative contextual sentiment and the rest positive.

Table 2.5: Motivational examples with different representations of quotations in different newspapers.

Guardian Tory MP Michael Fabricant, shadow minister for industry and technology, told PA: "It is quite clear that they are planning to mount a desperate dirty tricks campaign."	**Telegraph** Michael Fabricant, the Tory MP who discovered the sites, said: "It is quite clear that the Labour Party is planning to mount a desperate dirty tricks campaign. **How desperate they have now become.**"
Independent Both Jeremy Corbyn and the SNP let rip at the Prime Minister after the Foreign Secretary told a Czech newspaper that "Britain will probably leave the EU's customs union."	**Telegraph** He said "We probably will have to come out of the customs union, **but that's a question I am sure will be discussed.**"
Telegraph Boris Johnson says the "Brexit transition must last not a second more than two years" as **he laid out a series of challenges to Theresa May on the eve of the Conservative party conference**.	**Independent** Philip Hammond also slapped down the Foreign Secretary's insistence that "a transitional period must last not a second more than two years," **calling it a rhetorical flourish**.
Daily Mail Sources told a German newspaper that **the meeting had been a disaster** and that Mr Juncker told Mrs May "I'm leaving Downing Street 10 times more sceptical than I was before."	**Independent** While Ms May's **officials described the face-to-face as "constructive"** just after it ended, Mr Juncker is reported to have said "I leave Downing Street 10 times more sceptical than I was before."
Guardian Desmond Swayne MP asked about the "new lovefest with the benches opposite", which he suggested "**it was akin to making a deal with the devil**". "Given the record of the leader of the opposition on the Counter-Terrorism and Security Act does she possess a long spoon?"	**Independent** Desmond Swayne, a former minister, criticised "this new lovefest with the benches opposite", **urging the Prime Minister to approach Jeremy Corbyn with a very long spoon**.

2.1.5 Bias Detection in Reported Speech

In this section, we describe how we classify reported speech statements to their respective news source, as a way to show how predictable media can be. Hence, we model the bias detection problem in media as a multi-class classification task. Table 2.5 demonstrates motivational examples for our study, where the differences in the presentation of the quotations are pointed out. Each row shows how the same utterance is reported in two different newspapers. In the first two rows, the same

statement is reported partially by the first source (left) and specifically in the second row, the frame of the quote in the Independent is opinionated (“SNP let rip at the Prime Minister"). In the third row, the two outlets report about Boris Johnson’s transition statement and after this, they add more but different information about it. The example in the fourth row shows a very different context for the same utterance, namely one outlet calls a political meeting “disaster" and the other “constructive". Moreover in the last row, the newspapers report different segments of the utterances of Desmond Swayne and also with different framing verbs (“asked", “suggested" in contrast to “criticised", “urging").

Our goal is to understand which characteristics the journalists base their decisions on, e.g., the content of an announcement, the party it originates from, the speaker etc. Our motivation stems from that the more feasible it is to predict the origin of a political quotation (based on its characteristics), the stronger the political reporting pattern of the respective newspaper is expressed. To this end, we propose several features of the quotes that could help us distinguish where a quote comes from. They are inspired by the various ways media bias is expressed in the news (as discussed in Section 2.1.1), especially in the context of reported speech.

We define groups of features in Sections 2.1.5.1 to 2.1.5.4 and each set is given to a machine learning model with the same architecture that classifies quotes. For every input feature, we compute its numerical representation with one-hot encoding of dimension 500. In each experiment, we compute the average vector of all relevant individual features to one unified vector (e.g., as discussed in Section 2.1.5.4, we average the vectors of the quote text, the cue verb and the speaker’s party to train a classifier that detects partisan bias in reported speech). In addition, we standardize our input features to compensate that their values are in different scales and to ensure the best possible performance of our classifier. In Section 2.1.6 we describe how we build our classification model.

2.1.5.1 Selection Bias

As discussed in Section 2.1.1, there are several ways that bias is conveyed in a news article. The primary bias kind is selection bias, i.e., whether a news piece (in our case a politician’s statement) will be included in the news or not. In the context of reported speech, we aim to discover whether the content of a statement is a determining factor for the media to cite this statement. In order to capture this basic decision, we consider the **quotation text** as feature of our method. The quote text is represented with Glove word embeddings of dimension 50 [125], as

opposed to our previous technique with a bag-of-words representation [84]. We consider the maximum word number in each reported statement to be 500, since quotations can be paragraphs with multiple sentences (a sentence typically consists of maximum 20-30 words), but they are also shorter than full documents. This is our baseline approach, which naively considers that the quotation text is the only deciding factor for whether a quotation by a politician will be published or not. Since each political news report contains several quotations in its text[12] and this method will not consider any meta-information about the quote as assistance (e.g., cue verb, political party of the speaker, etc.), we consider this baseline simple and bias-unaware.

2.1.5.2 Coverage Bias

This type of bias refers to the extent that a story is covered by the media. In journalism, during the fact collection process, reporters are called to decide among different politicians that comment on an event who to quote in their article. Our goal is to find out whether the politicians' statements and their political affiliations are decisive factors for the journalists and if so, to which extent. Our motivation stems from our statistical analysis in Figures 2.4, 2.6 and 2.7, where politicians are sometimes mentioned and quoted disproportionately in each newspaper. Thus, we wish to leverage the discrimination that some news sources might show towards different political figures and parties. In our preliminary study [84], we used simple coverage bias metrics in the context of reported speech, i.e., the article length and quote length. In our latest contribution [86], we consider a more concrete signal for coverage bias in political news, namely we use two feature combinations: a) **the quotation text combined with the quote speaker**, and b) **the quotation text combined with the party** that the speaker is affiliated with. The speaker and the party are categorical features that are transformed into numerical ones with one-hot encodings of dimension 500.

2.1.5.3 Framing Bias

Our goal at this step is to discover which contextual information around a quotation is the most appropriate for our task, that is the most indicative of a news source. Newspapers write with different styles and frame the news with different views. We wish to discover which kind of context each outlet is using when citing politicians. In this way we can distinguish between them and understand their patterns. We consider as context (frame) of a reported statement the sentence that contains the

utterance, i.e., the remaining words before or after the quote. We perform three different experiments with three different feature combinations to show framing bias. Initially, we classify using **the quote text and context** together. The numerical representation of the context is the same as the quote text's (described in Section 2.1.5.1). In this step, we aim to see whether additional contextual information will improve our baseline approach. In other words, does the context of a quote depend of the content of the utterance? Would quotes with different topics, sentiments, viewpoints, etc. provoke different context in the media's citations? Next, we classify using **the quote text and the sentiment** around the quote. The sentiment is a binary feature that is one-hot encoded, similarly to the above-mentioned features. This experiment will show us whether there are indeed differences in the emotions conveyed in the quote context in every newspaper. We also use **the quote text together with the introductory verb** of the quote, because the cue verbs are a key tool of the article author to predispose the reader about the following reported utterance. The cue verb is a categorical feature of the quotes that is one-hot encoded as mentioned above in the introduction of Section 2.1.5. We perform our previous experiment as well [84], i.e., classifying the quotes only by leveraging their context without considering the quote text or any other quote attribute. Note that the surrounding of a quotation contains the introductory verb and the speaker's name, but also other possible bias indicators, such as adverbs (e.g., "announced proudly", "stated provocatively", "said arrogantly", etc.).

2.1.5.4 Partisan Bias

We perform a separate experiment to discover whether different newspapers treat certain parties differently when it comes to reporting their political statements. We combine **the quote text with the cue verb and the political party** that the quote speaker is affiliated with. In this way, the potential cases where a news source is systematically promoting or discriminating against a certain political party will be revealed. For instance, one example could be if a newspaper is constantly using negative introductory verbs to frame the reported speech by the *labour* party, and positive ones for the *green* party.

2.1.6 Classification Method

We build a feed forward neural network to classify quotations to their respective news source with the Keras library [27] for our experiments [86] – in our previous work [84] we use a Random Forest classifier [18] with the Weka library [174].

Table 2.6: Number of quotations per newspaper in the training and test set of our quotation classification task.

	Guardian	**Telegraph**	**Independent**	**Daily Mail**	**Total**
Training set	18,260	5,327	17,446	3,352	45,489
Test set	4,652	1,359	4,482	882	11,375

We refrain using traditional machine learning algorithms in our latest work [86], because deep learning approaches have shown the potential to perform better in complex natural language processing tasks. The network contains a pre-trained embedding layer at the beginning followed by two hidden layers (32 and 16 neurons respectively) and then a softmax activation layer with four outputs that correspond to our classes (the Guardian, the Telegraph, the Indepedent and the Daily Mail). We use the categorical cross entropy as loss function and the Adam optimizer [76]. Our model is trained with batch size 16 and for 200 epochs. We have experimented with a wider/deeper network and higher batch size, but the current setting achieved the best results.

Evaluation: We train our model with eight different input types: firstly only with the quote text (our baseline approach) and then with the feature combinations described in Sections 2.1.5.1 to 2.1.5.4 that tackle specific kinds of media bias in reported speech. For every newspaper, we use 80% of its quotations to train and the rest 20% is used for testing. Thus, our final training and test sets consist of the individual sets for each news outlet combined (as shown in Table 2.6). We also shuffle the data in advance and use 10% of the training set as validation data. Due to the underlying class imbalance in our dataset, we oversample quotations from the underrepresented newspapers (the Telegraph and the Daily Mail) in our training set – we prefer this to undersampling so that we do not miss any important patterns in our data. For each quotation in these two news outlets, we create two additional copies of the original one and include them in the training set, in order for the two minority classes to contain approximately the same number of data after the oversampling process. We evaluate the performance of our models using the micro-average F-1 score (class-weighted harmonic mean between recall and precision), because in this way we aggregate the contributions of all classes to compute the average metric. This measure is preferable to macro-average, which gives equal weight to each class regardless of size.

Table 2.7: Micro-averaged classification results for all newspapers using different input feature groups for focused bias detection. *Text* is the content of quote, the *speaker* is the politician that is being quoted and the *party* is his/her affiliation. The *verb* is the cue verb that a quote is being introduced. The *context* consists of the remaining words in the sentence that contains a quote.

Bias kind	**Input data**	**Micro Avg F-1**
Selection	Text	0.33
Coverage	Text + speaker	0.35
	Text + party	0.34
Framing	Text + context	0.46
	Text + verb	0.35
	Text + contextual sentiment	0.34
	Context	**0.54**
Partisan	Text + verb + party	0.33

2.1.7 Results

Table 2.7 presents our first quote classification results for all newspapers. Every row in the table corresponds to an experiment with different input features and the same neural network architecture. The experiments are grouped based on the kind of bias they help us reveal. We can observe that in most cases adding more information along with the quote text, either improves or at the minimum maintains our ability to predict the origin of a reported statement. For instance, **combining the quotation text and its surrounding context enhances our performance** significantly in comparison to when we take into account only the quote content (from 33% to 46%). However, combining only a single categorical feature (the quote speaker, the cue verb or the speaker's party) with the quote text yields the same outcome as when using only the quote text. This indicates that in general reported speech features and metadata are not as useful for this task as the overall words around a quotation can be.

A very interesting outcome is that using the quote context as input we can achieve better results than combining it with the quote text (54% in contrast to 46%). Considering that this is 4-class classification task, 54% F-1 score improves significantly upon a random decision. This model focuses only on the journalist's words rather than the politician's words. This indicates that **the frame that each newspaper describes a quote can vary substantially** among the news outlets, regardless of the reported statements. That is, a quote does not necessarily need to be opinionated or subjective to provoke a opinionated frame in return. For instance, Donald Trump's announcement of the "travel ban", which restricted citizens of seven countries to en-

ter the United States, has caused many controversial reactions in the media, but that does not make the citation “Donald Trump announces new US travel restrictions" itself politically toxic[19]. The Financial Times has characterised this announcement as “his latest attempt to fulfil a campaign pledge"[20] and the Guardian described the ban as “contentious and chaotic"[21]. Thus, it appears that **the context can be more helpful than the actual quote content** and that it can assist the learner to understand the reporting patterns of each newspaper in a more effective way.

It is also notable that the addition of the contextual sentiment did not improve the current results. On the one side, this is surprising, since we have already confirmed that the context itself is very helpful for the classifier, both in Table 2.7 and in our previous work [84], and thus we assumed that its sentiment will be a meaningful addition. On the other side, it is commonly known that traditional news articles do not contain as many opinionated and sentimental terms as user generated content can contain. However, we were hypothesizing that in the rare cases that a journalist uses loaded language in a news article, it would be in the context of reported speech, as a way of argumentative or speculative discourse [12, 149]. On the other side, the amount of the somewhat opinionated words out of all words in a quote’s context can be very low. For instance, on the topic of the conservative leadership election in 2016, the context “*George Osborne has entertained Gove and Vine at Dorneywood more than once even after Gove announced that*" introduces the quote “*the would be advocating Leave*" in an article in Telegraph. In the same article, the context “*For a man who has espoused unfashionable causes all his life – Scottish Thatcherism , the Iraq war and Brexit – Gove now hopes that"* frames the quote “*the unfashionable candidate is going to surprise them all once more"*. **There are few sentimental** [22] **terms** in this texts (e.g., unfashionable, entertained), but they are very few in comparison to the rest of the words in the quote surroundings. Note that this particular article contains other opinionated words as well further away in the text from these quotations (e.g., “*Michael Gove seemed wholly out of place in modern Britain*").

[19] https://www.theguardian.com/politics/2016/jul/31/how-did-the-language-of-politics-get-so-toxic

[20] https://www.ft.com/content/fe65520c-a194-11e7-9e4f-7f5e6a7c98a2

[21] https://www.theguardian.com/us-news/2018/jun/26/muslim-americans-trump-travel-ban

[22] http://nlp.stanford.edu:8080/sentiment/rntnDemo.html

2.1.7.1 Newspaper predictability

Since the above-mentioned results are averaged over all newspapers in our dataset, it is yet unclear what the individual patterns are of every newspaper. In order to discover the bias types that each news source potentially exhibits, we show our classification performance for every newspaper in our collection in Table 2.8. It is

Table 2.8: Classification results per newspaper using different input feature groups for focused bias detection. *Text* is the content of quote, the *speaker* is the politician that is being quoted and the *party* is his/her affiliation. The *verb* is the cue verb that a quote is being introduced. The *context* consists of the remaining words in the sentence that contains a quote and *sentiment* refers to the sentiment around the reported statement.

Input data	**Guardian**	**Telegraph**	**Independent**	**Daily Mail**
	15,577 articles	*4,956 articles*	*10,712 articles*	*1,415 articles*
	P R F-1	P R F-1	P R F-1	P R F-1
(*selection bias*)				
Text	0.43 0.49 0.46	0.13 0.24 0.17	0.42 0.26 0.32	0.09 0.09 0.09
(*coverage bias*)				
Text + speaker	0.43 0.50 0.46	0.12 0.22 0.16	0.42 0.30 0.35	0.10 0.02 0.03
Text + party	0.43 0.50 0.46	0.13 0.25 0.17	0.42 0.28 0.34	0.08 0.02 0.03
(*framing bias*)				
Text + context	0.50 0.58 0.53	0.16 0.24 0.19	0.61 0.45 0.52	**0.53 0.34 0.41**
Text + verb	0.43 0.51 0.47	0.13 0.25 0.17	0.42 0.28 0.33	0.11 0.02 0.04
Text + sentiment	0.43 0.47 0.45	0.12 0.27 0.17	0.42 0.28 0.34	0.12 0.03 0.05
Context	**0.61 0.74 0.67**	**0.20 0.33 0.25**	**0.66 0.47 0.55**	0.41 0.16 0.23
(*partisan bias*)				
Text + verb + party	0.42 0.47 0.45	0.12 0.25 0.16	0.41 0.28 0.33	0.12 0.03 0.04

evident that the underlying class imbalance in our dataset is affecting the results. For instance, the lack of access to premium articles in the Telegraph and also the lack of a sufficient amount of political articles in Daily Mail prevent the model to learn the patterns of these two newspapers as efficiently as in the Guardian and the Independent. This phenomenon occurs also in our earlier time-evolving study [84], where our performance exhibits peaks during the years that general elections were held in the UK (2001, 2005, 2010 and 2015), partially because there were more news articles published in these periods. This effect could be also explained if we consider the elections as an opportunity of the media to deviate from each other. That is, prior to crucial political events, media potentially make their endorsements more obvious and thus differentiate from each other in terms of the news they report. In addition, it is worth noting that **selection bias does not manifest as much as the other bias kinds** in all newspapers. This is on par with related work [136], especially

considering that the four newspapers we analyze belong to the same geographical region and hence they have similar publishing interests.

Additionally, the classification results show that coverage bias is also not prominent in our data in the context of reported speech. Note that we have already seen in Figure 2.4 that the least popular political parties are quoted disproportionately in the newspapers, e.g., the Scottish national party is cited almost exclusively by the Telegraph. This signifies that the reporting choices of the outlets likely depend on the popularity of the parties, the frequency of their public announcements, the announcement time, etc., and not necessarily on the information that the parties publish and the speakers themselves. Thus, it appears that the combination of the quote text and the speaker or the speaker's party is not a criterion for the newspapers to decide whether they report a statement or not.

The most promising results are given by our framing bias detectors. In addition, the Guardian and the Independent seem to be the most distinct newspapers so far, because we can reach the highest precision for them (61% and 66% respectively). Both of these precision values are achieved by our model solely based on the quotation context. We hypothesize that the Guardian might be the most distinguishable news source among all, because the achieved recall values of almost every bias detector are significantly higher than the ones of the other outlets. This is in line with our previous work [84], where our comparative results for the Guardian are better than for the Telegraph with our bias-aware model that combines all features– note that we use accuracy as a classification metric in this case, due to balanced class distributions. Surprisingly, the best results for the Daily Mail are shown when both the quote text and context are used. This is not true for the rest of the newspapers and it could signify the correlation between the reported statement and its frame for this specific news outlet. That is, the frame the utterances are described with might depend more on the quote content than the newspaper's or journalist's political beliefs or writing style.

Moreover, we hypothesize that the existence of a high amount of false positives could be justified, because newspapers have oftentimes common vocabulary in their narratives and mis-classifying a quote to the wrong newspaper is neither surprising nor of high risk. However, very low recall (e.g., in Telegraph and Daily Mail) signifies that the model is truly suffering to distinguish a newspaper. We hypothesize that apart from the small data size, the main reason for this is that there is no systematic quoting pattern for our model to learn. This is a positive outcome for the respective newspapers, signifying that they are not as predictable,

and potentially as biased, as others. However, it is also notable that there is not much variety in the data to assist the model to learn the hidden patterns. Regarding the small amount of Brexit related articles in the Telegraph, our observation is on par with recent related work that discovers that the Telegraph did not cover the Brexit news as thoroughly as the Guardian did [109].

2.1.7.2 Newspaper and party correlations

An additional aspect that we are interested in is the extent of the potential media bias towards different political parties. Thus, we investigate the classification precision of our quote classifier based on the quote context for the four most discussed parties in our UK news corpus. Our various framing bias detectors in the context of reported speech are superior to the rest of the approaches. Hence, for this party-focused experiment, we use the context-based model that takes into account all the words in the immediate quote surrounding to decide the news source that the quote belongs to. In Figure 2.8, we observe that **every party is treated by each newspaper differently**, e.g., in the case of the *SNP* the Telegraph has more consistent reporting patterns than the other outlets have, and the Daily Mail seems to have no patterns at all when citing politicians from the *SNP*. Note that the performance of the model is close to zero for the Telegraph and the liberal democrats, and also for the Daily Mail and the *SNP*, because the newspapers published almost no quotation from these parties, respectively.

In Figure 2.8(a), we see that the Guardian is the most predictable in the way it reports quotes by the *labour* party and the *liberal democrats*. Overall, the political party that is framed in the most characteristic way across multiple news sources is the *conservative* party for the Daily Mail, the Independent and the Guardian, and the quotes from *SNP* are more predictable when the Telegraph reports them. We also observe that the article lack in the Telegraph does not prevent us to achieve good results for the *SNP* (60% precision) and in the case of the few seats of the Liberal Democrats (20 MPs), our model can still conclude that the newspaper which reports about them in most particular way (over 70% precision value) is the Guardian. Thus, even though one can assume that the classification performance could correlate with the amount of articles in each outlet (depending on the newspaper popularity, accessibility and media coverage), we conclude that this inherent nature of the data does not appear to skew our results.

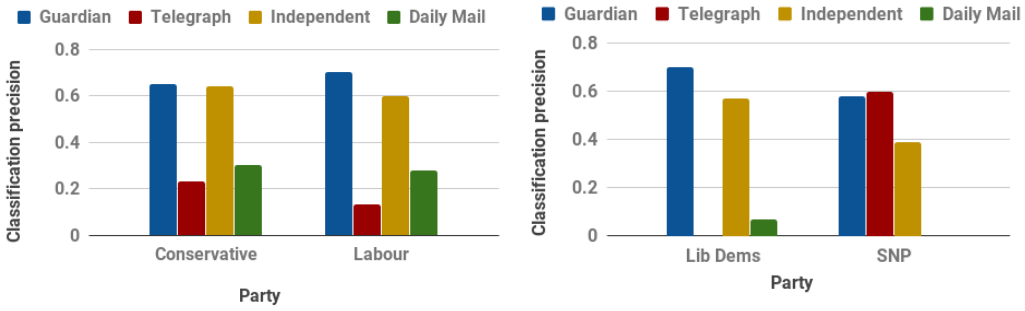


Figure 2.8: Classification precision individually for the most discussed political parties.

Table 2.9: Classification results (using two framing bias detectors) for an additional test set. This set consists of the quotes by the same speaker that appear in more than one newspaper in our corpus. The two models detect framing bias with two different feature sets.

Input	**Guardian**	**Telegraph**	**Independent**	**Daily Mail**
	P R F-1	P R F-1	P R F-1	P R F-1
Text + context	0.42 0.60 0.49	0.17 0.31 0.22	0.70 0.44 0.54	**0.72 0.29 0.42**
Context	**0.57 0.78 0.66**	**0.17 0.34 0.23**	**0.74 0.48 0.58**	0.53 0.18 0.27

2.1.7.3 Newspaper predictability for widely published quotes

In a real-world scenario like ours, some quotations by certain speakers are usually repeated in different news articles by different media – the content that is cited is the same and by the same speaker, but the context can differ in each newspaper. We consider duplicate quotations to be statements that contain the same text and the speaker of the utterance is also the same. These are specifically interesting cases, because they enable us to examine in a more straightforward way the relative differences among news media, given a common political statement. As previously mentioned, approximately 2-3% of our data contains duplicate quotations, which were removed before performing our experiments. A very small set of them contains quotes by the same speaker that appear in more than one news outlet, i.e., 1,099 quotations. We apply two of our framing bias detectors on this dataset and show our classification results for this new test set in Table 2.9. Even though, the micro-average F-1 score over all news outlets remains approximately the same as with our default test set (53%), it is interesting to observe the performance improvement for the Independent and the Daily Mail. Namely, the precision is almost 10% and 20% higher, respectively. This outcome implies that **the newspapers can be more predictable for certain quotations and less for others**, especially when the quotes

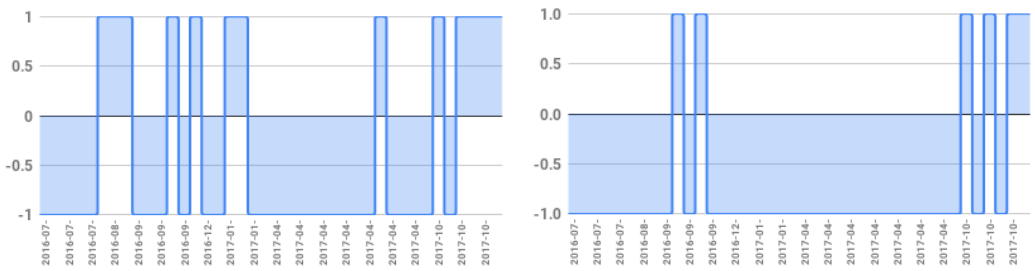


Figure 2.9: Sentiment values in the Guardian (left) and the Telegraph (right) around the common utterances of Theresa May that are reported in all newspapers from 2016 until 2017. 1 stands for positive and -1 for negative score.

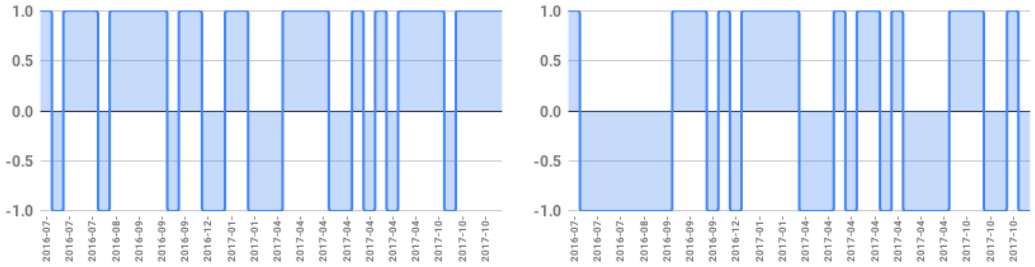


Figure 2.10: Sentiment values in the Independent (left) and and the Daily Mail (right) around the common utterances of Theresa May that are reported in all newspapers from 2016 until 2017. 1 stands for positive and -1 for negative score.

are reported by multiple newspapers – and thus they are potentially about important or breaking news.

Regarding the Guardian and the Telegraph, even though they are not significantly more predictable for this test set, they write about Theresa May's statements negatively more often than not. We shed more light into this experiment by visualizing the contextual sentiment around the quotations by Theresa May that are reported in all four news outlets. As shown in Figures 2.9 and 2.10, as time evolves from the left to the right of the horizontal axis, **each newspaper reports Theresa May's statements with a different frame** or viewpoint. Specifically, the Telegraph puts her quotations in a negative context for almost one year starting in September 2016, when she began her first major public statements about Brexit. The Independent and the Daily Mail do not show an obvious pattern in their sentiment, and the Guardian is more negative in 2017 than in 2016. Overall, given this set of quotations, the Guardian frames positively 32% of them (and the rest 68% negatively), the Telegraph 15%, the Independent 67% and the Daily Mail 45%. Hence, we conclude

that even though the Independent and the Daily Mail are more predictable (in terms of precision) in this test set, their contextual sentiment is relatively balanced for the quotes of Theresa May.

2.1.8 Summary and Findings

In this section, we summarize our methods and discoveries, and we discuss future work ideas later on in Section 2.3, collectively for all contributions of this chapter. In this line of work, we study political bias manifestations in reported speech, specifically in the news domain. Our hypothesis is that different newspapers cite and frame politicians' statements in different ways, potentially conveying their own political agendas. Chronologically, we first obtained initial insights of the problem [83], we contributed to media bias detection in reported speech via quote classification next [84], and lastly we extended our proposed approach by considering additional news corpora and features [86]. Moreover, we formulate our automatic unsupervised approach to detect the speaker of a reported text based on the context of the statement and the overall news article that the statement appears in. We extract indirect and direct reported speech, and propose groups of the quotation features that correspond to specific bias kinds. For instance, analyzing a quotation's context gives us evidence about the frames and lenses that news sources describe the news with. Note that our focus in not the information extraction process itself, e.g., extracting quotations, their speakers, the sentiment etc., but rather, given this information to analyze the patterns that it exhibits in each newspaper. We classify the quotes to their original newspaper with a neural network as a means to discover how predictable and unique each newspaper is in the way they cite politicians. We argue that when our model performs well with a specific feature as input, then this feature (and by extension bias type) is more apparent in a given newspaper.

Our findings include that features that are based on news framing are more powerful than the rest, which indicates the existence of framing bias in reported speech in UK media. Moreover, the context around politicians' utterances in the news varies varies significantly among the outlets. The reporting patterns of the newspapers for different parties also varies, with certain outlets being more predictable when discussing a political party than others. In addition, we present focused results for each news source and discuss which features are more relevant for every newspaper. We discover that the Guardian and the Independent are more predictable than the rest. Interestingly, for the Daily Mail the frames of the quotes depend more on the quote content itself than for the other outlets. In addition, we discover party and

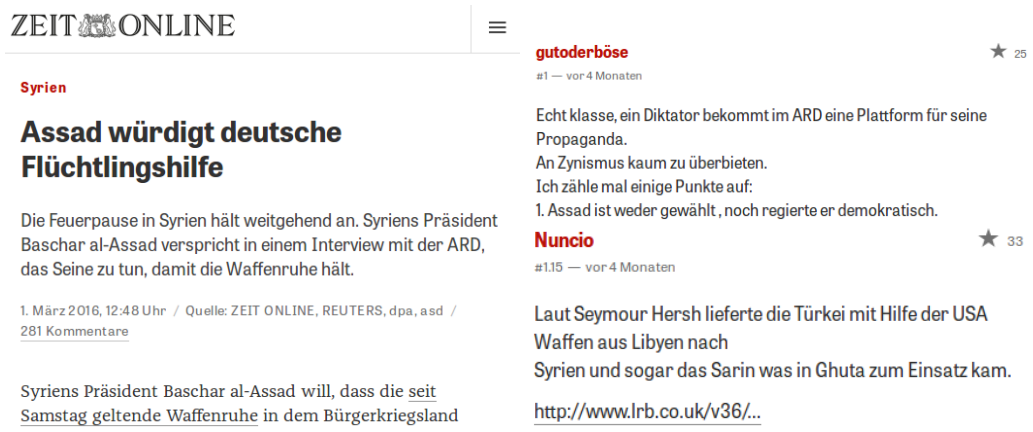
newspaper correlations, e.g., the Telegraph presents the quotes by *SNP* in the most distinguishable way, and the Guardian the ones by the *liberal democrats* respectively. We also perform a more focused classification experiment where we consider the common quotes that belong to multiple newspapers and we observe improvement in the results for the Independent and the DailyMail. That could signify that the outlets are even more discernible given a popular or important statement by a prominent speaker regarding breaking current news. In this setting, we also investigate further the quotes by Theresa May that are reported in all four news sources and we show that the frames' sentiment varies significantly. Namely, the Telegraph reports in the most negative and the Independent in the most positive way the statements of Theresa May that are published by all four news outlets.

2.2 Classifying News Comments

Online news has gradually become an inherent part of many people's every day life, with the media enabling a social and interactive consumption of news as well. Readers openly express their perspectives and emotions for a current event by commenting news articles. They also form online communities and interact with each other by replying to other users' comments. Due to their active and significant role in the diffusion of information, automatically gaining insights of these comments' content is an interesting task. We are especially interested in finding systematic differences among the user comments from different newspapers. By finding patterns in the different viewpoints of the commenters, we can also infer conclusions about the newspapers' beliefs and the kind of users they attract. To this end, we propose the following classification task: Given a news comment thread of a particular article, identify the newspaper it comes from [50]. Our corpus consists of six well-known German newspapers and their comments. We propose two experimental settings using SVM classifiers build on comment- and article-based features. We achieve precision of up to 90% for individual newspapers.

2.2.1 Online Comment Sections

Many online news sites offer their readers the possibility to comment on news articles either directly below the article in a forum-style way, or via Twitter or Facebook. While the latter is more suitable for sharing news, the former is more appropriate for discussion of the articles' contents. These online comments are huge reservoirs of user generated content with readers expressing opinions on various news-related topics. These range from comments on the article's style, specific arguments of the article, to general opinions about greater questions. Note that processing and



(a) Excerpt of article

(c) Excerpt of reply

Figure 2.11: Example of an article, comment, and reply from the newspaper “Zeit”.

evaluating the quality of comments is also a challenging task in terms of the amount of data and the uncivil content they might contain [106].

Not only does the discussion in these comment sections often reflect the readers’ opinions about the article itself, but also about the overall topic and beyond, with readers referring to each other or introducing new arguments. Figure 2.11 shows excerpts of an article together with a comment and a reply to this comment. In general, the discussions are not limited to the specific article’s topic and often introduce new arguments and opinions. Sentiments are expressed as well, towards either the content of the article or statements of other users. The content of one individual comment is not easily machine-understandable. It needs to be evaluated in the context of the surrounding thread and associated article. Nevertheless, we argue that discussion style and topics may differ between various news providers, depending on their respective audience and possibly bias in the article’s coverage. For example, German newspapers and the majority of their readers are traditionally associated with a certain political alignment. If this is true, the political leaning should be reflected in the comment sections of the respective news sites as well. Even if the bias in the articles themselves is minimal, the reaction of the readers to the covered event may be much more diverse, which in return could be used to infer arguments for the political alignment of the news sites.

In this work, we analyze the user comments on six major German news sites regarding their differences in discussion focus, language and sentiment. Based on the assumption that user comments on various news sites differ in these characteristics, we propose a classifier to predict the source of specific comments, that is, the

news site on which the comments have been posted. To analyze this, a prediction method is developed and evaluated, which, given a set of user comments, predicts the originating news site.

2.2.2 Related Work

User comments can be found in different online platforms and communities. Social media platforms, such as Twitter, Facebook, and Youtube, are the most popular environment for users to generate personal content, share pieces of news, build social relations etc. Recent research focuses on analyzing comments' content on these platforms, as well as analyzing the commenters. An extensive analysis [147] of comments in social media communities investigates comments' sentiment, rating and popularity in Youtube videos and Yahoo! News posts. Momeni and Sageder [107] perform a comparative analysis of comments in Flickr and Youtube. The authors point out different textual, semantic, and topical features of the comments, which are later used to predict the comment's usefulness. Towards identifying the characteristics of influential users, Martin et al. [102] introduce an emotion lexicon-based technique that predicts the helpfulness of reviews posted on Trip Advisor and Yelp.

In addition to social media, related research focuses on news media as well. Here, understanding and potentially predicting the user characteristics and preferences is the main goal. The problem of user profiling in news media is tackled by introducing the notion of *comment-worthy* news articles[8]. The authors predict the comments' interestingness in blogs and news sites using an adapted topic model aiming at personalized recommendation of news articles to users. Similarly, Shmueli et al. [146] address the problem of ranking news comments according to the reader's personal interests in Yahoo! News using a factor model. Instead of analyzing existing comments, Cao et al. [20] extract relevant microblog posts to news articles and use them to automatically generate user comments for these news articles.

Moreover, since users shape the general public's opinion with their comments by often supplementing the news stories with new facts and expertise, approaches that automatically evaluate the comments' quality have received high interest in the literature. To this end, tools distinguishing the (in)appropriate and (ir)relevant comments could assist media to improve the news quality they offer. Related work includes the analysis of the quality of comments [40], and the measurement of the comment sentiment in order to conclude about the media's political leaning [119]. Additionally, the problem of comment relevance is also addressed [33, 41, 107], with the latter assessing the degree of pertinence of comments by comparing their

tf-idf vectors to the articles' in News York Times. Detecting the comments that shift the main article topic and change the article's focus at Digg.com is tackled by Wang et al. [168], while Zhang and Setty [179] identify sets of topic-wise diverse user comments in Reddit news articles. Recent research focuses not only on the comment moderation, but also on identifying how user comments can be helpful to journalists [97]. The authors develop a framework for the newspapers to analyze their user comments. They discover that classifying toxic content is as important as understanding the crowd's perspectives, which user agrees/disagrees with which article, etc.

Finally, multiple interesting prediction tasks emerge from news comments analysis. Among others, the volume of news comments is predicted with a random forest classifier by Tsagkias et al. [158] using a variety of comment and article metadata, as well as textual and semantic features derived from the comments. A recent journalism study [79] analyzes user comments from various aspects, in order to understand user engagement and how journalists can promote it. Rizos et al. predict news stories popularity based on users' comments and the properties of the social graph they form [135]. Since users abuse the commenting mechanism frequently by stating offensive or hate comments, Kant et al. [73] compare an SVM classifier to a pattern mining approach in order to detect spam comments in Yahoo! News articles. In addition, a novel approach for toxic comment classification is proposed by Aken at al. [162], where an ensemble of deep learning techniques classifies inappropriate user comments in Wikipedia and tweets.

In contrast to the above works, we analyze comments to investigate differences in readership and bias among different German newspapers. Automatically gaining insights in the huge amount of user-generated content in media will help us discover people's opinion over several issues. More specifically, the way readers perceive reality regularly depends on the different writing styles of different news outlets and their respective journalists. For instance, it would be interesting to discover that users tend to leave more informative or insightful comments, when a newspaper is being brief and doesn't discuss thoroughly certain topics. Alternatively, a user may post funny or hate comments, when an article criticizes openly a person or an event.

Furthermore, the ability to identify a comment's origin is a step towards detecting correlations between the news providers and the news consumers. We share the intuition of Park et al. [119] regarding media bias detection in news articles, that is, users tend to leave negative comments to articles that oppose their perspective and positive otherwise. Additionally, as introduced by Groseclose [52], readers often

Table 2.10: Characteristics of the articles and comments in six German newspapers.

Source	Articles	Articles w. ≥ 1 Comments	Comments	Average Comment Length	Articles w. ≥ 5 Comments
Bild	1,358	316	11,332	21.6	186
Focus	1,764	965	2,651	58.0	80
Welt	1,852	1,782	31,125	31.7	830
Spiegel	1,654	664	5,771	61.8	188
Zeit	1,045	1,032	8,553	46.1	642
Faz	1,656	458	1,329	71.3	61

choose to be informed by the sources that share their beliefs. Namely, one is more likely to perceive bias the further the slant of the news is from their own political position.

2.2.3 Predicting Comments' Original News Source

Our motivation stems from the idea that readers from different newspapers might use unique language and present different commenting patterns. There are indeed differences among users in news media in general: some users tend to be objective and include new facts to the articles, others leave subjective messages (e.g., supporting a party, an opinion), others may attack the journalist or comment writers with hate comments, etc. We are interested in whether the user writing style is indicative of the comments' source or not. Hence, we aim at identifying the comment features that distinguish the users of different news outlets. This will allow us to classify comment threads belonging to certain newspapers. To this end, all the direct comments and comment replies in a given article are considered as a single document in our prediction task. That is, one document is the complete news comment thread of a given news article. We then use an SVM classifier to classify each instance to its respective newspaper. The feature selection and the parameter setting are described below.

2.2.3.1 Datasets

We analyze six popular German newspapers, namely *Bild*, *Focus*, *Welt*, *Spiegel*, *Zeit* and *Faz*, which all allow user comment sections. The dataset characteristics are shown in Table 2.10. We crawled political news articles from March 2016 until June 2016 from all six news sources. The fifth column depicts the average comment length for each source after removing stop words[23]. It appears that *Spiegel* and *Faz*

[23] https://sites.google.com/site/kevinbouge/stopwords-lists

readers tend to leave longer comments than users from other sources. Additionally, we also observe that *Bild* commenters could be characterized as more active in comparison to the rest of the outlets, as the average number of comments per article in *Bild* is higher than in the rest of the newspapers.

Although the number of articles does not vary significantly among the newspapers, we can observe that *Welt* is the outlet with the most comments and commented articles in total. In our experiments, after considering all articles having at least 1, 5 or 10 comments in separate configurations, we conclude that the threshold (H) of 5 yields the best precision results and thus we only report on results using this threshold. The last column in Table 2.10 represents the number of articles with at least 5 comments for each source.

2.2.3.2 Classification Features

This subsection describes the comment-based and article-based features that we use for our SVM classifier.

Number of Comments and Average Comment Length. The number of direct comments and comment replies are summed up representing the first dimension of the feature vector. In addition, the average comment length is calculated for each article after filtering out the terms that appear in our stop word list. As shown in Table 2.10, there are significant differences among the outlets regarding the volume of comments and their length. Hence, our intuition is that the above-mentioned features will constitute an important indicator for the respective news source.

Direct Comment/Reply Ratio and Distinct Authors. The next two features refer to the users, regarding their activity and commenting behavior. The ratio between the direct comments and the nested ones is a numerical indicator of how interactive the commenters are and whether discussions are initiated by them or not. For instance, as illustrated in Figure 2.12, *Zeit* and *Bild* appear to have a higher number of user discussions than the other sources.

Moreover, the distinct number of authors per article is interesting as well, as it informs us about the comment availability and potential diversity. Articles with multiple commenters should contain a variety of opinions and statements, in comparison to stories that do not attract high user interest. Figure 2.13 presents the news articles that are covered by certain numbers of commenters. That is, e.g., around 90% of *Bild* and *Faz* news articles would be covered, if the top-30 commenters were

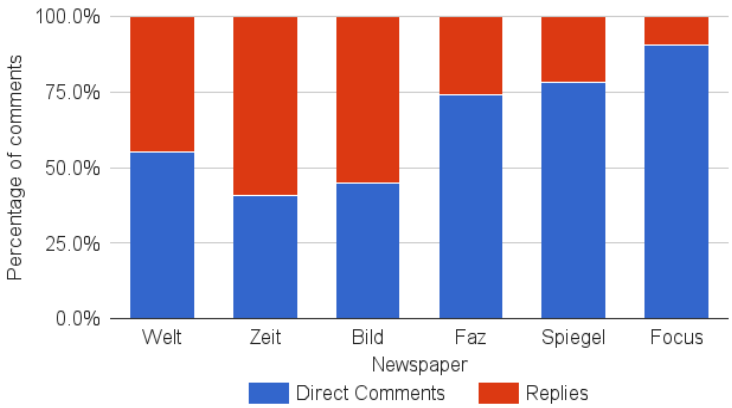


Figure 2.12: Direct comments and nested replies for six German news sources.

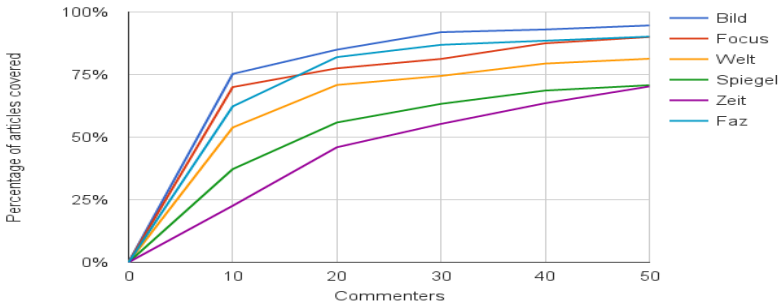


Figure 2.13: Percentage of articles covered by k user commenters.

considered. It should be also noted that for this plot we only use articles with H equal to 5. Our findings are in line with the work of Park et al. [119], where 50 commenters appear to cover around 80% of the overall dataset (when also considering solely articles with more than 5 comments).

Word Choices in Online Comments. The current feature corresponds to the comments' content as a bag of words, which potentially contains opinionated language. We argue that the kind of language used in the comments is the most representative feature of the users' perspective. Some comments aim at pointing out neglected facts from the articles and others might criticize the article's position or a politician's behavior, etc. Figure 2.11 illustrates an example of a comment in *Zeit* and one of its replies, where the two users express two different sides of the same story. It is notable that we only consider the terms' *tf-idf* scores that are not stopwords, since only these provide semantic and meaningful information about the users' interests.

Newspaper Uniqueness Metric. Apart from user features, newspapers' characteristics play a key-role to our prediction task as well. Towards discovering rep-

Table 2.11: Representative words of the topics discussed in German news media (ordered by descending popularity).

Topic Id	Frequent Terms
15	leben, politik, land, frage, deutschland, sagen, steht, kinder, sogar
0	prozent, deutschland, regierung, deutschen, zahl, land, praesident, frankreich, millionen
19	polizei, polizisten, frauen, demonstranten, koelner, maenner, verletzt, silvesternacht, koeln
7	euro, milliarden, deutschland, schaeuble, griechenland, geld, spd, gesetz, integration
3	spd, cdu, merkel, prozent, gabriel, afd, csu, seehofer, partei
5	russland, putin, usa, russischen, russische, praesident, obama, ukraine, nato
12	syrien, getoetet, stadt, waffenruhe, syrischen, terrormiliz, staat, aleppo, syrische
14	hofer, oesterreich, prozent, stimmen, fpoe, partei, wahl, parlament, van
4	afd, partei, deutschland, petry, islam, gruenen, cdu, kretschmann, npd
8	tuerkei, erdogan, boehmermann, tuerkischen, merkel, tuerkische, ankara, tayyip, recep
18	nordkorea, kim, journalisten, regierung, gericht, duendar, verurteilt, urteil, land
2	bruessel, anschlaegen, paris, anschlaege, flughafen, bruesseler, polizei, abdeslam, terroristen
13	panama, rousseff, papers, bundeswehr, zeitung, briefkastenfirmen, leyen, praesidentin, temer
6	cameron, khan, buergermeister, honecker, duterte, grossbritannien, london, johnson, britischen
16	trump, clinton, donald, sanders, republikaner, demokraten, hillary, cruz, vorwahlen
17	trump, clinton, sanders, donald, obama, hillary, prozent, cruz, trumps
10	the, waehler, and, twitter, primaries, staat, you, com, pic
1	trump, trumps, kasich, cruz, republikaner, senator, new, york, partei
11	fluechtlinge, tuerkei, griechenland, deutschland, grenze, fluechtlingskrise, migranten, fluechtlingen, europa
9	trump, sanders, clinton, cruz, rubio, donald, prozent, hillary, ted

resentative and specific language used by different newspapers, we measure the similarity between comments and news articles of all sources, in terms of their common words. We compare the comments' terms with the articles' terms from all sources and measure their *overlap coefficient*. That is, for each comment thread to be classified, we compute the *overlap* (or also known as Szymkiewicz-Simpson) *coefficient* between its terms and the overall vocabulary from the articles of each newspaper, which results in six separate numeric counts as individual features. Our intuition is that this metric indicates whether the journalists and the readers from a given newspaper mention the same words.

Since commenters are often subjective and emotional, the current feature might also contain words that are not expected to be found in news media. This word set is a possible bias indicator, considering that news articles are expected to publish objective and well-rounded news pieces, so that readers are adequately informed.

2.2.4 Topic Analysis

To ensure that all articles/comments are comparable across media outlets, we analyze the topics discussed in each news outlet.

As a first step towards understanding the discussions in our data, we are interested in detecting the topics mentioned in the newspapers' articles during our given time frame. For this purpose we use the latent Dirichlet allocation (LDA) implementation

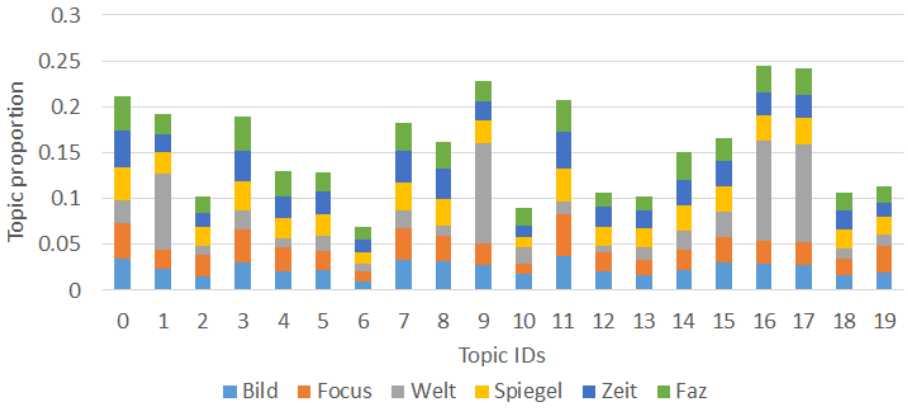


Figure 2.14: Topic popularity in six German newspapers.

in Mallet[24], a Machine Learning Java Toolkit. We experiment with different values for the number of topics, namely 10, 20 and 40, but report only our findings for 20 topics, since the results are rather stable with varying topic numbers. The current feature of an input news comment to our model is the topic distribution vector of the comment text, provided by LDA.

As shown in Table 2.11, the most discussed topics (15, 0) among all newspapers are focused on local affairs, with topic_0 touching upon financial issues. The least mentioned topics (9, 11, 1, 10, 17, 16) concentrate more on foreign politics, especially U.S. politics, which is an emerging topic as the general elections are approaching in the U.S.

In addition, Figure 2.14 presents the topic distributions across all newspapers. The x-axis represents the topics and the y-axis the volume of the discussion. One could infer that there are no extreme differences in the topic distributions among the outlets, that is, the same events/issues are covered by all newspapers. However, one notable exception are the comments in *Welt*, where the U.S. election topics (9,16,17) are clearly over represented. Our future work includes incorporating this topic-related information in the classification task and discovering whether it can improve our results, i.e., the users' commenting behavior differs for different combinations of topics and newspapers.

2.2.5 Results

The main goal of our work is to identify the newspaper that a certain comment thread comes from. Due to the small length of a single comment and the absence

[24] http://mallet.cs.umass.edu/

Table 2.12: Confusion matrix for one-vs-one news comment classification.

classified as →	**a**	**b**	**c**	**d**	**e**	**f**
a = Bild	19	0	0	0	1	0
b = Focus	0	8	0	0	5	7
c = Welt	2	0	12	0	5	1
d = Spiegel	1	2	2	11	3	1
e = Zeit	1	1	2	1	13	2
f = Faz	0	1	0	4	2	13

Table 2.13: Our news comment classification results.

Newspaper	**Precision**	**Recall**
Bild	0.82	0.95
Focus	0.66	0.40
Welt	0.75	0.60
Spiegel	0.68	0.55
Zeit	0.44	0.65
Faz	0.54	0.65
Average	0.65	0.63

(a) One-Versus-One

Newspaper	**Precision**	**Recall**
Bild	0.85	0.80
Focus	0.83	0.80
Welt	0.73	0.72
Spiegel	0.74	0.70
Zeit	0.80	0.75
Faz	0.90	0.90
Average	0.80	0.77

(b) One-Versus-All

of rich content, we classify all the comments for a given article at once, instead of considering them separately. For this purpose, we use the implementation of SVM classifier in Weka [174] with the default parameter settings.

Regarding the training phase, we initially perform one-versus-one classification, training m=k*(k-1)/2 classifiers (one for each pair of newspapers) and output the majority vote among all classifiers for each input instance. Namely, we train the model with 40 documents per source and tested it on 20 documents per source — all randomly selected from our original dataset. Our second experiment is a one-versus-all classifier that is trained and tested on articles from all outlets, but it performs binary classification for a single given source. In particular, six different classifiers are built (one for each outlet) using 40 articles from the target source and 40 random articles from the remaining sources. The test set consists of 20 articles from the target news outlet and 20 arbitrary ones from the other outlets.

The above numbers of articles are set after examining the last column of Table 2.10. The maximum possible numbers are considered, in order to obtain a sufficient and equal amount of comments per source in the training and test sets. Our future work includes obtaining more articles and subsequently more comments, fairly distributed to all six outlets, to achieve a higher comment quantity and diversity.

2.2.5.1 One-versus-one Classification.

The results of our first experiment are depicted in Table 2.13a and Table 2.13b. We can observe that the classifier performs best for *Bild* and yields inadequate results for *Focus* and *Zeit* with low recall or precision values respectively. The confusion matrix illustrated in Table 2.12 reveals that there is at least one comment from each source that is incorrectly classified as originating by *Zeit*. Considering that *Zeit* is the top-2 news outlet regarding the published number of articles with more than 5 comments, one might argue that highly popular and centrist newspapers, such as *Zeit*, contain a variety of comments and commenter behaviors. This makes such news sources a good candidate for an unseen comment, as they could contain a wide range of different commenting styles.

Additionally, *Bild* articles are largely classified successfully. According to Table 2.10, *Bild* is also one of the sources with the most overall comments, whereas the average comment length is relatively very low. Observing Table 2.10 and Table 2.12 concurrently, one can distinguish that when taking into account the most conservative sources, namely *Bild*, *Welt* and *Focus*, the lower the average comment length is the higher our precision result becomes. Since short user comments can often be sharp or toxic, this is an interesting observation for readers of newspapers in this spectrum. The average achieved precision is 65% and average recall 63%. Although the average performance score is a promising start, there is significant room for improvement, which we will further discuss in the following paragraph.

2.2.5.2 One-versus-all Classification.

Our next experiment is a one-versus-all classification. As previously mentioned, we build six different classifiers considering 40 articles from the target source and 40 random articles from the rest for the training set. The results are shown in Figure 2.13b. Surprisingly, although for the *Faz* articles the previous classifier achieved the worst results regarding precision, the current classifier performs best for this particular outlet. The overall results vary from 73% (*Welt*) to 90% (*Faz*) precision. Moreover, recall is significantly higher, ranging from 70% (*Spiegel*) to 90% (*Faz*). This leads to an average precision of 80% and an average recall of 77%.

2.2.6 Summary and Findings

In this work, we are interested in systematic differences among user comments in various newspapers. Our assumption is that the language news commenters are using on various news sites differs, e.g., based on the user's or newspaper's political

preferences. We hypothesize that the comment content differs enough to indicate the news source it is originally posted on. We leverage this distinctiveness of the news commenters to measure the distinctiveness of the newspapers in political news stories. Thus, we model the media distinctiveness problem via news comment analysis, i.e., we identify the original newspaper that a comment thread of a given article belongs to with our machine learning classifier. We develop an SVM classifier with different comment- and article-based features and consider six well-known German newspapers, namely *Bild*, *Focus*, *Welt*, *Spiegel*, *Zeit* and *Faz*. We discover that our best performance is achieved by six one-versus-one classifiers, one for each newspaper pair, where our precision scores range from 70% to 90%. This method (one-versus-all classification) is not as affected by our underlying class imbalance as the one-versus-one classification technique. Note that our distribution of our news articles is balanced between all sources, but not the comment distribution. Finally, we also observe a correlation between our achieved precision values and the average comment length in given newspapers.

2.3 Future Work

Our future work ideas in the area of media bias and distinctiveness in reported speech span different directions. Regarding our evaluation, instead of predicting whether a newspaper will report a utterance, one could predict how it will be described (extensively or briefly, in positive or negative context, with or without adverbs/adjectives in the context, etc.). Other improved approaches would be to use context-dependent word embedding techniques[37] to represent our input text in order to capture more meaningful and contextual insights of our news sources. In addition, another direction would be to utilize neural network models for text classification that apply attention mechanisms in the text, so that we discover additional signals about the hidden bias in the quotations context. Regarding our data collection, we believe that it is interesting to consider news articles from more countries, e.g., the USA, and test our hypotheses for other political scenes with different ongoing affairs. Especially for the USA and Barack Obama's statements, the related framework to our study QUOTUS [113] showed that liberal outlets are less likely to cite his quotes that are also reported by conservative outlets. It would be also interesting to examine whether there are dependencies among the newspapers themselves. For instance, if the Guardian publishes a quote by Boris Johnson, does this make it more probable that the Daily Mail will also cite this statement, and will it be reported in the same context?

Cross-newspaper topic-based analysis is also an interesting future work direction. Given a topic or a short time frame when some breaking news piece emerges, one could analyze the quotation distribution among the outlets from different perspectives. For instance, not all American sources report national security news (and by extension the relevant quotations) to the same extent (e.g., certain newspapers do not cover the live Benghazi hearings in the USA at all[25]).

Finally, there are also significant differences in the media coverage of female politicians in online news articles (e.g., about Hillary Clinton [58]) and mainstream news broadcasts [63], signifying gender bias. There is also evidence of race bias in American news when it comes to covering stories about white or African Americans politicians, with the latter receiving disproportional attention [114]. Hence, analyzing additional bias kinds hidden in the reporting speech choices of newspapers could reveal how fair media coverage is to politicians in minority groups and whether it conforms to their seats in the parliament.

With regard to our study on news comments, it would be beneficial to apply more modern supervised learning approaches, e.g., artificial neural networks for our classification task in order to achieve better performance. More semantically meaningful text representations are also of interest to us, especially sentence and paragraph embeddings, considering that user comments are short textual documents. An interesting addition to the problem dimensions we examine would be to include user-related information as an additional signal (e.g., user embeddings [123]). One could also attempt to take into account the levels of subjectivity in the news text, as an indication of the writing style. Moreover, the potential polarity (positive, negative, neutral) of each comment is a valuable information as well. It might hold that users in certain newspapers express their emotions more than in others or that users from specific outlets tend to express more their disapproval and criticism to certain issues than in other sources. Note that instead of using all comment terms in our data, we also experimented with using only the named entities found in the comments. The results were slightly worse than the reported ones, therefore we proceeded to use all terms of the comments, as presented in this work. Although the named entities along with different feature combinations may work in the future, it is interesting to note that not just named entities are crucial for this problem, but verbs, adjectives and adverbs as well. Named entities mainly reveal a text's subject, whereas adjectives and adverbs represent the author's perspective and discussion

[25] https://www.washingtonpost.com/blogs/erik-wemple/wp/2015/10/23/why-fox-news-ditched-the-benghazi-hearing-and-msnbc-didnt/

style. Finally, we also intend to tackle the problem of class imbalance in our news dataset, e.g., with oversampling the minority classes instead of down-sampling the majority classes that might result to information loss.

In general, the problem of understanding how unique news media are and what reporting differences they have is challenging, and it could be dealt with other approaches than news source classification. Matching a expert of text (from quotes or comments, etc.) to the respective news source, it is indeed one of the possible indicators for media's distinctiveness or even bias, but there could be more. For instance, analyzing the news in a comparative and unsupervised manner can also bring further insights, i.e., with applying cross-newspaper document similarity techniques. One could also visualize these similarities over time and show how media perspectives differ and also how they shift over time with different governing parties. Incorporating additional datasets, such as news articles from the original news agencies that the stories come from, or documents from think tanks, can also be helpful. In this way, one could focus on the how media select their original sources and whether they agree or disagree with certain public opinions.

Chapter 3

POLITICAL MEDIA BIAS DETECTION

Unbiased and fair reporting is an integral part of ethical journalism. Yet, political propaganda and one-sided views can be found in the news and can cause distrust in media. Sometimes we also come across fake news articles or opinion pieces with extreme and even hateful language. Both accidental and deliberate political bias affect the readers and shape their views. Note that bias in mainstream media should not be expected and it is often very subtle, which makes it very difficult to detect.

Bias is also not well defined and certainly not perceived in the same way by all readers. Related work includes mainly manual studies in political science or computer science works that reduce the problem complexity by focusing on opinionated documents, e.g., political blogs or tweets. In addition, classifying a news article as (un)biased is particularly challenging due to the vague problem definition, which leads to unreliable and noisy annotations.

We contribute to a trustworthy media ecosystem by automatically identifying politically biased news articles [87]. In this Chapter, we classify news articles for their bias with deep learning techniques, while keeping humans-in-the-loop during the model training process. This is a joint work between the Hasso-Plattner-Institute, the Beuth University in Berlin and Factmata[1]. Moreover, it is also the first study that introduces humanly annotated articles for media bias detection and compares expert to non-expert annotators. We intuit that each dataset has unique characteristics and we investigate this with an extensive quantitative and qualitative analysis. Our goal is two-fold: to discover whether domain expertise is necessary for this task, and to show whether deep learning techniques can tackle such a challenging classification problem even for humans.

We share our mission against misinformation and one-sided views in the news with Factmata, a UK-based company that evaluates the quality and credibility of online content and helps other companies to avoid unsafe and biased content. During our collaboration in 2018, we have interacted with media experts to understand the media bias problem definition in the context of politics and to acquire valuable domain expert annotations of news articles for their bias.

[1] https://factmata.com/

The first part of our joint project that contains the data manipulation and statistical analysis is conducted while the author of this thesis worked as a summer intern for natural language processing and engineering at Factmata. The rest of the study and the conference paper were developed later on under the supervision of Prof. Alexander Loeser and Prof. Felix Naumann, and partially by Dr. Maria Mestre. The internship was supervised by Dr. Maria Mestre at Factmata and the overall objectives were defined by Dhruv Ghulati, CEO and founder of the company. The annotation tasks where news articles were marked as biased or unbiased by experts and non-experts were conducted at Factmata by Lusine Mehrabyan and Dr. Emmanuel Vincent respectively, with additional assistance by the author of this thesis and Dr. Maria Mestre.

The rest of this chapter is organized as follows: In Section 3.1 we present the existing challenges to detect media bias in newspapers and our motivation for solving this task. We also outline our contributions. In Section 3.2, we examine related work and point out our novel aspects of the problem. In Section 3.3, we introduce our novel datasets and Section 3.5 shows our data quality analysis. Furthermore, Section 3.6 introduces our media bias detection method and Section 3.7 presents the achieved results. An error analysis and interpretation is also contained in Section 3.7. Lastly, Section 3.8 consists of a summary of this chapter and Section 3.9 outlines our conclusions.

3.1 Motivation and Challenges

Given the vast amount of news we consume on a day-to-day basis, ensuring information quality and credibility [127] becomes increasingly crucial, because we need access to accurate and reliable news stories. This way, we can form well-rounded views and make informed choices for our votes. Unfortunately, between the emergence of fake news articles, political propaganda in the media, and also hateful language around the Web, it is important to be alert and potentially show mistrust to the providers of information. We consider the following definition of media bias (based on the Oxford University Press definition): A biased news article leans towards or against a certain person or opinion by making one-sided, misleading or unfair judgements. An unbiased news article reports fair, impartial and objective information.

Media bias can be expressed in multiple ways [136], for instance it can be present in word choices: some use the word “terrorists" vs. “freedom fighters" or “death tax"[2]

[2] https://thenewdaily.com.au/news/national/2019/04/24/bill-shorten-death-

vs. “inheritance tax"[3]. Even though such phenomena are present and can introduce bias in the news, reliable labelled corpora are missing to learn automatically the hidden patterns in the text. In fact, while there are relevant studies in political science [56], works that investigate the scope of bias [52], how it is generated [47] and others that detect it in different domains [30, 69, 132], related work lacks automatic solutions for the binary classification task that classifies mainstream news articles as biased or unbiased.

Moreover, we have observed opinionated news pieces that are not marked as “Opinion" or “Editorial" at the beginning of the article and they do use extreme political language. For instance, an article from *Right Wing News*[4] describes the Barack Obama administration as awful and another one from *Red State*[5] writes that liberals are regressive leftists with mental health issues, respectively. Even though the domain names reveal a stance in this case, other examples cannot always be captured by the commonly accepted newspaper stances [161]. Hence, we do not rely on predefined and commonly accepted slants of media [121], but we identify the importance of human labels for news media bias detection and introduce them here.

Detecting politically toxic content on the Web can prepare and protect both news readers and online social network communities from misleading or false information. Journalists can also benefit from such content evaluation in order to reflect on their work. News aggregators, such as Google News, can incorporate this feature along with others (e.g., fake claim, missing citations, etc.) to facilitate the user’s briefing and remove the lenses that certain news sources write their articles from. To the best of our knowledge, this is the first work that introduces news data with domain expert annotations for media bias, and compares them with crowd-sourced and silver standard (automatic) annotations as well.

Our goal is two-fold: to discover whether domain expertise is necessary for this task, and to show whether deep learning techniques can tackle such a challenging classification problem even for humans. Although our first research question might sound trivial, the complex nature of this problem and the lack of related work in

tax/

[3] https://www.independent.co.uk/money/spend-save/hmrc-inheritance-tax-bill-rise-23-per-cent-inland-revenue-treasury\protect\discretionary{\char\hyphenchar\font}{}{}a7860626.html

[4] www.rightwingnews.com/chelsea-clinton/chelsea-clinton-attempts-burn-republicans-tweet-instead-massively-insults-michelle-obama/

[5] https://www.redstate.com/setonmotley/2018/01/03/reversing-obama-trump-protecting-thus-promoting-intellectual-property/

computer science on news media bias leads us to investigate the differences between expert and non-expert annotations in a qualitative and quantitative manner. As a second step, we focus on the automatic prediction of media bias and aim to overcome the challenge of the vague media bias definition [56]. Our work includes the following steps:

- We introduce novel and reliable annotated datasets for media bias detection
- We are the first to compare experts and non-expert annotators for this task
- We classify the news articles with a deep learning model and a self-supervised curriculum learning technique
- We perform an error analysis of our results for further insights of the problem

During our collaboration with Factmata[6], we have interacted with several native English speaking journalists that helped us assess the quality of online news, by labeling news articles, giving us feedback on labels they find helpful for media bias detection, etc. Note that it is challenging to confidently define what is biased and what is not, because bias can be perceived differently by different individuals [52], even by experts. For instance, 80% of the journalists we collaborated with define political media bias as the act of writing the news so that they fit a specific political agenda, view or party. Few of them believe that the bias is often inevitable and it should be explicitly declared to avoid confusion.

3.2 Related Work

The problem of political bias in the news is originally and mainly tackled in **political science**, though lately it has gained attention in computer science as well. The survey by Hamborg et al. outlines the creation stages and effects of media bias [56]. The authors also outline the different forms of selection bias that social science studies. Very few computer science works exist that study news media bias and they mainly solve related sub-problems, e.g., source, topic, sentiment and event detection. Due to the difficulty to classify articles for their bias and the lack of training data, there exist **approximations** to understand this problem, e.g., examining the outlets' quoting patterns [113], leveraging information in social media [133, 182] and the political orientation of news readers [81].

[6] https://factmata.com

Other studies reduce the complexity of the bias detection problem by focusing on the **sentence level**, namely analyzing the choices news outlets make for the statements they publish and the politicians they mention [83], and also the news headlines they write [25]. In addition, Yano et al. annotate biased sentences in American political blogs and compare the perceived bias of the labelers to the commonly-accepted slant of the blogs [177]. In contrast, we aim to classify automatically political bias in traditional news articles on the article level (noted as *spin bias* [56]), whose text contains mainly subtle manifestations of political viewpoints that are not encouraged as they are in political blogs.

Furthermore, reporters often change their narrative in order to focus on a certain aspect, a technique that is called *news framing*. Related work analyzes specific types of framing in the media [108]. Another line of research performs a linguistic analysis of *hyperpartisan* (extremely biased) and fake news and shows that the latter are often politically biased [128]. **Writing style** features and readability scores are used to predict hyperpartisanship, political perspective and fake content. In general, **linguistic analyses** could reveal many interesting patterns in the text, but one might need to perform complex argument mining, opinion holder detection, or to identify direct and indirect reported speech (so that it is not attributed to the article author), etc. Political perspective detection is also studied on blogs [2, 95] and news outlets [7, 121]. However, we focus on the binary categorization of news articles into “biased" and “unbiased", rather than on particular cases of bias, e.g., left-wing/right-wing, conservative/liberal, unreliable/trustworthy etc., Moreover, recent studies propose **textual features** for the problem of deception detection on the Web in order to find unreliable information [166]. The authors utilize features such as biased language lexicons, connotation frames, writing style, etc. Opposed to this setting, we do not perform any cumbersome feature engineering, but we rely only on the content of the articles we classify.

To the best of our knowledge, there is not an existing automatic solution for classifying a news article in a binary manner as biased or unbiased, mainly due to the unavailability of reliable document-level labels by trustworthy annotators. Another reason is the noise of the existing labels inferred from the commonly accepted stance of the newspapers [161]. These inferred assumptions could potentially change over time due to trends or new owners and reporters joining the news outlets. In contrast, human labels are more reliable and potentially explainable, e.g., by looking into the annotator agreement or the notes annotators leave while labeling. In this work, we focus only on mainstream news media without engineering textual features

Table 3.1: Data characteristics: Number of news articles in each collection, number of annotations per article, labels, number of articles in each class and number of unique newspapers in each dataset.

				Classes		
Dataset	**Articles**	**Annotations/Article**	**Labels**	**Biased**	**Unbiased**	**Newspapers**
Experts (E)	1,154	3	0, 1	523	631	306
Non-experts (NE)	2,993	3	1, 2, 3, 4, 5	1197	1230	961
Publishers (P)	750,000	1	0, 1	375,000	375,000	1194

and predefined media slants, but we guide and improve our classification model by applying curriculum learning [11]. This technique has been shown to improve the classification performance and the training process in machine learning. It is also reported to outperform non-curriculum approaches in multiple tasks, such as language modeling, especially when the task is particularly challenging like ours [172].

3.3 News Corpora for Bias Detection

In this section we describe our political news datasets, which are scraped in a random manner from 2015 until 2018. All datasets are presented in Table 3.1. We gathered randomly selected news articles from the *politics* sections from a broad variety of English-speaking news sources in terms of size and credibility for our annotation task.[7] In addition to these humanly labeled articles (*E* and *NE*), we use the training data given to the participants of the Semeval 2019 task [75] for hyperpartisanship detection (denoted as *P*) to compare our performance against it. These publisher-based labels are produced based on newspaper credibility scores.

Articles annotated by journalists. As shown in Table 3.1, this is a rather small collection (*E*). However, due to the experience of the annotators in their field and their ability to identify one-sided text even in cases where bias is very subtle, we hypothesize that this dataset is very valuable. This set of news articles is included in the non-expert data as well (*NE*), in order to facilitate their comparison. The platform that was used is an internal annotation tool of Factmata, where the users (eight journalists) were asked to read a set of political news articles and mark at

[7]Example news sources: AbcBusinessNews, Associated Press, Albuquerque Journal, Baptist News Global, BBC, Breitbart, Chicago Reporter, Circa News, CNN, CounterCurrents, Daily Banter, Ethics and Public Policy Center, Fair, Federalist Press, Fox Business, Free Beacon, Greensboro, Guardian, Heavy, InfoWars, Intrepid Report, In These Times, Lima Charlie News, MotherJones, MSNBC, NBC News, NewsMax, New York Times, Occupy, OpsLens, Political Insider, Poynter Institute, Raw Story, Real News Network, Reuters, San Jose Mercury News, Seattle Times, Slate, Times of India, Townhall, Upworthy, Valley News, Vox, Washington Blade, 21st Century Wire, The Whim.

least one biased or unbiased text snippet that they find in each article, following the bias definition in Section 3.1 (i.e., the author is favoring or discriminating a certain view or person). The labelers were asked to read the entire article, identify the bias of the overall text and then highlight the evidence for their decision. By extension, the annotations can be words, sentences, paragraphs or entire documents. We chose this setting, because these low-level annotations can give more concrete evidence of bias and can be used as ground truth for explaining our model in the future [3]. This annotation exercise was a joint effort between the author of this thesis, Dr. Maria Mestre, and Lusine Mehrabyan.

We propagate these fine-grained labels to the article level and we assume that each article that contains at least one annotated biased (or unbiased) sentence is biased (or unbiased respectively). We exclude articles that contain both biased and unbiased marked text. This filter prunes less than 1% of the data, because the journalists were asked to not annotate each document exhaustively. It is obvious that regardless the annotations, a biased article will contain neutral text as well (and vice versa). However, the labelled text, either perceived as biased or unbiased, constitutes only the supporting evidence of the annotation, i.e., it corresponds to exemplary biased or fair content. We manually examined the aforementioned 1% of articles and we observed that sometimes in these cases the text contains the relevant facts, but also a few opinionated words that one might identify as biased. It also occurs that such articles are biased towards a given perspective, but they are well-written and cite the appropriate sources. We regard them as unclear, but we are interested in gaining insights into these potentially controversial news pieces in our future work.

Articles annotated by the crowd. The next dataset consists of annotations from our two crowd-sourcing tasks for media bias detection, launched in the Amazon Mechanical Turk (AMT)[8] (1,979 documents) and the Figure Eight[9] (1,014 articles) platforms in Factmata. The author of this thesis contributed to the Amazon Mechanical Turk task with enhancing the annotator instructions and preparing the task conceptually, e.g., the names of the labels, with Lusine Mehrabyan and Dr. Maria Mestre, while Dr. Emmanuel Vincent launched this annotation exercise in AMT. Note that the Figure Eight dataset is originally introduced in 2018 [164], though in this work we consider the full dataset, instead of the proposed filtered version based on an in-house evaluation of the data. In both datasets, the crowd workers evaluated each article using a score range similar to related work [177], where 1 meant

[8] https://www.mturk.com/
[9] https://www.figure-eight.com/

“unbiased" and 5 signified “biased". Similarly with the experts, they were asked to follow the media bias definition in Section 3.1 and read the full article before they annotate. Both the crowd and the experts were asked to be mindful of bias manifestations, such as loaded or subjective language, opinionated text, one-sided claims, or unsupported arguments. As we can observe in Table 3.1, the combination of these two non-expert (*NE*) data collections contains almost 3,000 news articles labeled for their political bias. We have combined the annotations from these two tasks into one unified dataset. The expert and non-expert document collections are available via our industry collaborator for further details and research purposes.

3.4 Data Preprocessing

In this section, we explain how we aggregate and transform our datasets.

Article transformation. There are at least three annotations per article in *E* and *NE*, and the class distribution in each case is fairly balanced. In order to aggregate the labels of multiple annotators for each article, we apply the *Dawid Skene* algorithm [34], specifically an optimized variation of it [148]. This model produces one final label for each document and it improves on simpler methods, because it considers the annotators' bias and competence. It is assumed that each worker corresponds to a confusion matrix that shows the joint probability distribution over correct and reported labels. The correct labels are initialized with the *Majority Vote* method, which outputs the label that was reported most often. For a N-way classification task (in our case $N = 2$), a worker w and a data instance d, the Dawid-Skene assumption is as follows:

$$P(X_{wd} = l) = p^*_{wlx_d}$$

where X_{wd} is the random variable that models the reported label l of annotator w and for the document d, and all X_{wd} are mutually independent. After generating one annotation per article, we still face the challenge that *E* contains binary labels, but *NE* corresponds to a multi-class classification setting. For this purpose, we binarize the non-expert data, following the literature in similar tasks where five star ranges were used [99]. We take into account only the two ends of the scale, namely only the highly polarized text. That is, we consider the articles with bias score 1 and 2 as unbiased (negative class), and the ones with bias score 4 and 5 as biased (positive class). Similarly to *E*, we exclude ambiguously labeled data (bias score is 3).

Unambiguous test set for media bias detection. We construct a reliable and independent of our training data test set in order to compare the achieved classification performance with training data labeled by different communities. We take into account all articles that are annotated by experts and non-experts as well — namely, all articles in E as it has a smaller size, and then use a subset of them for our tests. That is, we consider the subset of articles that are marked with the same label both by the experts and the crowd, because we hypothesize that these articles have low uncertainty and controversy regarding the underlying media bias. From this unambiguous dataset, we randomly sample 40% of it and use it as our final test set. We leave the rest 60% in E and NE respectively. We do so in order to maintain our training data sufficiently large, given that in our experiments we remove from the training sets any article that appears also in the test set. Hence with this setting, our training data contain "diverse" articles, whose labels might or might not be the same in E and NE.

3.5 Label Quality Assessment

In this section, we describe our annotation analysis as an effort to determine the quality of the datasets and improve our classification results later on.

3.5.1 Per-dataset Agreement

As a first step to examine the quality of the human labels, we measure the inter-annotator agreement (*ITA*) within each collection. That is, we calculate the agreement for the expert dataset, the Figure Eight dataset and the MTurk dataset separately. Note that for this experiment we consider the original labels in the raw data, without binarizing them first (we transform the labels as described in Section 3.4 only later on for machine learning purposes). We chose Krippendorff's α coefficient[10], which is independent of the sample size, the categories, and numbers of annotators and measurement levels. Krippendorff's α for a text document is defined as follows:

$$\alpha = \frac{p_a - p_e}{1 - p_e}$$

where p_a is the weighted percent agreement and p_e to the weighted percent chance agreement. According to this metric, the documents and the agreement scores assigned to them are statistically unrelated. When $\alpha = 1$, this indicates perfect reliability and when $\alpha = 0$, there is absence of reliability. Moreover, α is zero when disagreements are systematic and exceed what can be expected by chance.

[10] https://en.wikipedia.org/wiki/Krippendorff%27s_alpha

Table 3.2: Inter-annotator agreement (Krippendorff's α) for each of the three humanly labeled datasets.

Dataset	**ITA**
Crowd workers (Figure Eight)	0.21
Experts (Journalists)	0.59
Crowd workers (MTurk)	0.66

We present our findings in Table 3.2. Considering how challenging the given problem is, we observe the expert (*E*) and MTurk annotators to agree sufficiently well internally in each collection. However, the data produced via Figure Eight seem more ambiguous. Chronologically, we have performed these annotations tasks starting with Figure Eight, continuing with the journalists and then completing our study with MTurk. That is why the differences in the agreement could be justified due to the continuous improvement of our instructions to the annotators, which potentially makes the annotations' quality higher at the later rounds in contrast to the earlier ones. For instance, we discover that we had to explicitly emphasize to all annotators the difference between when a reporter's words and viewpoints are toxic themselves, to when a politically toxic event or statement is reported, and that we are only interested in the first case.

Furthermore, the labeled dataset from Figure Eight is introduced earlier [164], where an in-house gold standard dataset based on fact-checking was used to evaluate the workers and disqualify unreliable ones. In our study we consider the full dataset (thus, we see a lower inter-annotator agreement), in order to maintain a more generalized setting without constraints. Note that both crowd-sourced datasets use a numerical range for the bias score. We leverage the numerical distance between the labels when computing the *ITA*, which is not possible in a binary setting, e.g., in the expert dataset. Taking this range into account, we have significantly improved the inter-annotator agreement (from 0.14 to 0.21 in Figure Eight and from 0.44 to 0.66 in MTurk), where both original agreement scores are lower than in *E*.

3.5.2 Cross-dataset Agreement

To investigate whether media expertise is necessary for our task, we compute the annotator agreement between *E* and *NE*. For the crowd-sourced data, the transformed annotations to binary labels are used as described in Section 3.4. We apply a well-established method for expert versus non-expert analysis in natural language processing tasks [150], using the articles that both *E* and *NE* annotated. The authors calculate how (non-) experts perform within their community and against all involved

annotators (experts and non-experts combined). Given two communities A and B, for every individual a_i in A, they compute the *ITA* with all the individuals b_j in B and then average the results. In the following step, they average across all individuals a_i in order to obtain how well A agrees with B in total. The authors use the Pearson correlation coefficient (*PCC*) as agreement metric. Given two vectors (the labels of two different annotators), the computed *PCC* has a value between 1 (positive correlation) and -1 (negative correlation).

Table 3.3: Inter-annotator agreement and standard deviation based on the method of Snow et al. [150]. *E* refers to the experts, *NE* to the non-experts and and *All* to both. For the crowd-sourced data, the binarized annotations are used as described in Section 3.4

Compared sets	**Agreement** %	**STDEV** %
E vs. *NE*	73.68	17.13
E vs. *E*	65.46	15.46
E vs. *All*	67.39	7.87
NE vs. *NE*	64.76	19.51
NE vs. *All*	64.32	20.3

For our task, this agreement metric is not appropriate, because not all user pairs annotated exactly the same amount of articles and this makes *PCC* not work as expected: it yields a high score when two annotators have many common articles, and very low score (close to zero) when the shared articles are few. We have worked with a limited number of eight journalists, as it is cumbersome and expensive to obtain domain expert annotations, but the crowd-sourcing platforms are generally low-cost and employ a very high number of annotators for their tasks (in our case eighty). Thus, the non-experts have annotated generally more articles and also more articles in common with each other – the latter could make their agreement scores more robust. We apply a simpler method instead of *PCC*, i.e., the percentage of times that two annotators agreed on the article bias.

Our findings are presented in Table 3.3. Surprisingly, the expert community and the crowd-workers appear to agree on what is biased and what is not at approximately 70% of the time. Thus, in the majority of the articles the individuals in *E* and *NE* recognize the evidence in the text to mark it as biased or unbiased. We hypothesize that in the majority of the agreement cases the articles are either very obviously hyperpartisan or very fair and balanced news, and potentially the disagreement occurs when the article topics are more controversial and ambiguous. Interestingly, the STDEV in *E* vs. *All* is much lower than in *NE* vs. *All*, which can be an indicator

of the consistency and reliability of the journalists. Thus, given the lower variance, one might not need as many expert annotators as crowd workers to obtain a high quality media bias detection dataset. Moreover, journalists do not agree with each other significantly more than non-experts agree with one another. This could be potentially explained by the fact that media bias can be a very sensitive and often times subjective topic for journalists. Therefore, so far we observe unexpected yet not entirely conclusive results regarding the superiority of either annotator group.

3.6 Article Classification

In this section, we describe our approach to detect media bias automatically and how we apply a curriculum learning technique to improve our results.

3.6.1 Baseline Method

We use the FastText classifier [71], a basic neural network that uses averaged bag of n-gram features, which is a model well-known for competitive results to state-of-the-art approaches for supervised text classification. We run all our experiments with the learning rate set to 0.1 and for 500 epochs. In these experiments, we shuffle our input data and perform ten iterations and then report the averaged results.

3.6.2 Curriculum Learning Method

To enhance our baseline model, we leverage the data quality assessment we performed in the Section 3.5, and apply a *curriculum learning* approach, which is based on *transfer learning*: Given a target classification task T_1 and an external one T_2, transfer learning techniques that solve T_1 could leverage information derived by T_2 in different ways. For instance, one can use word embeddings or losses of output layers trained on T_2, or take an entire network designed for T_2 and train it on T_1 to improve the classification performance. Unlike traditional transfer learning approaches, our external information is not provided by another classifier, dataset or task. In contrast, it is derived by the humans that share our mission to fight misinformation on the Web and contribute to our task by labeling our political news articles. Ultimately, using their wisdom, we aim to guide our classifier during training with some initial data instances (“easy to learn examples") and perform better in the next steps.

We follow the definition of *curriculum* as introduced by Bengio et al. [11], i.e., sorting the training examples from “easy" to “difficult" and introducing them to our classifier in this order during training to avoid confusing the learner. This method can not only speed up the training process, but it can improve the classification results and model generalization as well. The authors perform experiments on

shape recognition and language modeling in their work. For the latter task (which is more relevant to ours), the curriculum learning strategy is to grow the vocabulary size gradually, i.e., starting from the most popular words in a Wikipedia corpus and then considering more words in each training pass.

Our learning difficulty definition. In our proposed approach, we leverage our previous agreement analysis and build a curriculum that stems from the quality of the article annotations. That is, we compute the inter-annotator agreement (*ITA*) in *E* and *NE* with the Krippendorff's alpha coefficient as shown in Section 3.5.1 We consider the agreement score to be the learning difficulty of an article. This choice is based on the assumption that an "easy" article is an article that causes very low to no disagreement between its annotators regarding its bias. We hypothesize that these news pieces are either very objective or very subjective, and hence this makes the decisions of the annotators simpler. On the other side, newspaper articles with high label disagreement may indicate controversy and potentially contain a mixture of facts and opinionated words. We leave these difficult-to-learn examples to be given to our model after the clearer examples have been introduced. We split the training data into 10 parts, namely we first consider the top-10% of the documents with the highest agreement score, then the top-20% and so on and so forth. We build a classifier with each of these data chunks and every time we load the latest calculated weights from the previously trained model. We then fit the current training set and predict the media bias of our test set. Our technique is also similar to the *stochastic curriculum learning* definition [172], which is a variation of stochastic gradient descent, where the model imports training data instances gradually based on their difficulty score. Unlike our approach, the authors define their curriculum without the presence of human knowledge.

Evaluation setting. Since the annotator agreement results for the Figure Eight dataset were not satisfactory, we use only the MTurk dataset for the rest of our experiments. Thus, when we refer to the non-expert annotations, only the dataset from MTurk is considered. We have actually trained our model on the Figure Eight dataset and the performance was similar to a random decision – we do not report the detailed results here. Hence, we concluded that it is necessary for the inter-annotator agreement to be at least 60% for our task and the Figure Eight labels did not achieve it. Related work also reports similar results (0.55 Cohen Kappa score) for crowd-sourcing biased sentences in blogs. Furthermore, even though the precision achieved with the Figure Eight dataset in the work of Vincent [164] is promising (approx.

Table 3.4: Article sizes of our three different training sets and our unambiguous test data.

Training sets			Test set
E	*NE*	*P*	$\subset(E \cap NE)$
759	1,805	750,000	237

70% on their own test set), the in-house manual improvements during training and testing that the authors perform raise the question of generalization potential of their approach. Thus, we leave investigating this data for future research.

We show the size of our training and test sets in Table 3.4. As mentioned earlier in Section 3.4, our test set is balanced and it consists of a random sample of the news articles for which both the domain experts and the crowd workers agreed on their labels in order to eliminate noise. After removing these 237 articles from the training data, there are 759 articles in *E* and 1,805 in *NE* remaining for our model to learn (*P* does not exhibit an overlap with *E* and *NE*). Note that a similar setting is used at the Semeval competition, where the unknown test set is also a small balanced (and crowd-sourced) dataset of 645 news articles. We also prefer this predefined unseen test set instead of cross validation (which is appropriate for small datasets like ours), because we can maintain the same test set across all our experiments with different training datasets. Especially for the Semeval data that we compare against, cross-validation would not work, as we only aim to test on humanly annotated documents.

3.7 Results

In this section we describe our experimental evaluation and we show the qualitative results of our error analysis.

3.7.1 Domain Expertise Stands out

We use the expert and non-expert annotated articles (*E* and *NE* respectively) for training a FastText classifier. In the first two lines of Table 3.5 we can see that the articles annotated by journalists are a more appropriate dataset for this task, because when our model is trained with it, it achieves significantly higher precision. The model trained with crowd-sourced labeled articles constitutes a promising dataset that achieves competitive results with the expert model, though it does not outperform the performance of the model trained with *E*. Note that even though the consensus in MTurk is higher than in the expert data (see Table 3.2), the prediction

Table 3.5: Classification results of our model trained with: expert data, non-expert data, expert data with curriculum learning and non-expert data with curriculum learning. Our test set is a sample of the articles where both experts and non-experts agree.

Training	**Precision**	**Recall**	**F-1**
E	0.90	0.89	0.89
NE	0.85	0.89	0.87
E_c	0.93	0.95	0.93
NE_c	0.79	0.86	0.82

power of MTurk is lower. Thus, higher inter-annotator agreement does not automatically lead to higher classification results in this case. In the following lines we see the classification results of the same models, but this time trained incrementally with a curriculum created based on the learning difficulty of each data instance. The achieved F-measure with *E_c* is significantly higher than the one with *E*, namely 93%. Note that recent related work on similar tasks [128] that uses linguistic features of news articles achieves a maximum of 86% precision for hyperpartisanship and 75% precision for political orientation classification.

Unfortunately, the curriculum constructed by the knowledge of the crowd is not as useful for this task as the one by experts. In fact, it worsens the performance of *NE* by decreasing the achieved precision from 85% to 79%. Note that the training dataset constructed by crowd workers is more than twice the size of the one by journalists and overall it still shows a lower F-1 measure for our task, with or without curriculum learning. We hypothesize that this outcome signifies the limits of mass labeling in crowd-sourcing platforms for tasks that are often not clearly defined and easily solvable even by humans, e.g., bias, irony and sarcasm detection. Furthermore, we also performed experiments with a dense feed forward neural network and a network with long-short memory units (*LSTM*) that we do not report in detail here. Our results were not as satisfactory as with FastText (approximately 20% worse). We hypothesize that these networks are potentially too big and too complex for our small humanly labeled datasets (which is why transformer architectures would also likely not work). Traditional news articles are also less noisy datasets in contrast to text that is user generated, and thus a word-based input is appropriate for our task. For future work, we are interested in applying attention mechanisms that might capture specific biased terms in the text.

Table 3.6: Comparison of our models (*E_c* and *NE_c*) with anti-curriculum learning (*E_rc* and *NE_rc*) and with learning from automatically labelled data (*P*). Our test set is a sample of the articles where both experts and non-experts agree.

Training	Precision	Recall	F-1
E_c	0.93	0.95	0.93
E_rc	0.85	0.90	0.88
NE_c	0.79	0.86	0.82
NE_rc	0.78	0.87	0.83
P	0.54	0.89	0.67

3.7.2 Bias Detection Requires Expert Curriculum

We compare our proposed solution to different methods in Table 3.6. In order to confirm the usefulness of an expert curriculum, we compare our approach with an *anti-curriculum* approach. Namely, we rank our training data instances in an ascending order of their learning difficulty as defined in Section 3.6.2. In this way, we introduce the most ambiguous and hard to learn examples to our classifier first, and then proceed with the rest of the training data, completing the learning process with the easiest examples. We show in Table 3.6 that, as expected, this "reverse" curriculum technique (*E_rc*) worsens the results of our expert-based model significantly. In addition, it produces almost the same outcome for the non-expert data (*NE_rc*). We hypothesize that for this reason the labels of the crowd are not of the same potential as the ones by journalists. That is, the non-expert consensus for a given article does not provide additional intuition or help to a media bias detector.

In Figure 3.1 we show how our precision increases while we increase the training set size using *E_c*, *E_rc*, *NE_c* and *NE_rc*. We observe that *E_c* outperforms the rest during the whole training process, and it starts approximately at the same precision value as *NE_c* does. It is remarkable that only the top 20% of the expert data with the lowest learning difficulty can already achieve 80% precision. Furthermore, all four models improve as the training set size increases, however the curves of *NE_c* and *NE_rc* almost overlap. This indicates that a crowd curriculum does not prevail over its anti-curriculum version, and thus it is not as helpful to the learner as the expert-based one. Lastly, we see a significant difference between *E_c* and *E_rc* both in the starting point and during training.

3.7.3 Quality is More Important than Quantity

We perform an additional comparison of our approach to a model trained with automatically labelled articles for their media bias. As briefly mentioned in Section 3.3,

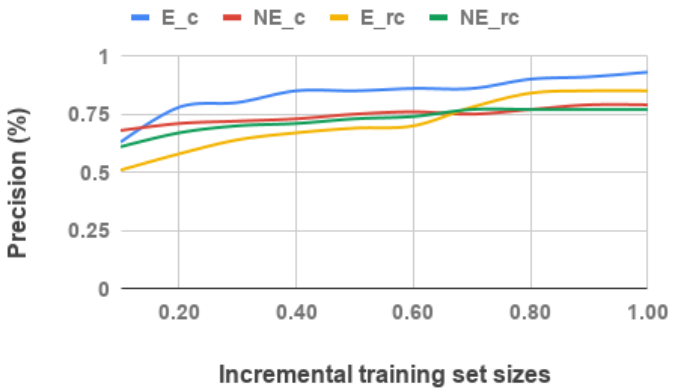


Figure 3.1: Achieved precision by our model trained with expert curriculum (*E_c*), non-expert curriculum (*NE_c*), compared to the respective reverse settings, *E_rc* and *NE_rc*. We train incrementally with step=10%, starting with the top 10% of the training set that is the “easiest" (*c*) or “hardest" (*rc*) to learn.

we consider articles with inferred publisher-based bias from a Semeval competition [75]. In this dataset (*P*), each document is marked as hyperpartisan (or not) if the news outlet where the article originates from is considered extremely politically biased (or not). As shown in Table 3.1, this dataset is considerably larger than ours. Note that a few hundreds of annotated articles by the crowd were included in the Semeval training set by the organizers, which we removed from *P*, because we aim to compare our small manually labeled training datasets to a massive silver standard dataset.

The comparative results are shown at the bottom in Table 3.6. It is evident that small amounts of human labels and domain expertise are more essential for our task than the size of training data, which is not true for every machine learning problem. Large amounts of weak labels are sufficient for other tasks, such as sentiment classification [36]. Other similar works that miss correct labels include tracking and matching individuals in images with transfer learning [124]. The outcome can also be justified, because we consider the Semeval dataset to address a slightly different task, namely the prediction of the newspaper’s bias and not the article’s (similarly to Aires et al., where domain level labels are used [121]).

It is expected that the Semeval dataset can achieve competitive recall values (almost 90%) due to its very large size, but the precision is still suffering from the uncertain quality in the training set. Our manual qualitative examination shows that there is significant noise in Semeval the data, which is on par with the results (60–70% classification accuracy) of recent studies based on this dataset [139]. For instance,

every news article by *Pjmedia* is considered hyperpartisan in this dataset, but not all articles from this news source in the MTurk dataset are labeled as such. Note that although our test set is small, after comparing our approach with the classifier trained on *P*, we consider our results indeed significant. That is, we assume that if our test set was easy to classify, then the baseline with the Semeval data would be able to outperform or at least compete with our proposed expert-based approach.

3.7.4 Qualitative Analyses Bring Further Insights

In this section we analyze the errors of our expert curriculum model (*E_c*) and also apply it to a new dataset.

False predictions. Approximately 9% of our predictions are incorrect. Over 50% of the articles that are misclassified contain **loaded language**. Heavy words can be either the journalist's (an article calls Donald Trump "misogynist") or could describe a sensitive topic (the same article is discussing "sexual assault"). Hence, sentiment detection alone would be a rather inconclusive approach, because such words are not always chosen by the journalist, but are often contained in cited text (this is one of the multiple reasons that sentiment analysis performs poorly on the news corpora [56]). Moreover, in 60% of our errors, the content of each misclassified article splits evenly into facts and opinions. Some of them are essentially opinion pieces with factual information and verified sources, but are **disguised columns**, i.e., there is no declaration of this in any part of the news article page. Even though they are annotated correctly and they are almost all classified correctly by the model, there is still a very small set that is very hard to classify automatically. In addition, in around 30% of the errors we observe humor and satire in the text, thus a filter or another model could be used to avoid such cases.

A very interesting error class with approximately 60% of errors appears when essentially the **news topic is a politician**, and not a political event. Among these errors, over 85% of them are false positives and the rest are false negatives. Such articles generally report a politician's statement or action, and at the same time describe them with endorsement or criticism. An example is the article of *MSNBC*, where the journalist describes Michelle Obama's standpoints on Donald Trump's taped comments about women. This article is a representative error, because it has somewhat subjective tone ("Michelle Obama *slammed* Donald Trump"), discusses a sensitive subject and it is about two politicians.

A very challenging task for our model is to distinguish the presence of bias when

the article is about very **sensitive topics**, e.g., incidents of racism, sexual assault, terrorism, brutal crimes. These errors (13%) are all false positive predictions, which indicates that loaded language can sometimes lead the model to confuse tragic news stories with biased reporting. Note that this is not a rule, because we also classify relevant unbiased articles correctly (e.g., an article in *Circa News* about the domestic terrorism attack in Charlottesville was a true negative). Furthermore, in about 10% of all errors, the article author is using first-person pronouns, which could be discovered with claim/argument mining. First-person expressions could serve as an indicator that an article contains the author's/newspaper's subjective point of view [46] or that it is an editorial [16].

Results on independent dataset. We additionally apply our model to a small recent set of news articles from the New York Times. We use the newspaper's *Most Popular API* to get the most read articles in mid August 2019. Out of the 17 articles in this test set, our model classified 13 as unbiased (including an opinion article that our algorithm missed), and only four of them were classified as biased. Among these four, one article is an opinion piece and another one is self-help guide giving relevant professional opinions, which justifies the decision of our model. The other two articles are about brutal crimes in Afghanistan and New York, respectively, and we consider them falsely classified as biased. The first article about a suicide bomber who killed dozens of people in the capital city of Afghanistan describes the tragic event with factual reporting. However, the language is somewhat loaded (mainly due to the nature of the news story) and the title is described by one commenter as too dramatic. Similarly to our error analysis of our own test set, we see that such tragic event reports are harder to classify correctly.

Moreover, in the article about a crime committed by a police officer in the New York region, we observe only factual and fair reporting. Thus we regard this as false positive prediction as well. According to MediaBiasFactCheck[11], the New York Times is a highly factual and reliable unbiased source, that occasionally publishes articles with loaded language that moderately favors liberal views. We find our qualitative study to be on par with MediaBiasFactCheck, because our model labels the majority of the articles unbiased, captures almost all the opinion pieces (which are explicitly declared and do not belong to our focus) and understandably misclassifies the articles on hard-to-classify topics due to the presence of emotional words.

[11] https://mediabiasfactcheck.com/new-york-times/

3.8 Summary and Findings

In this chapter, we describe our contribution on classifying political news articles as a whole for their potential media bias, i.e., we solve a binary classification task [87]. Thus, we tackle the problem of media bias detection in the news with machine learning techniques. We introduce two novel humanly labeled article sets and use them to build very competitive deep learning models for our task. Our work is the first to consider and compare human labels (by domain experts and crowd-source workers) to automatically derived labels for media bias detection. We classify news articles successfully for their bias and also give human knowledge to our model as a *curriculum*, by introducing the articles incrementally during training. We also contribute further insights with our manual error interpretation and discover challenging corner cases to be aware when annotating or classifying new media bias.

One of our main findings is that human labels are more suitable than automatic labels for this task, with both models trained on crowd and expert data respectively achieving higher F-1 scores than models with automatic labels. The expert knowledge can be used in the form of a curriculum to boost the classification performance further, e.g., a model trained with the top-20% articles with the highest consensus among experts can already achieve 80% precision on this task. An interesting finding is that few hundreds of domain expert labels are more efficient for this task than almost one billion automatic/weak article labels. Moreover, we also observe that the inter-annotator agreement score for media bias detection should be at least 60% and that the amount of training data is not as influential as its quality. Our conclusion is that human expertise and data quality are essential for this problem – more than the amount of annotated documents, and the classification performance can increase significantly when expertise is being used in a transfer learning setting, i.e., as a curriculum that assists the learner. We also come to the conclusion that there is still room for improvement, especially in regard to the very difficult to classify cases. For instance, we observed the phenomenon of “disguised" editorials and columns, i.e., opinion pieces with a mixture of loaded language, but also facts and citations of relevant sources. These articles are not only challenging for a model to classify, but also for the annotators to decide in which class they belong to.

3.9 Future Work

News producers and reporters have the liberty to decide what will become publicly known and in which context this news will be given. Media bias or perceived media

bias can likely be found in various news articles and stories even if it is subtle and accidental, or it happens due to editorial policies, marketing purposes, political beliefs, etc. Our findings intuitively suggest that the kind of experts with the highest potential to detect media bias in the news are the ones that are often responsible for it – the journalists. However, we conduct our study for a small set of expert and non-expert annotations, which indicates that working with larger datasets might bring further insights into the problem. That is, there is still room for improvement regarding the recruitment of domain experts based on their professional/personal background and credibility, in order to create a media bias detection model that is stable and can always generalize. We hypothesize that the highest potential would be achieved when computer scientists and political scientists collaborate to create clearer requirements for the training data of machine learning model for media bias detection.

Moreover, we aim to shed more light into our ambiguous human annotations (articles marked with score=3/5 in the non-expert data and articles with biased and unbiased snippets in the expert data). This set of articles could become a very difficult and interesting test set for our task, or a training set for controversy detection in the news. We also intend to grow the overlapping articles that both experts and non-experts have annotated to obtain more information on their (dis)agreement cases and to obtain more evidence on the potential of each annotator group. To this end, we supervised a bachelor thesis and the Beuth University that created a browser plugin for annotating and identifying text in webpages that is either biased or hateful[12]. This thesis was supervised by Prof. Alexander Loeser, Betty van Aken and the author of this thesis.

We also plan to experiment with stricter learning difficulty scores, e.g., the global annotator agreement in all collections instead of the internal agreement within each collection. We intuit that this could result to an even more robust and powerful expert curriculum. Furthermore, we are also interested in transferring external political knowledge (texts from presidential debates, press conferences, opinion articles and editorials, etc.)[13] and using it to train our own document embeddings as means to improve our data representation. We hypothesize that such embeddings would be potentially richer in terms of capturing typical or even loaded political language used by parliamentary members and hence, this in turn could make it feasible to detect such language in the news.

[12] https://github.com/s61211/Finder-for-Hate-Speech
[13] https://manifesto-project.wzb.eu/

In addition, we also believe that a universal, scientific and straightforward definition of the media bias problem along with a variety of representative examples given by domain experts, is a very important future work direction for media analysis. Such a definition can assist the general public to be alert of misinformation, the journalists to reflect on their work and the scientific community to advance the solutions for this problem further. Lastly, our future work also includes working on the explainability of our bias detection model. This is a very challenging and essential task, since both the journalists and the readers should be aware of the reasons that our model labels a news articles as fair or unfair. We believe that neural networks are very powerful algorithms for solving media bias detection, but they always come with the cost of being black boxes, which we would like to address next.

Chapter 4

POLITICAL DISCOURSE IN USER GENERATED CONTENT

Online news and social media have revolutionized communication and information sharing. With online journalism emerging in the late nineties, and social networks making their appearance with the turn of the century, a new environment for social interactions had been created. Nowadays, not only do social network users log in to such tools (e.g., Twitter) to communicate across great distances, but also to engage in political discussions. Similarly, users are also interacting via comment threads under news articles and thus contribute to online political debates. Political parties and politicians also use social media to propagate their messages or to promote their political agendas. Analyzing such short messages can assist us to understand the general public's interests, especially in the context of politics, where we can often encounter online debates over breaking or controversial news and the involved politicians.

The convenience and anonymity in this online setting can motivate individuals to engage in the ongoing discussions and thus share their perspectives and beliefs [176]. On the other hand, this freedom comes with the risk of encouraging polarization, hateful content and uncivil behavior. User generated content is anticipated to be biased by default towards the opinion of the author, but it could also contain toxic, one-sided or hateful text, which can harm the media's balance [159, 178]. Hence, analyzing such short online documents, e.g., the topics and opinions in news comments, user reviews and social media posts, can bring us a step closer to promoting and protecting the quality and credibility of content shared in online media.

News often spread faster than in mainstream media, along with additional context, facts and aspects about the current affairs. Hence, many works focus on finding breaking news and topics on Twitter [103, 141]. Additionally, users in social networks are up-to-date with the details of real-world events and the involved individuals. Examples include crime scenes and potential perpetrator descriptions, public gatherings with rumors about celebrities among the guests, rallies by prominent politicians, concerts by musicians, etc. This chapter is motivated by and focused on the vast amount of online social information and its up-to-dateness. Specifically in the case of the prompt political discussions and debates during election periods,

the shared information about the involved individuals is often very current and insightful. We believe that social media is a very powerful instrument for information flow as the news sources also are, and we use its unique characteristic of rapid news coverage in this chapter for two applications.

We analyze Twitter messages and debate transcripts during live political presidential debates in 2016 to predict the topics that Twitter users discuss. Our goal is to discover the favoured topics in online communities on the dates of political events as a way to understand the political subjects of public interest. This work is implemented by Jaqueline Pollak during her master thesis [1], supervised by Prof. Felix Naumann, Dr. Toni Gruetze and the author of this thesis. With the up-to-dateness of microblogs, an additional opportunity emerges, namely to use social media posts and leverage the real-time verity about discussed individuals to find their locations. That is, given a person entity, we use the wisdom of the crowd to track her physical locations over time. We evaluate our approach in the context of politics. More specifically, we examine the political discourse on Twitter during the last presidential elections in the U.S.A. in 2016 (we analyze the same dataset as the above-mentioned master thesis) and we predict the locations of US politicians. We identify our work as a proof of concept for important use cases, such as to track people that are national risks, e.g., warlords and wanted criminals.

The poster paper with our initial baseline solution for person tracking in social media was developed by Dr. Toni Gruetze in 2017 [54]. This solution was jointly extended and improved in our subsequent research paper [85] in 2018. In this work, Gruetze focused on the data quality evaluation approach (pruning the irrelevant data to our task) and the author of this thesis worked on the machine learning approach for person tracking and an extensive experimental evaluation. The approaches and results achieved by Jaqueline Pollak is briefly described in Section 4.7.

The rest of the chapter is organized as follows: Section 4.1 introduces our work and Section 4.2 presents related research on detecting topics, locations and events on Twitter. Section 4.3 introduces our method to discover relevant tweets for person tracking and Section 4.4 discusses how we determine the individuals' locations. Section 4.5 shows our evaluation results for noise filtering and person tracking. Section 4.6 summarizes this line of work and in Section 4.7 we present the problem that Pollak addressed in her master thesis while using the same dataset we crawled

[1] https://hpi.de/naumann/projects/completed-projects/politics-on-twitter.html

for our work [54, 85] and also show an overview of her approach and findings. Lastly, we conclude this chapter in Section 4.8.

4.1 Challenges and Use Cases for Location Detection in Social Media

Millions of people publish their thoughts and experiences on various social networks and microblogs, such as Facebook, Twitter, etc. Users share real-time information via text messages, geo-located images, live videos etc. An example of the speed and brevity specifically of Twitter is the shooting outside the Texas Irving mall in 2011. The incident was reported by a very short tweet immediately after the shooting, in contrast to newspapers, which reacted with a 3-hour delay [93]. Similarly, users are also likely to inform their peers about a natural disaster outbreak online, even before the first news story is published [66]. Hence, Twitter can be seen as a fast and decentralized news media.

Furthermore, users keep their peers up-to-date, by retweeting, quoting and engaging in discussions about the current affairs. When considering that most of the posts on Twitter have no visibility restrictions, it is reasonable to claim that this platform "breaks down the communication barriers" [141]. According to Kwak et al., regardless the popularity of the original account, any random retweet spreads over the network almost instantly [82]. This means that every retweet is expected to reach 1,000 users on average, imposing its impact to the rest of the network. Thus, Twitter users can be extremely influential by sharing real-time ongoing news, including civil unrest, entertainment activities, earthquakes and floods, etc. This vast amount of information has attracted various Twitter analyses, particularly related to the problem of event [55] and location [130] detection in social media, with the latter being essential due to the very low amount of geo-tagged tweets.

In this work, we are interested in a location detection problem that leverages the up-to-dateness of social media (e.g., microblogs), that is: the task of *person tracking*. Unlike related work on user location detection, we consider the individuals to be mentioned in discussions in the Twittersphere, rather than assuming that they hold a user profile. We prefer to rely on the wisdom of the crowd that discusses about a given person p, because we hypothesize that it brings many more tweets as evidence on p's locations than p might potentially share him/herself. We also do not assume that a location mentioned in a user post is identical to this user's current position. Thus, we allow users who discuss event locations asynchronously.

As shown in Figure 4.1, by detecting where a music band (*Lovelyz*) is or plans to be, a user can decide to join their concert and browse people's comments and

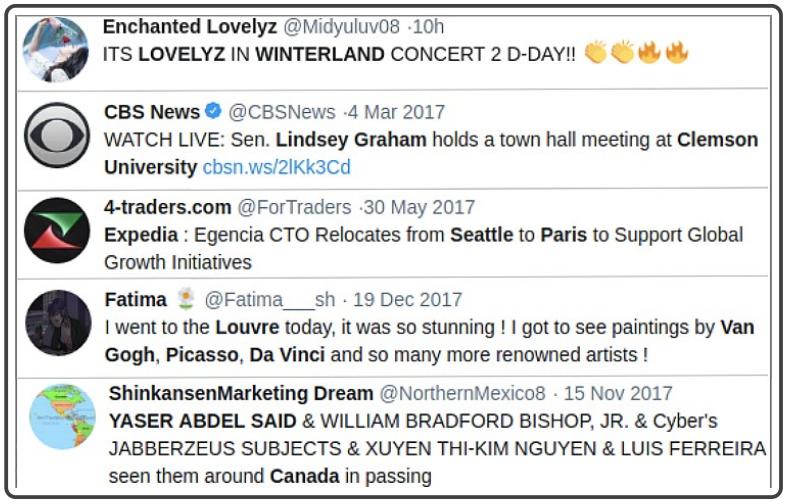


Figure 4.1: Tweets that indicate the locations of different entities.

anticipation about this specific event. Similarly for politicians (*Lindsey Graham*), we can leverage tweets discussing about them to discover the town hall meeting they hold. Additionally, target entities might also be companies that relocate (*Expedia*), or objects, such as famous art pieces that are moving to different countries over time (*Van Gogh, Picasso, Da Vinci*).

Moreover, an important use case covered by our approach is the ability to track people that are national risks, such as wanted criminals and warlords. An example of a well-known fugitive is *Yaser Abdel Said*, who is still missing and for whom FBI offers a high reward in exchange of valuable leads on his arrest. To demonstrate the benefits of a person tracking approach in social media, we performed a simple query in the Twitter Search API, namely, "Yaser Abel Said seen in". Only one tweet is returned by *NorthernMexico8* posted on November 2017 and as shown in Figure 4.1, it places him in Canada. We do not know whether that was his actual location, but it is commonly known that he has strong ties to Canada and Egypt, as well as Texas in the USA.

As we can observe, this basic test indicates the challenge of analyzing a limited amount of valuable data, yet the potential of tackling the (person) entity tracking problem via social network discussions. Note that, even when the available data is more, i.e., individuals are very popular and draw a lot of attention in the media, there are still important challenges to face. Particularly, the high amount of spam and fake messages makes it crucial to filter the data, in order to detect correct person locations and avoid any misinformation or chatter, e.g., false positives and farces.

Hence, our goal is to harvest the wisdom of the crowd that can potentially provide

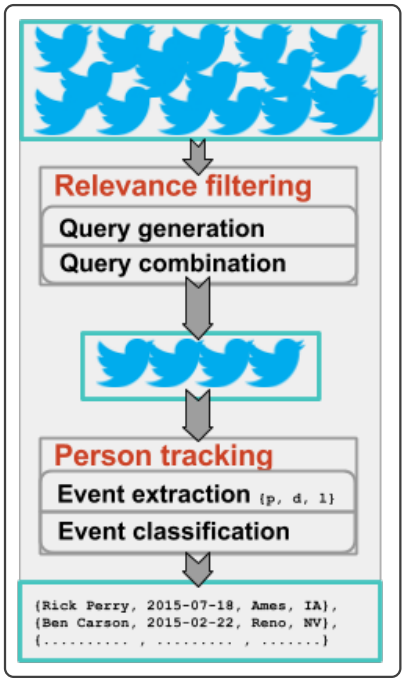


Figure 4.2: Our overall approach for person tracking in social media.

us with ongoing events, but at the same time we must make sure to avoid noisy tweets. In addition, users of such systems must take into account that even though tracking people that do not want to be found is useful in the case of criminals, locating other people, such as protesters and activists[2], might raise ethical concerns and have negative implications on the target individuals.

Unlike most prior work that mainly deals with location identification of social network users themselves [101], their home [129] or messages [157], we consider the tweets as a means to derive the physical position of mentioned individuals. We seek to answer the following question: Given a target person entity p, can we identify the locations of p over time only by observing what people say about p in social media? To address this problem, we analyze tweets that mention p to determine all p's physical locations and gain insights from a big tweet dataset spanning one year. Our approach uses millions of tweets relevant to the U.S. general elections in 2016 with the goal to track the presidential candidates and it considerably outperforms existing techniques and baselines. As illustrated in Figure 4.2, given a tweet stream, we apply a relevance filter that prunes the noise and supply the remaining tweets to a person tracker that outputs person locations. Hence, our contributions include:

[2] http://www.complex.com/life/2016/11/police-surveillance-activists-people-of-color

- A greedy two-phase algorithm that filters relevant tweets for person tracking
- A novel approach for predicting the locations of individuals by leveraging their third-person references in social media posts
- Evaluation results on tracing U.S. politicians' locations

4.2 Related Work

An extensive body of literature focuses on novel information discovery in user-generated content, such as news pieces, trending topics, popular events, etc. Related research includes *TwitterStand* [141], a framework that provides a geographic overview of breaking news on Twitter and *TwitterMonitor* [103], which performs real-time trending topic tracking. Topic models, such as Latent Dirichlet Allocation [15, 180], are also used to find for topic detection and hashtag recommendation on Twitter [51]. A recent and interesting work analyzes tweets for the topic of climate change and also detects users that are climate change deniers [26] with a feed forward deep neural network.

Since we are interested in detecting localized events and their timestamps, which we consider as the places a target individual visits, we find our work relating better to the task of event detection, rather than trend detection. An event usually appears as a bursty occurrence of novel information in a certain time period [1] and a sudden increase of the occurrence of certain words [55]. The attention that events attract typically fades over time as other significant incidents arise, e.g., in our case, the target entity moving to another location.

Event detection has been studied extensively for various application areas, e.g., predicting earthquakes [137], real-time discovery of sports competitions [1] and detection of event-related information [100]. However, prior work is mainly motivated by the need to keep the users up-to-date in emergency situations and few works identify and analyze events independently of their type [55]. The majority focuses on the cases of incidents of public interest [169] , e.g., natural disasters, instances of civil unrest, or disease outbreaks.

In contrast, we do not address the problem of event detection aiming at public awareness, but we solve the task of person tracking in social media. Given an individual as a user query, we show that social media can help us create a timeline of his/her locations. Each event in the timeline is independent from the others regarding its kind and duration, and the frequency of these events depends entirely

on the individual's profile. Hence, we are limiting our search to locations that these persons visit, yet at the same time we consider all possible types of events.

Another line of research that is closely connected to our work is the identification of locations in social media. Its emergence can be justified by the lack of geo-located user posts, especially on Twitter, since only 1% of the messages includes geo-tags [144], which might be totally irrelevant to the locations mentioned in the text. A recent work is using deep learning and multi-view learning to detect user geo-location on Twitter [42]. Related works also include *PETAR* [92], a time-aware point of interest (POI) extraction system and *TWILOC*, which determines the location of a tweet based on various content and network features [62]. Backstrom et al. study the relationship of social and spatial proximity and use the network properties to predict the location of users [4]. It is shown that social data, such as the location of a user's friends, can enhance prediction performance.

Moreover, interesting studies in different domains identify the locations of individuals in multimedia data, e.g., videos. For instance, Liang et al. predict future person activities and paths in videos using deep learning techniques [94]. Zhao et al. also develop deep neural networks in combination with visual tracking systems that detect people's location and motions [181].

Unlike the above-mentioned works, we take into account tweets by various users that are published in a certain time frame, instead of performing a user-focused analysis [153]. Thus, we are not interested in geo-locating either a tweet or its user. Instead, we analyze the location and person mentions that are contained in tweets, in order to track the mentioned individuals.

Our goal is to gain insights about a discussed entity p and hence, we treat any potential tweet posted by p as all other tweets that share information about p in the third person. A representative example of a tweet we wish to discover is: "'History is made by the dreamers, not the doubters'. *Donald Trump* just now in *Des Moines*. #Politics @POTUS @realDonaldTrump @IvankaTrump @FLOTUS". This is an appropriate post for our task regardless the account that it originates from. By mining the textual content of such messages, we cope with the lack of geo-tags on Twitter, as well as with location inconsistencies. For instance, users might also share their thoughts about an event they attended earlier this day, which means that their current location is not identical with the event's location anymore. Thus, we choose to find locations in the content of the tweets instead, by applying a named entity linking approach [53].

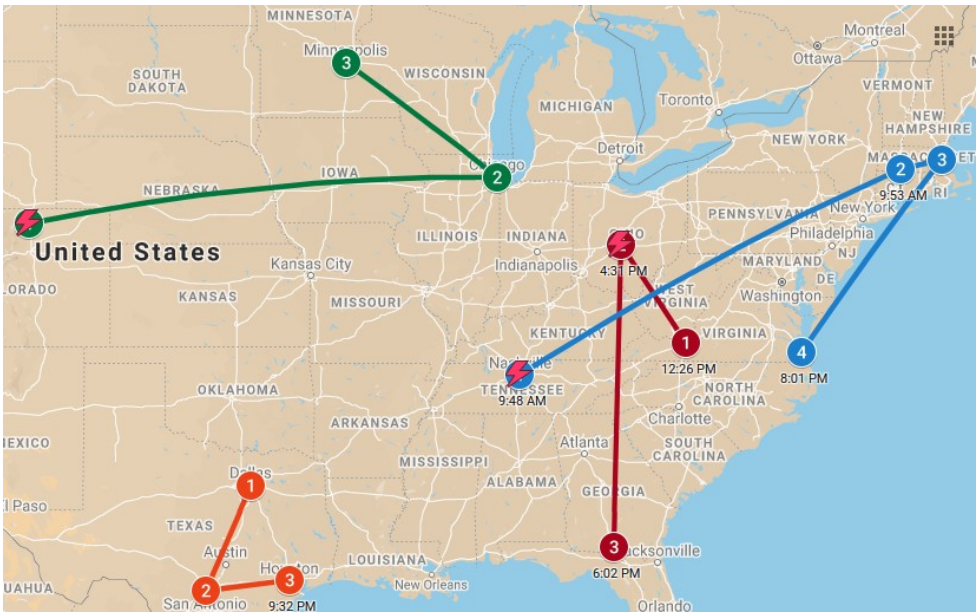


Figure 4.3: Trails of four presidential candidates extracted from Twitter on February 29th 2016. The red lightnings mark incorrect predictions.

The study most relevant to our work is our previously introduced basic approach that discovers naive "is-located-in" patterns in political tweets [54] and predicts the location of a person on a specific date, based on the relevant tweets of that date. Initially, we extract mentions of persons and locations in the text using Wikipedia, and then we consider the tweet text as a bag of words to discover possible "is-located-in" patterns. We apply the Apriori algorithm on a very small subset of our tweet dataset to discover frequent words that can be used as queries for the Twitter API to retrieve relevant posts to politicians' locations. These frequent term sets are then filtered based on the relative frequency among all locations of Donald Trump, Hillary Clinton, Bernie Sanders and Ted Cruz found in the tweets. The remaining frequent term sets are subsequently used to remove irrelevant tweets. Therefore, the final tweets can be interpreted as crowd-sourced textual indicators for the politicians' locations. The actual time of an event is estimated based on the tweet publication times. For each day in the dataset, we consider all tweet messages of a person-location pair and calculate the median of tweet times. For each day, we consider all tweets of a candidate-location pair and calculate the median of tweet times. Consequently, it is naively assumed that each event that yields a minimum support of ten or more tweets in the result set corresponds to an actual event. Figure 4.3 depicts the automatically retrieved trail of the above-mentioned four politicians two days prior to the Super Tuesday on Google Maps (the order is based on the estimated event time) – Donald J. Trump is depicted with red, Hillary Rodham Clinton with

blue, Bernie Sanders with green, and Ted Cruz with orange color. The numeric order is based on the estimated event time. We mark with a red flash sign the incorrect predictions. The two errors of the democratic trails (Nashville and Fort Collins) stem from campaign events from the previous day (February 28th), whereas Donald Trump actually visited Ohio one day later.

Drawing inspiration by these preliminary findings [54], we perform a large-scale analysis [85] on almost a billion tweets and various events and individuals. We introduce a novel approach for noise detection in the context of person tracking, which is based on recursive partitioning and carefully generates higher quality queries than our previously proposed method. Instead of solely relying on the popularity of the mentioned events, we use a supervised constraint-based approach to detect which of the event locations are valid.

4.3 Finding the Needle in a Haystack of Tweets

The first part of our person tracking approach is responsible for excluding noisy messages, which provide misleading information about the target entities and their associated events. We define an *event* as a triplet $e = (p, l, d)$, where an individual p appears in a specific location l on a particular date d. We model our noise detection task as an information retrieval task: given the tweets published in a certain time period, we wish to retrieve the ones that are relevant to person tracking. That is why we design a query for the Twitter API that will return suitable messages for our goal. Given the result set, we detail how we classify the discussed events into correct and incorrect in Section 4.4.

One can easily grasp that the terms {*rally*} or {*rally, today*} might be promising choices if one is searching for political campaigns in social media. However, given the almost infinite amount of words and hashtags that one can search with, choosing the right query is a cumbersome and complex task. The appropriate query terms depend on how users like to describe the locations of others, such as “live in”, “don’t miss the”, or “just saw”. Since the phenomenon of misinformation in media has risen in the past years [110], a naive query might return tweets that are fake or spam regarding the target entities.

Figure 4.4 depicts the number of tweets we found for four popular entities on four randomly selected dates: the U.S. politicians Donald Trump and Hillary Clinton, and the bands U2 and Red Hot Chili Peppers. The number of tweets that simply refer to an entity p is shown in *blue*, while the portion of them that contains a reference to an actual event location of p is depicted in *red*. The events are public

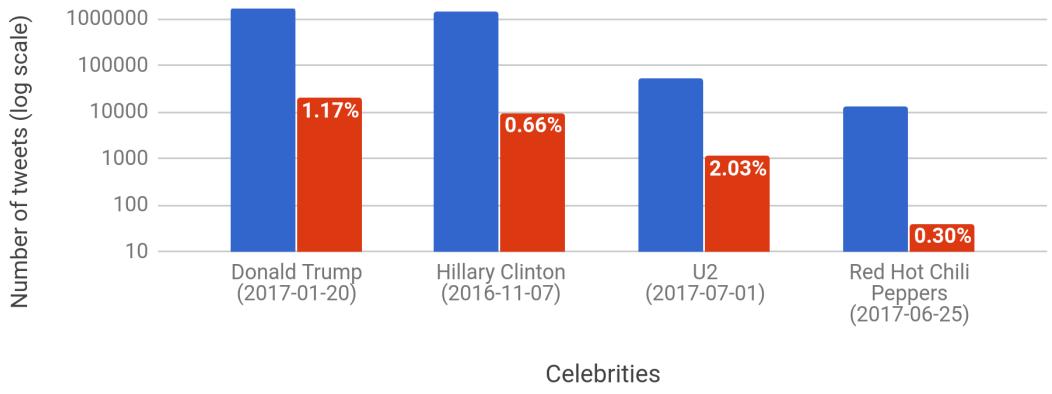


Figure 4.4: Tweets about a public figure (left) and its correct location (right).

speeches and concerts respectively. Although we depict a limited data sample[3] that is often biased by the medium's sampling process [140], we can already observe that irrelevant tweets are orders of magnitude more common than relevant ones are. Therefore, noise detection becomes an important, often domain dependent problem and a person tracking method is expected to distinguish which context provides correct person information and which not.

4.3.1 Problem Statement

Given a set of individuals we wish to detect, let us assume that the set of Twitter statuses mentioning at least one entity is denoted as T. Each tweet $t \in T$ represents a document that consists of a set of words, s.t., $t = \{w_1, w_2, ..., w_i\}$. All unique words in T form the existing vocabulary V. We aim to discover the tweets $T^+ \subset T$ that contain relevant information for our task. We refer to the rest of the tweets as $T^- \subset T$, where $T = T^+ \cup T^-$ and $T^+ \cap T^- = \emptyset$ hold.

A relevant tweet $t \in T^+$ is a message that refers only to correct event information, that is, contains an actual event triplet, $e = (p, l, d)$. For instance, during the U.S. election campaigns, the current U.S. president Donald Trump conducted a rally in Georgia on 29/2/2016. Thus, the tweet "LIVE Stream: *Donald Trump* Rally at Valdosta St. University in *Valdosta, GA*" belongs to T^+, whereas "It's Leap Day 2016. February has 29 days. And *Washington* is in an uproar. *Donald Trump* is trying to have the extra day deported" belongs to T^-. The second example is a tweet that refers to a false location of Donald Trump for that date and our approach makes it feasible to detect it, since it learns the context that individuals' locations are likely to be discussed on Twitter.

[3]The Public Streaming API is limited to a maximum of 1% of the overall traffic on Twitter (i.e., around 5 million tweets per day).

function FIND_CANDIDATES(T, V, P)
  $\Omega = \emptyset$
  for p **in** P **do**
    for i **in** $1..\lfloor\sqrt{|V|}\rfloor$ **do**
      $V_i = fold(i, V \setminus p)$
      $\Omega = \Omega \cup$ PARTITION($T, V_i, \{p\}$)
  return Ω
function PARTITION(T, V_i, q)
  $\Omega = \{q\}$
  if $|q| < \theta_{len}$ **then**
    $w = \arg\max_{w \in V_i}$ IG(q, w)
    if $\chi^2(T_q, w)$ **then**
      if $\frac{\left|T^+_{q \wedge w}\right|}{\left|T_{q \wedge w}\right|} > \frac{\left|T^+_{q \wedge \neg w}\right|}{\left|T_{q \wedge \neg w}\right|}$ **and** $\left|T^+_{q \wedge w}\right| \geq \theta_{supp}$ **then**
        $\Omega = \Omega \cup$ PARTITION($q \wedge w, V_i \setminus w$)
      else if $\left|T^+_{q \wedge \neg w}\right| \geq \theta_{supp}$ **then**
        $\Omega = \Omega \cup$ PARTITION($q \wedge \neg w, V_i \setminus w$)
  return Ω

Figure 4.5: Our recursive candidate query discovery algorithm that prunes irrelevant tweets.

The Twitter API provides an interface for *Boolean queries*, where a query Q is a combination of terms $w \in V$ and boolean operators $\neg$, $\wedge$, and $\vee$. Our goal is to create a filtered tweet set T_Q, s.t., $T_Q \cap T^+$ is maximized and $T_Q \cap T^-$ is minimized. This optimization task can be reduced to the knapsack problem, which is known to be NP-hard. Given the fixed-size knapsack (queries allowed by the Twitter API), we aim to fill it with the most valuable items (most promising queries). Because the number of possible queries is exponential to size of the vocabulary $|V|$, it is not possible to enumerate them and select the best one. Therefore, it is not feasible to find an optimal solution in reasonable time.

To design a good query, we propose a greedy approach that is based on recursive partitioning. We generate Q in a disjunctive normal form. That is, Q is defined as an $\vee$-combination of queries, i.e., $Q = q_1 \vee q_2 \vee ... \vee q_i$, where each q_x is an conjunction of words or their negations, e.g., $q_x = w_1 \wedge \neg w_2 \wedge ... \wedge w_j$. For instance, we discover that promising queries to trace politicians in the context of U.S. elections are "night $\wedge$ primary", "holds $\wedge in$" and "rally $\wedge \neg monday$". Our noise filtering algorithm consists of two phases: first, we discover promising conjuction

queries q_x that maximize the positive examples in T_{q_x} and second, we combine candidate conjunctions in a query Q. The retrieved tweets are further examined by our event classifier in Section 4.4.

4.3.2 Candidate Query Discovery

Inspired by the principle of boosting in machine learning, we construct a variety of term-conjunctions that are built on independent data portions. Figure 4.5 illustrates our approach for generating candidate queries, motivated by the principles of decision tree learners. Consider a set of pivot terms P (queries containing only one word) with a high coverage in T^+ (Line 3). Each seed term provides us with a high quality start, which propagates to the conjunctions that will be generated in the next recursive partitioning step (Line 6). For instance, let us assume the football player Luis Suarez and as pivot term the word *seen*. If *seen* is found in a high number of correct tweets (T^+) about Luis Suarez, e.g., "Just seen *#LuisSuarez* in Park Guell *#Barcelona*", this also increases the chances that the combination of *seen* and *in* would retrieve correct locations of the player.

Furthermore, for every pivot term, we split the vocabulary into $k = \sqrt{|V|}$ random and equally sized folds V_i (Line 4 and 5). In each iteration we expand the candidate query (that initially consists of p) with new terms from V_i. Note that every fold has the same size: $\forall i \in \{1, 2, ..., k\} : |V_i| \approx \sqrt{|V|}$, while $\bigcup_{V_i} = V$ and $\bigcap_{V_i} = \emptyset$ hold. The partitioning process (Lines 8- 17) works as follows: Assuming that the current q does not exceed the permitted length θ_{len} (Line 10), the algorithm expands it further. Although the length threshold is rarely hit, we adopt this constraint to prevent very long conjunctions that might lead to overfitting or conflict the restrictions of the Twitter API. Moreover, we perform a query expansion and select the term w (Line 11) that results in the highest information gain regarding the separation of the sets T^+ and T^-. We measure information gain as:

$$IG(q, w) = H(q) - \frac{\left|T_{q \wedge w}\right| * H(q \wedge w) + \left|T_{q \wedge \neg w}\right| * H(q \wedge \neg w)}{\left|T_q\right|}$$

where the *Shannon* entropy H(q) is defined as:

$$H(q) = -\sum_r \left(\frac{\left|T_q^r\right|}{\left|T_q\right|}\right) \log \left(\frac{\left|T_q^r\right|}{\left|T_q\right|}\right), r \in \{-, +\}$$

and the set T_x refers to the tweets that x satisfies. The expansion based on the information gain is inspired by the greedy feature selection of the $C.45$ decision tree

learner. It fits well to our task, because we leverage that the term conjunctions fulfill the monotonicity property.

Our overall goal is to distinguish between the vocabulary that users choose to discuss actual events (of the target entities) and the vocabulary in any other topic that is irrelevant to our task. Thus, in order to capture and successfully avoid words that typically appear in incorrect context, we allow either w or $\neg w$ to expand q (Lines 13 to 16).

For the purpose of avoiding overly specific queries that overfit the data associated to the current fold, we stop expanding when the improvement of w (or $\neg w$) over q is not statistically significant (Line 12). To quantify the significance, we consider the null hypothesis that q's application will not affect the distribution of T^+ and T^-. We perform a χ^2 test to test the hypothesis and reject it if it cannot be supported with the typical significance level of at least $\alpha = 0.05$. We also prevent the query expansions $q \wedge w$ (or $q \wedge \neg w$) to be too specific, by ensuring that the new partition yields sufficient support over T^+, denoted as θ_{supp}.

4.3.3 Candidate Query Combination

Armed with a valuable set of promising queries Ω, we now combine them to generate our final query Q in a disjunctive normal form that provides us with fewer noisy tweets for person tracking. Given Ω and our document collection T, Figure 4.6 illustrates our approach to greedily derive a good disjunction by maximizing the expected query quality *score*:

$$score(q, T) = \frac{\left|T_q^+\right|}{|T^+| + \left|T_q\right|}$$

It is evident that our *score* definition is proportional to the F-1 metric, given that T^+ is the set of relevant and T_q the set of retrieved documents. Therefore, COMBINE_CANDIDATES finds a local optimum for our problem.

Note that the number of possible combinations is exponential to the size of Ω and hence, enumerating all solutions is not feasible. If the maximum length of Q is reached or Q cannot be improved by adding further conjunctions $q \in \Omega$ (Line 6), the combination phase terminates. The monotonicity property of the disjunctions combined with the repeated improvement of the *score*, results in an extended query that covers a high number relevant tweets.

1: **function** COMBINE_CANDIDATES(Ω, T)
2: $Q = \emptyset$
3: **repeat**
4: $Q' = Q$
5: $Q = Q \vee \arg\max_{q' \in \Omega} score(Q \vee q', T)$
6: **until** $score(Q, T) > score(Q', T)$ **or** $|Q| > \theta_{len}$
7: **return** Q

Figure 4.6: Our query combination approach to minimize noise and maximize relevance.

4.4 Constraint-based Person Tracking

Given the relevant data we discover in Section 4.3, we can now address the question: How can one accurately extract people's locations by examining their references in social media posts? Inferring the places that individuals attend from social network discussions is a very challenging task. Realistic constraints should be taken into account, such as, any person cannot visit more than a reasonable number of locations per day, e.g., music artists usually schedule only one big gig per day, even during a tour. Additionally, many tweets are expected to talk about real-world events in contrast to incorrectly discovered events that won't dominate the online discussions. For instance, users share their experiences about various situations, ranging from popular global events (a concert by a famous band) to local community fairs that will most likely gain more attention in social than mainstream media.

We model this reasoning problem as a binary classification task and decide for each mentioned event on Twitter whether it is true or not. In order to ensure a good tracking performance, our constraint-based person tracking method leverages both the characteristics of the discussed events as well as the tweets themselves.

4.4.1 Event Extraction

Each discussed event $e = (p, l, d)$ in our tweet set T is associated with a person p, a location l and a date d. It is denoted as $e \in E^T$, while the messages about e are denoted as T_e ($T_e \subset T$). To infer the date of e_i from a tweet t that discusses e_i, we use t's publication date, inspired by the up-to-dateness of microblogs as Twitter [82]. Thus, we leverage the daily reactions on e_i, by considering asynchronous discussions about it within the course of a day. We leave more flexible temporal tagging for our future work.

We allow that a person can visit the same location on different dates and can appear in multiple locations on the same date. To identify l and p, we use a named entity

linking approach based on *CohEEL* [53] and apply it on the tweet text. Given a knowledge base, e.g., YAGO [154], CohEEL discovers potential mentions that are likely to be linked to a certain entity in the knowledge base. As a second step, the algorithm explores the entity graph derived from the knowledge base with a random walk approach and it extracts the final and coherent entity mentions.

We apply CohEEL with WIKIPEDIA and WIKIDATA (an open knowledge base) and extract from the tweets two different types of entities: persons (the target individuals) and locations (cities). We perform our analysis on a city level, that is, if an entity is found in n different city venues in n different tweets (various streets, buildings etc.), we map the venues to the appropriate city name and consider each of them as a visit to this particular city. Taking into account that CohEEL can also be used for other kinds of target entities (e.g., companies and organizations) and locations in different granularities can be allowed (e.g., states, countries), our approach is easily adapted to other tracking use cases.

4.4.2 Event Classification

After identifying the events mentioned in the tweets, we classify them into real or false event references using a number of features, inspired by the previously mentioned realistic constraints:

Popularity: The *popularity* or *prevalence* of an event on Twitter can be estimated based on the number of unique original tweets discussing about it (disregarding retweets). We refer to the popularity as $prev(e) = |T_e|$. This feature has already been proven as a good indicator for actual events in previous work [54]. For instance, on 31/8/2017 we found that the football player Cristiano Ronaldo was tweeted to be on a trip in the UK. There are more than 3 different tweets on that date all placing him in Manchester, as well as three others, about Tottenham, Longsight and Wolverhampton respectively. Thus, from a statistics point of view, Manchester seems more likely to be a true location.

Another interesting example is shown in Figure 4.7, which presents the city locations of the politician Jeb Bush during his South Carolina (SC) rally. The color indicates the number of tweets in a specific region. Despite the fact that many locations outside South Carolina are mentioned, the dominance of SC venues on Twitter gives a strong indication towards events in this particular region.

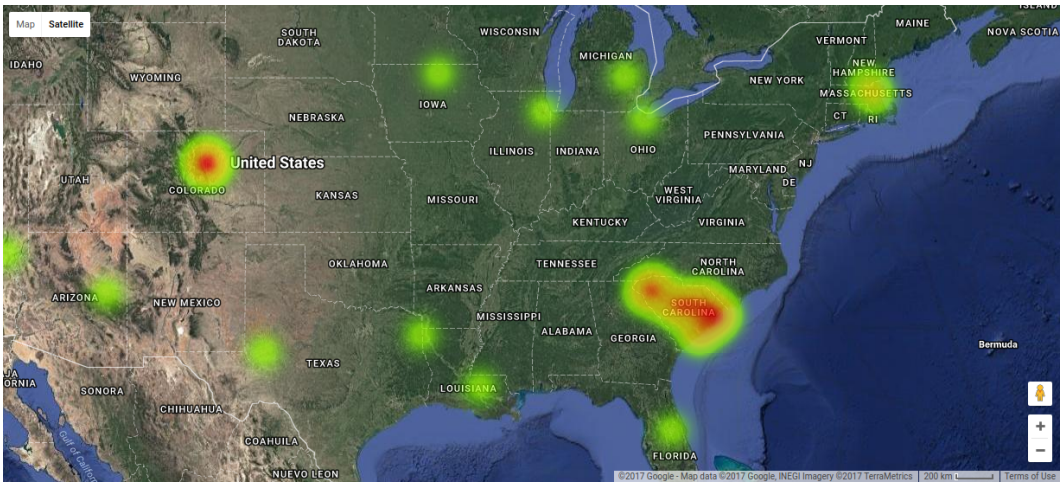


Figure 4.7: Heatmap of Jeb Bush's locations identified in tweets on 11/02/2016.

Distance: In the previously described example about Christiano Ronaldo, we observed that all tweets are published in a timeframe of only two hours, which raises the question of how far these three mentioned locations are from each other. Therefore, given all location mentions on a date, an event classifier should be able to understand how far an entity can travel within a certain time period.

We introduce the feature *distance*, i.e., the average pairwise distance of a certain location to the rest on a specific date. Similarly to the *popularity*, in a real-time experiment, this distance is updated as more locations are mentioned in newly published tweets. Moreover, each city is considered as a point on the earth and given its longitude and latitude, we calculate its Haversine[4] distance from the other cities. Namely, given the locations of a person p on a date d, the associated tweets are denoted as $T_{p,d}$. The events found in $T_{p,d}$ are defined as follows:

$$E^T_{p,d} = \left\{ e_i \in E^T \mid p_i = p, d_i = d \right\}$$

and the distance feature of an event e is:

$$dist(e) = \frac{1}{\left|T_{p,d}\right|} \sum_{e_i \in E_{p,d}} \left|T_{e_i}\right| \cdot \textsc{Haversine}(geo(l), geo(l_i))$$

By using this feature, we aim to preserve the events that are held reasonably close to each other and eliminate locations that are very far from each other. We extract the geo-locations of the cities from WIKIDATA: $geo(l) = \texttt{wd:P625}(l)$[5].

[4] https://en.wikipedia.org/wiki/Haversine_formula
[5] https://www.wikidata.org/wiki/Property:P625

Table 4.1: Twitter data extracted from November 2015 until January 2017.

All tweets	Tweets with entities	Correct tweets
903,239,572	29,208,457	321,530

Population: We hypothesize that the size of a location, such as the *population* of an event's city, can be indicative of whether this event is true or not. We test this hypothesis by including the city's population as a feature of our event classifier and expect that the popularity of a target individual might be correlated to the size of the locations he or she visits. The city populations are retrieved from WIKIDATA with the query: $pop(e)$ = `wd:P1082`(l)[6].

4.4.3 Datasets

We evaluate our approach for person tracking on a set of messages extracted via the Public Twitter API[7] during a period of approximately one year.

Tweets: Our dataset consists of millions of messages published by more than 33 million users. The posts mention various individuals related to the last U.S. presidential election (2016). In order to ensure a high coverage on discussions about political parties and their members, we used 241 queries with politicians' names and usernames, as well as popular hashtags related to the election. Since the language found in the Twittersphere can be eccentric, the queries we posed contain not only the individuals' names and Twitter user accounts, but also potential aliases (such as *Hillary Rodham Clinton*, *Secretary of State*, *@HillaryClinton* and *#HRC*), extracted from WIKIPEDIA.

As shown in Table 4.1, the overall amount of tweets we gathered from the Twitter API is approximately one billion. This results in an amount of 2 million tweets per day. Among this data, there are 29 million posts that discuss our target entities (contain mentions to presidential candidates and locations as discovered by *CohEEL* [53]). Furthermore, there are only 321,530 tweets revealing the actual locations of our target entities, i.e., mentioning a person's name and his/her actual location on the day the tweet was posted, which makes our task particularly challenging. The list of tweet ids for every discussed location and politician can be found in our homepage[8].

[6] https://www.wikidata.org/wiki/Property:P1082

[7] https://dev.twitter.com/streaming/public

[8] https://hpi.de/naumann/projects/web-science/social-media-analysis/politics-on-twitter.html

Events: To evaluate our approach for person tracking in social media, we collected a series of publicly available event records regarding the U.S. presidential candidates in 2016. Our ground truth is a set of events that presidential candidates hosted or participated prior and after the general elections extracted from the 4President blog [9]. This website contains information about events related to the past four elections in the U.S.A. We automatically extract the reported events related to the presidential candidates of the last general election in 2016. Many entries on the website are usually a single event, e.g., the title of the entry page explicitly refers to an event triplet, `person-location-date` (e.g., Donald Trump, Youngstown, Ohio, 25/7/17[10]).

In the cases where the title contains a broader location, i.e., a state[11], we apply *CohEEL* on the page's body text to determine the different cities within the state that a presidential candidate has visited. For the purpose of ensuring the validity of each event in our gold standard collection, any other blog entry[12] whose title does not describe an event triplet, namely `politician-city-date`, is disregarded by our extractor. The resulting gold standard consists of almost three thousand events for various candidates, such as Ben Carson, Lincoln Chafee, Chris Christie, Hillary Clinton, Lindsey Graham, Mike Huckabee, Rick Santorum, Jim Webb, etc.

4.5 Results

In this section, we evaluate our noise detector based on **Rec**ursive **Par**titioning (RecPar) and our **Co**nstraint-based **P**erson **T**racker (CoPT). RecPar leverages the wisdom of the crowd to discover relevant tweets and CoPT categorizes them into true or false person locations. First, we show the optimal setup of RecPar and compare it with our previously introduced approach that is based on **Freq**uent **Item**sets (FreqItem) [54]. Second, we demonstrate results on person tracking and compare CoPT with other approaches and baselines. In general, we conduct our experiments in consecutive monthly time intervals, namely we use the earliest months of our dataset to learn RecPar's query, afterwards we train CoPT, and in the last part of the dataset we test the performance of the overall approach.

[9] http://blog.4president.org/

[10] http://blog.4president.org/2020/2017/07/president-donald-j-trump-to-hold-rally-at-the-covelli-centre-in-youngstown-ohio-on-tuesday-july-25-2.html

[11] http://blog.4president.org/2016/2016/01/dr-ben-carson-visits-iowa-on-monday-january-11-2016.html

[12] http://blog.4president.org/

4.5.1 Relevance Filtering

RecPar consists of two consecutive phases: candidate query discovery and candidate query combination. The set of candidates, denoted as Ω, is generated by the first component and is also referred to as conjunctions or subqueries of the final query combination Q. In the current evaluation task, we show how RecPar behaves with different parameter settings. There are four parameters that we must consider in our approach:

- the maximum length of the final combined query Q (`maxQLen`)
- the maximum length of each subquery in Ω (`maxSubQLen`)
- the minimum support –number of tweets– that a subquery should exhibit to be included in Ω (θ_{supp})
- the minimum pivot support (θ^{p}_{supp}) that determines which terms will be the pivots

An example combination of the first two parameters could be a setting where `maxQLen` = 100 and `maxSubQLen` = 10. Herewith, RecPar would create a query Q with at most 10 subqueries, whose length will be 100/10=10 at maximum. For instance, a query combination that tracks art exhibitions of Picasso could be $(must \wedge see \wedge art \wedge exhibition \wedge Picasso) \vee (don't \wedge miss \wedge art \wedge work \wedge Picasso) \vee (interesting \wedge exhibition \wedge inspired \wedge by \wedge Picasso)$. Both parameters are influenced by the restrictions of the Twitter API, yet affect RecPar's performance as well. In a real-time setting, our system would query the Twitter Streaming API with the target's name and meaningful keywords, and as the tweets arrive, it would categorize each mentioned event as true or false. Thus, we take into account that as of today, the Twitter API allows searches with at most 400 terms. This means that at least one of the query terms needs to be the name of the target person (or its variants) and the rest will be generated by our model.

Our intuition is that the more queries we allow our model to generate, the better the chances to capture more helpful tweets. In contrast, experimenting with different `maxQLen` values (i.e., 100, 200, 300 and 400) showed that this aspect influences our final event classification results only up to approximately 1%! We conclude that selecting promising and relevant queries is more essential than their number. Hence, in all our experiments `maxQLen` is set to its potential maximum, i.e., 395,

Table 4.2: Precision and true negative rate of RecPar after the candidate query generation phase (Ω) and after the candidate query combination phase (Q) respectively.

θ^{p}_{supp} (%)	θ_{supp} (%)	PREC		TNR	
		Ω	Q	Ω	Q
10	0.25	0.133	**0.523**	0.019	**0.928**
5	0.25	0.133	0.515	0.012	0.926
1	0.25	0.132	0.516	0.002	0.924
0.50	0.25	0.132	0.513	0.002	0.923
0.25	0.25	0.132	0.510	0.001	0.921
10	10	0.133	0.519	0.019	0.931
10	5	0.133	**0.530**	0.019	**0.937**
10	1	0.133	0.500	0.019	0.927
10	0.50	0.133	0.512	0.019	0.927
10	0.25	0.133	0.514	0.019	0.926

leaving five terms to contain the person's name or alias (e.g., nickname) we aim to discover.

We also examine different values for `maxSubQLen` (between 2 and 20). Similarly to `maxQLen`, the results were not significantly affected for values higher than 5. Assigning a small number to `maxSubQLen` seems logical if we consider that tweets are limited to 140 characters, among which the name of the target person and a location have to appear. Therefore, we chose to set `maxSubQLen` to 5 for the rest of our experiments.

Support thresholds: As shown earlier in Figure 4.4, the number of tweets mentioning real-world events is extremely low, i.e., below 2% of all tweets. Thus, we experiment with low values for θ^{p}_{supp} and θ_{supp} and define these two thresholds as a percentage of the correct tweets in our training set. We train RecPar's query with the first 3 months of our dataset (2015-11-01 – 2016-01-31) and use the next month (2016-02-01 – 2016-02-29) as a validation set to optimize the parameters. The results are shown in Table 4.2. The maximum depicted values for θ^{p}_{supp} and θ_{supp} are 3,686 tweets (i.e., 10%), given that there exist 36,868 positive examples (out of 2,011,085) in our training set. Note that we exclude the messages that refer to multiple persons and locations as it is not clear how to assign one of the locations to one of the persons. Examining the word order in the text with the help of a syntax parser is a challenging problem and we leave this task for future work.

Initially, θ_{supp} is set constant and θ^{p}_{supp} is decreased, and then vice versa. By setting the pivot support higher than the overall support, we aim to be strict with our seed set so that limited ensemble models are created. The first conclusion we draw is that both thresholds affect RecPar's performance, but not drastically. For instance, in a strict setting where a pivot term has to appear in least 3,687 tweets (θ^{p}_{supp}=10%), the precision and the TNR are improved only by approximately 1% in comparison to the softest constraint (θ^{p}_{supp}=0.25%). Similarly for the θ_{supp}, its second highest value achieves the most successful result.

Another interesting finding is the crucial contribution of the candidate combination phase to RecPar's performance. It is evident that the naive usage of all subqueries would achieve poor precision results (first column under PREC). The reason behind this is that RecPar's first phase is recall-oriented and the candidates of this phase accomplish 95-99% True Positive Rate (TPR) and False Positive Rate (FPR). However, the combination phase improves the precision by a factor of 4 and the TNR by more than an order of magnitude. Additionally, our experiments show that the second phase diminishes the FPR and boosts the TNR significantly, leading to fewer noisy and irrelevant tweets in our dataset. To conclude, for the rest of our study, we use RecPar's best query combination, which is learned in 2015-11-01 – 2016-01-31 with θ^{p}_{supp} = 10% and θ_{supp} = 5%.

Comparison between filtering approaches: This tweet-based experiment is an intermediate evaluation of our overall approach, before the evaluation of the event discovery. We measure how many of the remaining tweets after the filter are correct (i.e., refer to real events). We compare against the previously introduced approach for person tracking [54]. Similarly to RecPar, we apply FreqItem's query to every tweet t in the test set and if t satisfies it, then we classify t to the correct class. We expect FreqItem to perform poorer than RecPar, due to the fact that it is trained with a very small set of correct tweets and because it does not support negative predicates ($\neg w$).

Furthermore, the recursive nature of RecPar and the higher diversity of its query candidates, originating from independent data partitions in the generation phase, should lead to better queries. In contrast, in this work, we leverage millions more tweets and anticipate that the recursive nature of RecPar will dominate the naively constructed queries of FreqItem. The test set for both approaches is March 2016 (subsequent to RecPar's validation set).

As depicted in Table 4.3, RecPar prevails in terms of precision, F-1 measure,

Table 4.3: Comparison of RecPar to FreqItem

Model	PREC	REC	F-1	ACC
RecPar	0.48	0.47	0.47	0.83
FreqItem	0.25	0.60	0.35	0.64

Table 4.4: Comparison of RecPar+CoPT to the variations RecPar+PoPT, CoPT, PoPT, the tracking approach FreqItem+Po [54], a naive baseline Po and the event detector MABED [55]

Approach	PREC	REC	F-1
RecPar+CoPT	**0.68**	0.43	**0.53**
RecPar+PoPT	0.64	0.37	0.47
CoPT	0.32	0.24	0.28
PoPT	0.17	0.23	0.19
FreqItem+Po	0.15	0.67	0.25
Po	0.01	**0.87**	0.02
MABED	0.14	0.00	0.00

and accuracy, since it generates more sophisticated and carefully designed queries, which guarantee that the result set will contain more relevant than irrelevant tweets. However, FreqItem achieves a higher recall, because it generates a very high amount of naive queries and hence many relevant (and irrelevant) tweets are covered by it. Note that this is not necessary to find a person's location though. Namely, a small and relevant subset of tweets is enough to correctly locate a person.

4.5.2 Person Tracking

We now evaluate our constraint-based approach (CoPT) on the promising filtered tweets. We initially show the necessity of our realistic constraints (*population*, *popularity* and *distance*) by comparing our proposed solution RecPar +CoPT to RecPar +PoPT (**Po**pularity-based **P**erson **T**racking), which considers only the *popularity* of an event on Twitter. We use a Random Forest classifier for both approaches. Our goal is to see whether this obvious and simple constraint is adequate to retrieve the locations of the target individuals.

In order to show the filter's necessity, we compare against CoPT and PoPT without filtering the tweets. As discussed earlier, our previous technique [54] applies FreqItem at first and then it assumes that each event that yields a *popularity* score higher than 10 corresponds to an actual event. We refer to this person tracking approach as FreqItem+Po (**Po**pularity) and we also compare simply against Po, as

Table 4.5: Monthly precision for all person location detectors in 2016

Approach	June	July	Aug.	Sept.	Oct.	Nov.
RecPar+CoPT	**0.62**	**0.66**	**0.66**	**0.67**	0.72	**0.80**
RecPar+PoPT	0.56	0.64	**0.66**	0.60	**0.80**	0.74
CoPT	0.14	0.26	0.45	0.34	0.31	0.47
PoPT	0.16	0.13	0.21	0.17	0.19	0.17
FreqItem+Po	0.13	0.13	0.16	0.14	0.21	0.15
Po	0.01	0.01	0.01	0.01	0.01	0.01
MABED	0.10	0.00	0.06	0.14	0.20	0.33

a naive baseline.

We train the above-mentioned models with events from April and May 2016 and test them monthly in a six-month period prior to the general elections in the US (from June till November). Various evaluation metrics are shown in Table 4.4, computed as an average of all test sets. RecPar+CoPT outperforms almost all techniques and competes closely to its variation, RecPar+PoPT, especially in terms of precision. That is, the *popularity* of a discussed event in social media is a very strong indicator about its validity, but not enough on its own. The importance of the RecPar phase is also evident, since CoPT and PoPT cannot outperform our overall proposed approach. Moreover, FreqItem+Po and Po achieve higher probability of detection (REC) than RecPar, due to their simplistic nature.

As time goes by: Multiple events related to our target persons happened prior to the US general elections[13], e.g., primaries/caucuses in June, e-mail leakage in July and October, the Green National Convention in August, the first presidential debate in September, etc. In order to explore how the models work on each occasion, we show the monthly precision values in Table 4.5. We see that our person tracker outperforms all competitors, while having similar results to RecPar+PoPT for certain tests sets. For instance, in August 2016, the two techniques perform the same and in October 2016, RecPar+PoPT outstrips RecPar+CoPT.

The performance of RecPar+PoPT increases in October 2016, which is the month with the highest amount of published tweets (i.e., 3,556,464 messages) in our dataset, considering that the election date was on November 8th 2016. Thus, we assume that the number of published tweets enhances significantly the performance of this model.

[13] https://en.wikipedia.org/wiki/United_States_presidential_election,_2016_timeline

However, our proposed solution RecPar+CoPT appears to be more consistent and robust, by always achieving a minimum precision of 60% and maintaining satisfying recall and f-1 scores levels, as depicted in Table 4.4 as well.

Person tracking as event detection: One can argue that tracking the locations of mentioned entities in social media is a problem that can be tackled by an event detection algorithm. We hypothesize that existing literature on event discovery will not be as successful for our task, since the works are not focused on the involved individuals and thus, they will discover other events in our dataset that the target entities did not attend. To verify our intuition, we consider another competitor, namely MABED, a mention-anomaly-based event detection algorithm [55]. MABED leverages the creation frequency of dynamic mentions to discover events. Noise is avoided by allowing fine-tuned and dynamic events, which do not have to fit to a predefined time duration. This setting serves as a helpful noise "filter", given our highly imbalanced dataset.

An event is defined in MABED by a starting and ending date, a main keyword, and a set of related terms. We are looking for person and location mentions in these keywords by applying *CohEEL* and we use the event timeframe to create event triplets. As long as the detected event exists in our ground truth, we consider it a true positive. In addition, the system is user-parametrizable and we tune it appropriately for our task. Namely, after experimenting with different parameter settings, we set the time window to 120 minutes to allow medium time precision and the number of words describing an event to 10. Increasing this number did not improve our results, because the longer the event summary is, the more are the chances that multiple politicians and locations are included in it and our evaluation setting does not allow such cases (as discussed in Section 4.5.1). The threshold for selecting relevant words is the default one (0.6). Since we perform monthly experiments and the most popular month in our dataset contains 400 events, we set k (the maximum number of returned events in MABED) to 400.

Unsurprisingly, we can see in Table 4.4 that MABED is not performing well, specifically it is unable to capture almost any event in our ground truth and it achieves similar precision to PoPT and FreqItem+Po. We observed that MABED can generally capture the political discussions and oftentimes, there exist mentions of presidential candidates and U.S. cities in the event descriptions. However, at least one item in the discovered event triplets (`person-location-date`) is usually incorrect and thus the triplet does not refer to an actual location that a person visited

on a certain date. This confirms our hypothesis that event detection models are not designed for predicting the precise locations of people mentioned in social media. The results are also not as consistent as of other models, e.g., there are no true positives discovered by MABED in July 2016, as shown in Table 4.5.

4.6 Summary and Findings

In this chapter, we tackle the problem of person tracking via online discussions in social networks. We show that social media posts reveal more than the obvious and they make it feasible to discover which places the discussed individuals visit and when. Our motivation stems from the vast amount of political information in social media discussions. We leverage the wisdom of the crowd in the context of politics, specifically the richness and up-to-dateness of the shared information. Our proposed approach extracts facts from tweet text and it could be applied to any domain whose entities move over time. The problem we study has several applications, such as detecting singers' concerts, politicians' speeches, companies' relocations, sport teams' games etc., but also in emergency situations, one can identify mentions of missing persons or any kind of threat, such as, fugitives, criminals etc.

We introduce RecPar, a recursive partitioning algorithm, which carefully generates queries for the Twitter API that return relevant information to the target entities and their locations. An extensive experimental analysis is conducted to examine RecPar's behavior and optimize its input parameters. We also propose a constraint-based person tracking approach (CoPT), which reasons over the filtered tweets and categorizes the mentioned events as true or false. Social media as well as location characteristics were used to classify the events. Our overall person tracking method (RecPar + CoPT) outperforms the previously introduced tracking technique [54], the event detection algorithm MABED [55] and multiple baselines.

4.7 Topic Analysis

In this section, we discuss an additional task we perform on our tweet dataset. Namely, the topic detection in online political debates by Jaqueline Pollak in the context of her master thesis[14]. This work analyzes our one billion tweet collection and also considers an additional political dataset, namely the transcripts of the presidential debates in 2016 in the USA[15], which we have annotated for their topics. Thus, we introduce a new annotated dataset of political speeches and another

[14] https://hpi.de/naumann/projects/completed-projects/politics-on-twitter.html

[15] https://www.kaggle.com/mrisdal/2016-us-presidential-debates

annotated tweet dataset that we test our approaches on. We also perform an initial analysis to discover insights about political discussions in social media, e.g. topics and sentiments in the text. This work is inspired by the vast of amount of shared thoughts in microblogs, whose analysis can help us understand the public opinion and the users' decision making in the context of politics.

4.7.1 Related Work

Similarly to our previous work in this chapter regarding person tracking [54, 85], this study also lies on the intersection of several research problems in social media mining, i.e., discovering topics and sentiments in social text messages, and dealing with the challenges of chatter [6] and bots [22] in such data. We have previously referred to many works focusing on finding breaking news and topics on Twitter [103, 141], since this is a very useful and widely studied task. For instance, recent work uses a combination of word embeddings and topic modelling to find local topics on Twitter [24]. Political tweets have also been receiving particular attention with studies that detect racism [98] in the ego networks of Hillary Clinton and Donald Trump. As mentioned in Chapter 3, identifying political perspectives in blogs and social media is also a related research topic that gained attention in the last decade [175]. Moreover, users express freely their opinions and emotions on Twitter while discussing political affairs, with several sentiment detection techniques focusing on predicting election results [13, 160]. Our work uses existing algorithms in these fields to propose a solution for finding topics from a given taxonomy on Twitter in the context of politics.

4.7.2 Datasets

As mentioned earlier in Section 4.4.3, the tweet collection is obtained by the Public Twitter Streaming API from November 2015 until April 2017. We use 241 queries and we retrieve half a million to 4 million tweets per day (there are high peaks on the election and inauguration dates). The most retweeted user accounts are, among others, the accounts of Donald Trump, Hillary Clinton, CNN and FoxNews. On the evening of the first presidential debate, Donald Trump is quoted [44] three times as much as Hillary Clinton and we also observe a high quotation amount in our dataset after the first presidential debate, with users quoting the statements the candidates made earlier. Regarding the debate transcripts and our annotations, the questions asked by the moderator in each debate were predefined and announced in advance on Wikipedia[16]. We watched the recordings of the debates and annotated every

[16] https://en.wikipedia.org/wiki/2016_United_States_presidential_debates

Table 4.6: Format of the US presidential debates in 2006 based on Wikipedia and the moderators' instructions. We merge/rename all topics in the last column. All dates are PM (after midday).

Date	**Debate topic**	**Unified Topic**
26-09-2016	**First presidential debate**	
9.02–9.23	Economy	Economy
9.23–9.41	Trade	Economy
9.41–10.03	Race relation	America's direction
10.03–10.24	War on terror	Foreign politics
10.24–10.30	Foreign policy	Foreign politics
10.30–10.36	Candidates experience	Candidates
09-10-2016	**Second presidential debate**	
9.02–9.13	Model behavior	Candidates
9.13–9.26	Donald Trump's behavior	Candidates
9.26–9.34	Health care	Inner politics
9.34–9.44	Islamophobia	America's direction
9.44–9.49	Wiki leaks	Candidates
9.49–9.59	Taxes	Economy
9.59–10.08	Syria	Foreign politics
10.08–10.10	Military forces	Foreign politics
10.10–10.21	President capabilities	Candidates
10.21–10.26	Supreme court justice	Inner politics
10.26–10.30	Energy policy	Inner politics
10.30–10.34	Candidates' characteristics	Candidates
19-10-2016	**Third presidential debate**	
9:01–9.16	Supreme court justice	Inner politics
9.16–9.29	Immigration	Foreign politics
9.29–9.32	Hacks	Foreign politics
9.32–9.48	Economy	Economy
10.06–10.20	Candidates' fitness	Candidates
9.48–10.06	Foreign hot sports	Foreign politics
10.20–10.32	Debt and entitlements	Inner politics

reply of the contestants based on the topic of the given question.

The resulting taxonomy of the topics can be seen in Table 4.6. In the *Economy* topic, frequent keywords include: debt, economy, jobs, pay, percent, plan, trade, work and deal. In the *America's direction* category, we often observe the words war, police, NATO, communities, believe, muslims and in the *Candidates* category, we see the terms things, campaign, never, last, emails, and election. The transcripts regarding

Foreign politics include frequent words, such as Russia, isis, mosul, Syria, Putin, Iran, Aleppo, border, Assad and the *Inner politics* texts discuss more the words court, second, amendment, supreme, insurance, health, care, energy, justice and Obama care. We propose several machine learning classifiers to categorize the tweets in these five topics, as well a noise classification classifier to remove irrelevant information to the topics. In addition, we have also manually annotate a set of 1,500 tweets for their topics in order to evaluate our approaches.

4.7.3 Topic Classification

After applying tokenization, stop-word removal and stemming to the tweets, our first task is to filter the noise (chatter) in our dataset. We define noise as the tweets that do not discuss any of the five given political topics in Table 4.6 (see "Unified topic" in the last column). We annotate 200 tweets for their noise and train a Random Forest classifier for this purpose with cost-sensitive learning due to the class imbalance in this dataset (we experimented with other traditional supervised learning methods and this performed the best). We use several features for this task, including textual features (tweet text, retweeted text, quoted text), meta text information (number of URLs, number of hashtags, etc.) and user information (e.g., ratio of followers and followees), and after calculating the importance for these features, we only consider the ones with sufficient information gain. We evaluate the noise classifier on our gold standard test set (1,500 tweets with 70% noise) and achieve a precision of 0.38, recall of 0.82 and F-1 score of 0.52, outperforming a random decision which would result to a precision value of 0.33, recall 0.50 and F-1 score 0.40.

After applying the noise detection model to our tweet corpus, we proceed to detect the underlying topics in the remaining tweets. Initially, we experiment with Latent Dirichlet Allocation (LDA) algorithm as a baseline. We try several settings to train the topic model (e.g., sentences of the presidential debates, Wikipedia articles, tweets) and test the resulting models on our annotated tweet dataset. The results are not satisfactory, with the best model (trained on the debates) being able to detect only some of the five political topics (e.g., the *Foreign politics* topic with frequent terms discovered by LDA being Iraq, Isis and Syria).

Moreover, we propose and compare three supervised learning models for topic detection on Twitter based on a given taxonomy using three different training sets. First, we build a *Debate classifier*, trained on the debate sections (a section starts every time the speaker changes), which are represented with a bag of words (we also experimented with an LDA representation, but it performed poorly). Then, we

Table 4.7: Multi-class topic classification results of our three different topic detection approaches on Twitter. The annotated test set contains 456 tweets, excluding 1,044 noisy tweets.

Method	Precision	Recall	Marco F-1	Micro F-1
Debate classifier	0.78	0.33	0.32	0.47
Tweet classifier	0.54	0.41	0.4	0.45
Improved tweet classifier	0.58	0.47	0.48	0.52

build a *Tweet keyword-based classifier*, which is trained on an silver standard tweet dataset. Namely, given the frequent terms of each topic in the debate transcripts, for each of the five topics, we consider tweets to belong to this topic, only if they contain at least one of the frequent keywords of this topic and none of the other topics. In this way we constructed a set of 60,000 tweets and we utilize all above-mentioned textual and meta features to represent this training data. Our last model, *Tweet keyword and time-based classifier*, takes advantage of the time dimension in the tweet and debate datasets. That is, our training data for each topic consists of the tweets that were posted only during this topic discussion in the respective debate and also satisfy the previous keyword-based criterion of the *Tweet keyword-based classifier*. This training set is smaller than the previous one, but more balanced in terms of the topic distribution. All three classifier are multi-class Random Forest algorithms that are tested on the remaining tweets of our annotated dataset after removing the noise (456 tweets).

4.7.4 Results

The tweet classification results are shown in Table 4.7. The *Debate classifier* achieves a high precision and a low recall value, as it performs very well for only two classes, namely *America's direction* and *Inner politics*. The micro F-1 score, which is more appropriate for multi-class classification problems than the macro F-1 score, has the highest value for the *Tweet keyword and time-based classifier*, with the other two methods performing similarly to each other. We believe that the combination of the keyword-based and time-based approach to create a training set automatically looks very promising for such a challenging and multi-class task that essentially aligns the topics of two datasets. There is certainly still room for improvement and we believe that it can also be applied in other use cases, e.g., finding discussed topics in social media during sport championships or the Olympic games.

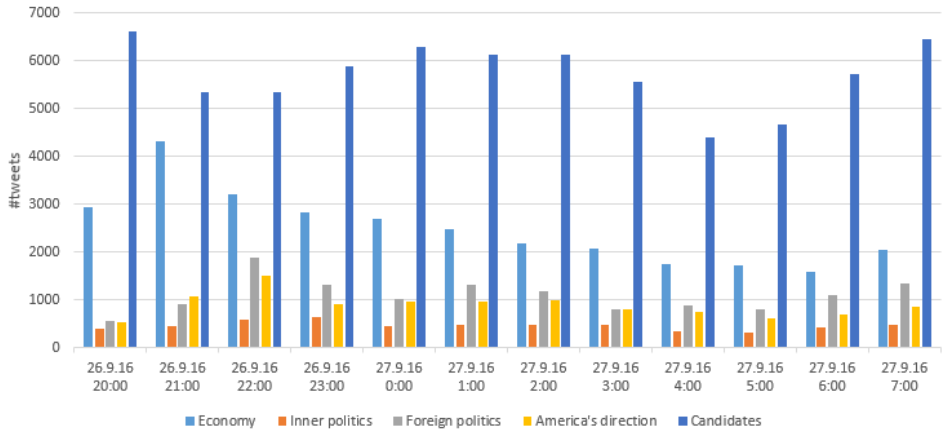


Figure 4.8: Topic development on Twitter as discovered by our *Tweet keyword and time-based classifier* on 26-09-2016.

Further insights about politics in the USA

We explore our tweet dataset further and try to discover insights for specific dates. For instance, in Figure 4.8, we apply the *Tweet keyword and time-based classifier* and visualize the topic distribution on Twitter during the evening of the first presidential debate (26-09-2016, 9:00-10:30 PM). We can observe that the candidates' capabilities is the most common topic in the Twittersphere, with the *Economy* topic reaching a pick during the time-frame that it was also discussed at the debate (first half of the debate). Tweets about *America's direction* and *Foreign politics* are also increasing in number during the time these topics were addressed at the debate (second half).

Considering the large amount of publicly shared opinions in this political dataset, we are also interested in other dimensions of the data apart from the discussed topics, for instance the sentiment expressed in the text. We use SentiStrength [156] to detect the underlying sentiment in the tweets, a sentiment detection algorithm that performs well for short social messages with casual writing style. It provides a positive and a negative value for a given text, with each score being in the range of 1 (no sentiment) to 5 (strong sentiment). Similarly to related work [80], we consider a tweet to be positive if the positive score by SentiStrength is higher than the negative score and vice versa. If both sentiments are equal, we regard this tweet as ambiguous and potentially neutral. Based on the two sentiment scores, we define an additional metric, namely *polarity*. A tweet is polar when the sum of the positive and negative score is equal or higher than 4. For instance, the following tweet is polar: "RT @realDonaldTrump: Loved the debate last night, and almost everyone

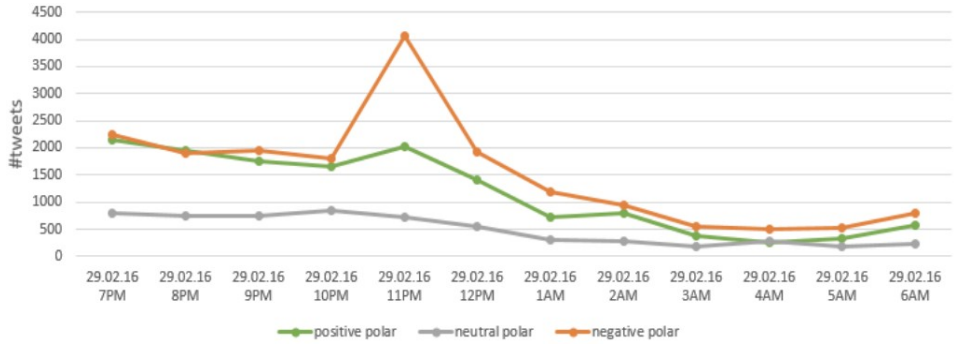


Figure 4.9: Sentiment development on Twitter as discovered by SentiStrength on 29-02-2016.

said I won, but the RNC did a terrible job of ticket distribution. All. . .". We use this metric to examine tweets with loaded language in our dataset.

In Figure 4.9, we see the sentiment development of the polar tweets that were published on the evening of the 88th Academy Awards (Oscars). It is interesting to see a high peak of negative and polar tweets at around 11:00 PM, which was the time that the politician Joe Biden and the singer Lady Gaga announced that they have joined a movement against sexual assault. Twitter users commented on these announcements and regardless of their expressed support or opposition, SentiStrength labeled their tweets as negative and polar, because the negatively sentimental words "assault" and "violence" were discussed in the text. Moreover, we have also calculated the sentiment during the week of the first presidential debate (26-09-2016) and discovered that most tweets are negative (criticizing the candidates), which is in line with related work [167]. In general, we observe that regardless of their topic (as discovered by the *Tweet keyword and time-based classifier*) the sentiment of our tweets is mostly negative (which is typical in political tweets [167]), with the exception of the *Inner politics* topic. For this particular subject, we observe more positive tweets, especially at the night after the first presidential debate, when users showed their support to Hillary Clinton addressing the problem of racism in the criminal justice system. Lastly, we briefly look at the sentiment on Twitter from another perspective as well. We detect bot accounts in our dataset with the Debot algorithm [23] and discover that the expressed sentiments on 26-09-2016 are more negative when the tweets are posted by humans and more positive when published by autonomous software.

4.7.5 Summary and Findings

In this section, based on the master thesis of Jaqueline Pollak, we focus on additional tasks to person tracking in social media and we utilize our previously introduced tweet dataset about the presidential elections in the USA in 2016. Based on a political topic taxonomy that is followed in the American presidential debates in 2016, we annotate a small tweet set for its topics and we also annotate the debate transcriptions accordingly. We use the first for testing our topic detection algorithm and the second for training it. We propose three different classifiers, i.e., the *Debate classifier* based on the politicians' speeches, *Tweet keyword-based classifier* based on tweets and *Tweet keyword and time-based classifier* based on tweets and their publication time. Our results show that the last model performs best and they all outperform the LDA algorithm after optimizing its number of topics and iterations. In addition, we apply our topic detection model in combination with a sentiment analysis tool (SentiStrength) and a bot detection tool (Debot) to find further insights into public opinions. We discover that similarly to related work most political tweets are negative regardless of their topic, and humans publish more negative posts than bots.

4.8 Future Work

Analyzing social networks is a very interesting problem, with applications in ethnography and sociology studies, crime detection, development of recommender systems, marketing campaigns, networking and product strategies for businesses, etc. We recognize this potential and contribute to the field of topic and event detection, while also discovering possible directions for future research. Regarding person tracking, we find out that the more messages are used for tracing the target entities, the more correct events can be discovered. Specifically, one can use more sophisticated methods for assigning a time to an event, i.e., temporal labeling of the tweets instead of considering the publication time of the message. In this way, more tweets would contribute to the detection of the events and our intuition says that the person tracking results can be further enhanced. We currently perform daily analysis, i.e., we use the tweets of a certain day to discover the events happened on that date. Thus, we allow users who discuss events asynchronously, but only within 24 hours. One can use temporal expressions [67], e.g., yesterday, 2night, tomorrow and also dates mentioned in the text. Given a more flexible temporal tagging approach, we can update our confidence not only about today's events, but also about other future and past events. Another future work direction is to estimate how many tweets are enough for our proposed solution to locate a person.

Moreover, in order to report the crowd's impression about an individual's event and also the anticipation for an upcoming event (e.g., how inspiring a TED talk was by an entrepreneur and how long-awaited the next event is), our system could additionally provide the average sentiment of the past and newly published tweets respectively. Another interesting improvement would be to consider various granularities of locations. We believe that this would be very helpful for counter-intelligence and law enforcement activities for decapitation attacks. For instance, considering fine-grained locations, such as towns and villages, can be useful when the individual of interest is a national risk and the accuracy of the reported information about him/her is essential. More abstract locations, such as on a country-level, might also be adequate for entertainment or business related activities, such as concerts and conferences.

Regarding topic detection during live events, other approaches are also promising, particularly sophisticated data representation techniques such as, Tweet2vec [39], which is more appropriate for the slang language and abbreviations used in social networks. An integral part of social messages is that the terms used to discuss different subjects is changing over time. Future work could adapt to this setting with different techniques, e.g., by evaluating our approach with forward-chain cross validation, which is applicable to time series data that also have this time-evolving nature. Lastly, we are interested in analyzing the public opinions and sentiments deeper, e.g., by discovering controversial tweets for given topics. A tweet could be considered as controversial when both the positive and negative scores are high. Another interesting question is whether users agree or disagree more for specific topics, which could help us understand which public matters and their respective solutions are important to the general public.

Chapter 5

CONCLUSION AND FUTURE WORK

The purpose of this thesis is to use and advance computer science, specifically machine learning algorithms, in a way that contributes to a more credible, transparent and diverse media ecosystem. We focus on political news and social media and extract insights about the shared content, the involved individuals and their opinions. That is, we examine our data from the perspective of the news sources, journalists, news commenters, politicians and social network users to discover how they discuss current affairs and hence how they shape collective opinions. Our motivation stems from the society's undeniable right to have a say in political decisions and access to reliable information that enables political accountability.

This thesis contains three main chapters, which address three individual research problems in the area of political text analysis. The first chapter discusses the characteristics of political language in news sources, the second one addresses the problem of bias in political news and the third one focuses locations and topics in political online debates in social media. We evaluate our hypotheses primarily with binary and multi-class text classification algorithms. We apply traditional machine learning models and engineer features based on the respective research problem, and we also build artificial neural networks and use semantic text representation techniques. Furthermore, the analyses in this thesis cover multiple countries, i.e., we investigate patterns in political news articles in the UK and USA, and news articles and comments in Germany.

Our first hypothesis in Chapter 2 is that the recent and possible mistrust towards the media could potentially be observed and proved in the differences of the media's reporting choices. That is, by showing the way each newspaper shapes reality, we can reveal evidence that justifies the media's suffering from a decline in public confidence. Thus, we examine how several media outlets discuss politicians' statements, each source choosing its own approach, and also the profile of their respective news commenters and the insights this additional data can bring into the media distinctiveness. We find that news media sources can be predictable and biased at times, e.g., the way they cite politicians and the kind of readers they attract is sufficient to makes us distinguish between them.

We believe that our work on political reported speech [83, 84, 86] in Chapter 2 can be applied in other domains as well, such as in any domain whose citations are an integral part and this leaves room for reporting discrimination. For instance, analyzing citations and their surrounding context in Wikipedia articles might reveal indicators for vandalism and hoaxes. Similarly, finding quotes in tweets and evaluating the opinionated context around them could show the political preferences of the general public. Regarding our news comment study [50] in the same chapter, one can utilize any webpage that allows comments to discover knowledge about the content of the page itself and its author. For instance, one could leverage the language style in the comments left under a scientific/informative blog post to gain insights into the concepts the author adopts and whether their commenters endorse them or criticize them. The limitation of our approach, i.e., given a text snippet from a news article (or a comment thread), predict its original news source, is that it is an indirect way to infer media bias indicators. Although our method is useful to overcome the absence of training data, it serves as a signal and not necessarily as proof of news bias. We believe that even though news source prediction is a valid and relevant task, future work should address the problem of media distinctiveness in a more straightforward manner. Such an approach can be similar to our proposed method in Chapter 3, but with data annotations on the publisher/newspaper and not article level.

In Chapter 3, given a political news article our goal is to discover its potential media bias. Being able to answer this research question would help both the journalists that might provide their own version of the truth (deliberately or accidentally) and the readers that consume this version. Namely, the journalists could reflect on their work and protect the public from misinformation, while the latter would be less biased towards the media's perspectives and more exposed to the actual facts. To this end, we take into account the lack of reliable news data annotations for its bias and introduce novel and labeled news corpora for this purpose. We analyze them qualitatively and quantitatively, and show that deep learning models can classify media bias successfully, and we also reveal a few challenging cases that our solution suffers to perform [87].

Media bias identification is very challenging task, especially due to its vague and often subjective definition. We believe that a clear and less ambiguous problem definition (with examples) in the media domain, but also in others, can facilitate the bias discovery in the text and it is thus a very relevant future work direction. There are additional types of bias that we did not consider in this thesis, but can be apparent in

text, such as gender bias [58, 63], race bias [114], and other agenda-setting attempts to influence the public and shift our focus on specific viewpoints. Sometimes different kinds of bias are connected (e.g., in racial politics) and in such cases, our approach could be applied at first to identify politically problematic content, and its results could be used as a signal for existence of additional bias types. Moreover, reliable media bias classifiers are the foundation for the next step, i.e., explainable and interpretable bias classifications. In the future, explainability techniques [134] could be applied on our model in order to reveal its internal behavior and understand how its conclusions are drawn.

Furthermore, user generated content, such as text in tweets and Facebook posts, is typically unfiltered and biased towards the user's perspective, since this kind of means are mainly used for opinion sharing among one's peers. Prominent public figures, political leaders and news providers also own social network accounts and often control information, by targeting specific online audiences with their messages and allowing these audiences to communicate with each other. All online political debates and discussions are an enormous data reservoir that can be leveraged, especially its up-to-dateness, for various applications. In Chapter 4, we utilize the wisdom of the crowd in the context of political social messages and the recency of the messages' content to gain joint insights into discussions on Twitter and also into other political datasets, i.e., the topics of the presidential debate transcripts and the locations of the presidential campaign events [54, 85]. In both cases, we utilize the time dimension of the datasets, i.e., in order to align the topics of public interest in tweets with the topics addressed in the debates, and in the latter work to analyze how people discuss politicians and their events, and eventually detect the politicians' locations. We build machine learning algorithms for topic detection and person tracking on Twitter. For the latter, we consider our evaluation as a proof of concept that social networks make it feasible to locate the mentioned individuals, which could be pivotal in emergency situations in order to find missing persons or capture criminals. Note that the amount of politicians mentioned in social media is far higher than the mentions of people that need to be found in the above-mentioned cases. This is a limitation of our current approach, which could be tackled in the future by transferring external knowledge about the target individuals into our models, e.g., from government websites, police reports and Wikipedia articles. We also see potential in a more flexible temporal tagging approach, where our model could update its confidence not only about today's events, but also about other future and past events. Especially in use cases where public safety is at stake, then gaining

more fine-grained insights about the target person is pivotal.

The media, although not officially a part of our political systems, hold the power to influence democracy by allowing access to information and thus shape our lives. This thesis aims to promote transparency in the political content of social and news media, so that they serve the global good. We take part in fulfilling this vision by introducing datasets and machine learning algorithms for reported speech analysis, media bias discovery, topic and event detection. We see potential in our proposed methods and their generalized applications, and at the same time, we acknowledge the improvements that should be addressed in future work. Our automatic solutions can reveal hidden and enlightening knowledge in online information. They can be used either individually or combined in future research in order to bring us another step closer to our vision of a better media ecosystem. As mentioned in Chapter 1, the Web has revolutionized our lives and the data that is shared on the Web has the potential to both enhance our decisions and mislead us. Hence, as Billy Don Moyers stated, the quality of democracy and the quality of journalism (in our days, either traditional or social) are fundamentally entwined[1].

[1] https://www.womensmediacenter.com/news-features/a-powerful-media-can-stop-a-war

BIBLIOGRAPHY

[1] Hamed Abdelhaq, Christian Sengstock, and Michael Gertz. EvenTweet: Online localized event detection from Twitter. In *Proceedings of the VLDB Endowment*. VLDB Endowment, 2013.

[2] Amr Ahmed and Eric P Xing. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010.

[3] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2017.

[4] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2010.

[5] Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. In *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association, 2010.

[6] Ramnath Balasubramanyan and Aleksander Kołcz. "w00t! feeling great today!": Chatter in Twitter: Identification and prevalence. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Association for Computing Machinery, 2013.

[7] Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019.

[8] Trapit Bansal, Mrinal Das, and Chiranjib Bhattacharyya. Content driven user profiling for comment-worthy recommendations of news and blog articles. In *Proceedings of the ACM Recommender Systems Conference*. Association for Computing Machinery, 2015.

[9] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2019.

[10] Eric Baumer, Elisha Elovic, Francesca Polletta, and Geri Gay. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2015.

[11] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the ACM Annual International Conference on Machine Learning*. Association for Computing Machinery, 2009.

[12] Sabine Bergler. Conveying attitude with reported speech. In *Computing Attitude and Affect in Text: Theory and Applications*, chapter 2. Springer Netherlands, 2006.

[13] Adam Bermingham and Alan Smeaton. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology*. Association for Computational Linguistics, 2011.

[14] Tim Berners-Lee and Mark Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. DIANE Publishing Company, 2001.

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[16] Alireza Bonyadi. The rhetorical properties of the schematic structures of newspaper editorials: A comparative study of english and persian editorials. *Discourse and Communication*, 4, 2010.

[17] Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. Selection bias in news coverage: learning it, fighting it. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2018.

[18] Leo Breiman. Random forests. *Machine learning*, 45, 2001.

[19] Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80, 2016.

[20] Xuezhi Cao, Kailong Chen, Rui Long, Guoqing Zheng, and Yong Yu. News comments generation via mining microblogs. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2012.

[21] Alessandra Cervone, Catherine Lai, Silvia Pareti, and Peter Bell. Towards automatic detection of reported speech in dialogue using prosodic cues. In *Proceedings of the Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2015.

[22] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Debot: Twitter bot detection via warped correlation. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE Computer Society, 2016.

[23] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Identifying correlated bots in Twitter. In *Proceedings of the International Conference on Social Informatics*. Springer, 2016.

[24] Junsha Chen, Neng Gao, Yifei Zhang, and Chenyang Tu. Local topic detection using word embedding from spatio-temporal social media. In *Proceedings of the International Conference on Neural Information Processing*. Springer, 2019.

[25] Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Learning to flip the bias of news headlines. In *Proceedings of the International Conference on Natural Language Generation*. Association for Computational Linguistics, 2018.

[26] Xingyu Chen, Lei Zou, and Bo Zhao. Detecting climate change deniers on Twitter using a deep neural network. In *Proceedings of the International Conference on Machine Learning and Computing*. Association for Computing Machinery, 2019.

[27] François Chollet et al. Keras. https://keras.io, 2015.

[28] Nupur Choudhury. World wide web and its journey from web 1.0 to web 4.0. *International Journal of Computer Science and Information Technologies*, 5, 2014.

[29] William S Cleveland. Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69, 2001.

[30] Raviv Cohen and Derek Ruths. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2013.

[31] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20, 1995.

[32] Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. Media bias in german online newspapers. In *Proceedings of the Conference on Hypertext and Social Media*. Association for Computing Machinery, 2015.

[33] Mrinal Kanti Das, Trapit Bansal, and Chiranjib Bhattacharyya. Going beyond corr-lda for detecting specific comments on news and blogs. In *Proceedings of the International Conference on Web Search and Data Mining*. Association for Computing Machinery, 2014.

[34] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics, Journal of the Royal Statistical Society, Series C*, 28, 1979.

[35] Orphee De Clercq, Sven Hertling, Veronique Hoste, Simone Paolo Ponzetto, Heiko Paulheim, et al. Identifying disputed topics in the news. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databasesy*. Online Proceedings for Scientific Conferences and Workshops, 2014.

[36] Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[38] Ralf Dewenter, Melissa Linder, and Tobias Thomas. Can media drive the electorate? The impact of media coverage on party affiliation and voting intentions. *European Journal of Political Economy*, 58, 2019.

[39] Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William Cohen. Tweet2vec: Character-based distributed representations for social media. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2016.

[40] Nicholas Diakopoulos and Mor Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery, 2011.

[41] Nicholas A. Diakopoulos. The editor's eye. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery, 2015.

[42] Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. Twitter user geolocation using deep multiview learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE Computer Society, 2018.

[43] Jonathan G Fiscus and George R Doddington. Topic detection and tracking evaluation overview. In *Topic detection and tracking*. Springer, 2002.

[44] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. Quote rts on Twitter: usage of the new feature for political discourse. In *Proceedings of the ACM Conference on Web Science*. Association for Computing Machinery, 2016.

[45] Kiran Garimella, Gianmarco Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2018.

[46] Michel Généreux and Marina Santini. Exploring the use of linguistic features in sentiment analysis. In *Proceedings of the International Corpus Linguistics Conference*. University of Brighton, 2007.

[47] Matthew Gentzkow and Jesse M Shapiro. What drives media slant? Evidence from us daily newspapers. *Econometrica*, 78, 2010.

[48] C Lee Giles, Gary M Kuhn, and Ronald J Williams. Dynamic recurrent neural networks: Theory and applications. *IEEE Transactions on Neural Networks*, 5, 1994.

[49] Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2007.

[50] Christian Godde, Konstantina Lazaridou, and Ralf Krestel. Classification of german newspaper comments. In *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen"*. Online Proceedings for Scientific Conferences and Workshops, 2016.

[51] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. Using topic models for Twitter hashtag recommendation. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2013.

[52] Tim Groseclose and Jeffrey Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 120, 2005.

[53] Toni Gruetze, Gjergji Kasneci, Zhe Zuo, and Felix Naumann. CohEEL: Coherent and efficient named entity linking through random walks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37, 2016.

[54] Toni Gruetze, Ralf Krestel, Konstantina Lazaridou, and Felix Naumann. What was Hillary Clinton doing in Katy, Texas? In *Companion Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017.

[55] Adrien Guille and Cécile Favre. Mention-anomaly-based event detection and tracking in Twitter. In *Proceedings of the International Conference Series on Advances in Social Network Analysis and Mining*. IEEE Computer Society, 2014.

[56] Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, 20, 2018.

[57] Jack Hanson, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, 33, 2017.

[58] Dustin Harp, Jaime Loke, and Ingrid Bachmann. Hillary Clinton's Benghazi hearing coverage: Political competence, authenticity, and the persistence of the double bind. *Women's Studies in Communication*, 39, 2016.

[59] Christopher Harris. Searching for diverse perspectives in news articles: Using an lstm network to classify sentiment. In *Proceedings of the IUI Workshops, ACM Conference on Intelligent User Interfaces*. Association for Computing Machinery, 2018.

[60] Zellig Harris. Distributional structure. *Word*, 10, 1954.

[61] Zaobo He, Zhipeng Cai, and Jiguo Yu. Latent-data privacy preserving with customized data utility for social network data. *IEEE Transactions on Vehicular Technology*, 67, 2017.

[62] Bahareh Rahmanzadeh Heravi and Ihab Salawdeh. Tweet location detection. In *Proceedings of the Computation Journalism Symposium*. Columbia University, 2015.

[63] Marc Hooghe, Laura Jacobs, and Ellen Claes. Enduring gender bias in reporting on political elite positions: Media coverage of female mps in belgian news broadcasts (2003–2011). *The International Journal of Press/Politics*, 20, 2015.

[64] Leonard Hövelmann, Stockholmer Allee, and Christoph M Friedrich. Fasttext and gradient boosted trees at germeval-2017 on relevance classification and document-level polarity. In *Proceedings of the GSCL GermEval Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*. Creative Commons, 2017.

[65] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in Neural Information Processing Systems*. The MIT Press, 2014.

[66] Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. Breaking news on Twitter. In *Proceedings of the Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2012.

[67] Ali Hürriyetoğlu, NHJ Oostdijk, and APJ van den Bosch. Estimating time to event from tweets using temporal expressions. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2014.

[68] Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30, 2018.

[69] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2014.

[70] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017.

[71] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv*, 1607.01759, 2016.

[72] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017.

[73] Ravi Kant, Srinivasan H. Sengamedu, and Krishnan S. Kumar. Comment spam detection by sequence mining. In *Proceedings of the International Conference on Web Search and Data Mining*. Association for Computing Machinery, 2012.

[74] Erin M. Kearns, Allison Betus, and Anthony Lemieux. Why do some terrorist attacks receive more media attention than others? *Justice Quarterly*, 36, 2017.

[75] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2019.

[76] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference for Learning Representations*. Conference Track Proceedings, 2014.

[77] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, System Demonstrations*. Association for Computational Linguistics, 2017.

[78] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2011.

[79] Thomas B Ksiazek. Commenting on the news: Explaining the degree and quality of user comments on news websites. *Journalism Studies*, 19, 2018.

[80] Onur Kucuktunc, B Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. A large-scale sentiment analysis for Yahoo! answers. In *Proceedings of the ACM International Conference on Web search and data mining*. Association for Computing Machinery, 2012.

[81] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. Search bias quantification: Investigating political bias in social media and web search. *Information Retrieval Journal*, 22, 2018.

[82] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2010.

[83] Konstantina Lazaridou and Ralf Krestel. Identifying political bias in news articles. *Special Issue of the Bulletin of the IEEE Technical Committee on Digital Libraries*, 12, 2015.

[84] Konstantina Lazaridou, Ralf Krestel, and Felix Naumann. Identifying media bias by analyzing reported speech. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE Computer Society, 2017.

[85] Konstantina Lazaridou, Toni Gruetze, and Felix Naumann. Where in the world is Carmen Sandiego?: Detecting person locations via social media discussions. In *Proceedings of the International Conference on Web Science*. Association for Computing Machinery, 2018.

[86] Konstantina Lazaridou, Ralf Krestel, Alexander Loeser, and Felix Naumann. Reported speech in political news articles: A media bias perspective. *Natural Language Engineering (in revision)*, 2019.

[87] Konstantina Lazaridou, Alexander Loeser, and Felix Naumann. Media bias detection with humans in the loop: Discovering biased news articles. In *Proceedings of the International Language Resources and Evaluation Conference*. European Language Resources Association, 2019.

[88] Gregor Leban, Aljaž Košmerlj, Evgenia Belyaeva, and Blaž Fortuna. News reporting bias detection prototype. Technical report, Institut Jožef Stefan, 2014.

[89] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2009.

[90] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets*. Cambridge University Press, 2020.

[91] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the European Conference on Machine Learning*. Springer, 1998.

[92] Chenliang Li and Aixin Sun. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2014.

[93] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. Tedas: A Twitter-based event detection and analysis system. In *Proceedings of the International Conference on Data Engineering*. IEEE Computer Society, 2012.

[94] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2019.

[95] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2006.

[96] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*. Springer, 2012.

[97] Wiebke Loosen, Marlo Häring, Zijad Kurtanović, Lisa Merten, Julius Reimer, Lies van Roessel, and Walid Maalej. Making sense of user comments: Identifying journalists' requirements for a comment analysis framework. *SCM Studies in Communication and Media*, 6, 2018.

[98] Estefanía Lozano, Jorge Cedeño, Galo Castillo, Fabricio Layedra, Henry Lasso, and Carmen Vaca. Requiem for online harassers: Identifying racism from political tweets. In *Proceedings of the International Conference on e-Democracy and e-Government*. IEEE Computer Society, 2017.

[99] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011.

[100] Debanjan Mahata, John R. Talburt, and Vivek Kumar Singh. From chirps to whistles: Discovering event-specific informative content from Twitter. In *Proceedings of the International Web Science Conference*. Association for Computing Machinery, 2015.

[101] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of Twitter users. *ACM Transactions on Intelligent Systems and Technology*, 5, 2014.

[102] Lionel Martin, Valentina Sintsova, and Pearl Pu. Are influential writers more objective? In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2014.

[103] Michael Mathioudakis and Nick Koudas. TwitterMonitor: Trend detection over the Twitter stream. In *Proceedings of the International Conference on Management of Data*. Association for Computing Machinery, 2010.

[104] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Workshop Track Proceedings of the International Conference on Learning Representations*. arXiv, 2013.

[105] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems*. Annual Conference on Neural Information Processing Systems, 2013.

[106] Fahim Mohammad. Is preprocessing of text really worth your time for online comment classification? *ArXiv*, abs/1806.02908, 2018.

[107] Elaheh Momeni and Gerhard Sageder. An empirical analysis of characteristics of useful comments in social media. In *Proceedings of the International Web Science Conference*. Association for Computing Machinery, 2013.

[108] Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1, 2018.

[109] Islam Muhammad. Media coverage of the 2016 brexit referendum: An analysis of the Guardian and the Telegraph coverage using social responsibility theory. School of Governance, Tallinn University, Law and Society, 2018.

[110] Eni Mustafaraj and Panagiotis Takis Metaxas. The fake news spreading plague: Was it preventable? In *Proceedings of the International Web Science Conference*. Association for Computing Machinery, 2017.

[111] Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. A two-stage sieve approach for quote attribution. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017.

[112] Nona Naderi and Graeme Hirst. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Association for Computational Linguistics, 2017.

[113] Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015.

[114] David Niven. A fair test of media bias: Party, race, and gender in coverage of the 1992 house banking scandal. *Polity*, 36, 2004.

[115] Mark Nixon and Alberto Aguado. *Feature extraction and image processing for computer vision*. Academic Press, 2019.

[116] Kezban Dilek Onal, Ye Zhang, Ismail Sengor Altingovde, Md Mustafizur Rahman, Pinar Karagoz, Alex Braylan, Brandon Dang, Heng-Lu Chang, Henna Kim, Quinten McNamara, et al. Neural information retrieval: At the end of the early years. *Information Retrieval Journal*, 21, 2018.

[117] Sean Papay and Sebastian Padó. Quotation detection and classification with a corpus-agnostic model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. INCOMA Ltd., 2019.

[118] Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. Automatically detecting and attributing indirect quotations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2013.

[119] Souneil Park, Minsam Ko, Jungwoo Kim, Ying Liu, and Junehwa Song. The politics of comments: Predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. Association for Computing Machinery, 2011.

[120] Souneil Park, KyungSoon Lee, and Junehwa Song. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2011.

[121] Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. A link-based approach to detect media bias in news websites. In *Companion Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2019.

[122] Dario Pavllo, Tiziano Piccardi, and Robert West. Quootstrap: Scalable unsupervised extraction of quotation-speaker pairs from large news corpora

via bootstrapping. In *Proceedings of the International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence, 2018.

[123] John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. Improved abusive comment moderation with user embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.

[124] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2016.

[125] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014.

[126] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2018.

[127] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *Proceedings of the ACM International on Conference on Information and Knowledge Management*. Association for Computing Machinery, 2016.

[128] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2018.

[129] Adam Poulston, Mark Stevenson, and Kalina Bontcheva. Hyperlocal home location identification of Twitter profiles. In *Proceedings of the ACM Conference on Hypertext and Social Media*. Association for Computing Machinery, 2017.

[130] Søren B. Ranneries, Mads E. Kalør, Sofie Aa. Nielsen, Lukas N. Dalgaard, Lasse D. Christensen, and Nattiya Kanhabua. Wisdom of the local crowd: Detecting local events using social media data. In *Proceedings of the International Web Science Conference*. Association for Computing Machinery, 2016.

[131] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised

learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the Annual International Conference on Machine Learning*. Association for Computing Machinery, 2009.

[132] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2013.

[133] Filipe N Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence, 2018.

[134] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2016.

[135] Georgios Rizos, Symeon Papadopoulos, and Yiannis Kompatsiaris. Predicting news popularity by mining online discussions. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2016.

[136] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, 2013.

[137] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2010.

[138] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2010.

[139] Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection. In *Proceedings of the International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2019.

[140] Justin Sampson, Fred Morstatter, Ross Maciejewski, and Huan Liu. Surpassing the limit: Keyword clustering to improve Twitter sample coverage. In

Proceedings of the ACM Conference on Hypertext and Social Media. Association for Computing Machinery, 2015.

[141] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. TwitterStand: News in tweets. In *Proceedings of the International Conference on Advances in Geographic Information Systems*. Association for Computing Machinery, 2009.

[142] Christian Scheible, Roman Klinger, and Sebastian Padó. Model architectures for quotation detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2016.

[143] Jonathon P Schuldt, Sara H Konrath, and Norbert Schwarz. "Global warming" or "climate change"? Whether the planet is warming depends on question wording. *Public Opinion Quarterly*, 75, 2011.

[144] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence, 2013.

[145] Robert P Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53, 2012.

[146] Erez Shmueli, Amit Kagian, Yehuda Koren, and Ronny Lempel. Care to comment? Recommendations for commenting on news stories. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2012.

[147] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web*, 8, 2014.

[148] Vaibhav B Sinha, Sukrut Rao, and Vineeth N Balasubramanian. Fast dawid-skene: A fast vote aggregation scheme for sentiment classification. In *Proceedings of the Workshop on Issues of Sentiment Discovery and Opinion Mining*. Association for Computing Machinery, 2018.

[149] Alla Vitaljevna Smirnova. Reported speech as an element of argumentative newspaper discourse. *Discourse and Communication*, 3, 2009.

[150] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.

[151] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the International Conference on Machine Learning*. Omnipress, 2011.

[152] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng, and Christopher Potts. Parsing with compositional vector grammars. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2013.

[153] Yangqiu Song, Zhengdong Lu, Cane Wing-ki Leung, and Qiang Yang. Collaborative boosting for activity classification in microblogs. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2013.

[154] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2007.

[155] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Proceedings of the Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2012.

[156] Mike Thelwall. The heart and soul of the web? Sentiment strength detection in the social web with sentistrength. In *Cyberemotions*. Springer, 2017.

[157] Dennis Thom, Harald Bosch, Robert Krueger, and Thomas Ertl. Using large scale aggregated knowledge for social media location discovery. In *Proceedings of the Hawaii International Conference on System Sciences*. IEEE Computer Society, 2014.

[158] Manos Tsagkias, Wouter Weerkamp, and Maarten de Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, 2009.

[159] Paraskevas Tsantarliotis, Evaggelia Pitoura, and Panayiotis Tsaparas. Defining and predicting troll vulnerability in online social media. *Social Network Analysis and Mining*, 7, 2017.

[160] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2010.

[161] Khonzodakhon Umarova and Eni Mustafaraj. How partisanship and perceived political bias affect wikipedia entries of news sources. In *Companion*

Proceedings of the International Conference on World Wide Web. Association for Computing Machinery, 2019.

[162] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the Workshop on Abusive Language Online*. Association for Computational Linguistics, 2018.

[163] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*. The MIT Press, 2017.

[164] Emmanuel Vincent and Maria Mestre. Crowdsourced measure of news articles bias: Assessing contributors' reliability. In *Proceedings of the Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and of the Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management*. Online Proceedings for Scientific Conferences and Workshops, 2018.

[165] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics, 2014.

[166] Svitlana Volkova and Jin Yea Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018.

[167] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time Twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, System Demonstrations*. Association for Computational Linguistics, 2012.

[168] Jing Wang, Clement T. Yu, Philip S. Yu, Bing Liu, and Weiyi Meng. Diversionary comments under political blog posts. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, 2012.

[169] Shiguang Wang, Prasanna Giridhar, Hongwei Wang, Lance Kaplan, Tien Pham, Aylin Yener, and Tarek Abdelzaher. Storyline: Unsupervised geo-event demultiplexing in social spaces without location information. In *Proceedings of the ACM/IEEE International Conference on Internet of Things Design and Implementation*. Association for Computing Machinery, 2017.

[170] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering.

In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017.

[171] Jeffrey N Weatherly, Thomas V Petros, Kimberly M Christopherson, and Erin N Haugen. Perceptions of political bias in the headlines of two major news organizations. *Harvard International Journal of Press/Politics*, 12, 2007.

[172] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. *ArXiv*, 1802.03796, 2018.

[173] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005.

[174] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *ACM SIGMOD Record*, 31, 2002.

[175] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on Knowledge and Data Engineering*, 28, 2016.

[176] Tai-Yee Wu and David J Atkin. To comment or not to comment: Examining the influences of anonymity and social support on one's willingness to express in online news discussions. *New Media and Society*, 20, 2018.

[177] Tae Yano, Philip Resnik, and Noah A Smith. Shedding (a thousand points of) light on biased language. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010.

[178] Mengfan Yao, Charalampos Chelmis, and Daphney?Stavroula Zois. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *Proceedings of the International Conference on World Wide Web*. Association for Computing Machinery, 2019.

[179] Hang Zhang and Vinay Setty. Finding diverse needles in a haystack of comments. In *Proceedings of the International Web Science Conference*. Association for Computing Machinery, 2016.

[180] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and traditional media using topic models. In *Proceedings of the European Conference on Information Retrieval*. Springer, 2011.

[181] Yang Zhao, Ming-Ching Chang, and Peter Tu. Deep intelligent network for device-free people tracking: Wip abstract. In *Proceedings of the ACM/IEEE*

International Conference on Cyber-Physical Systems. Association for Computing Machinery, 2019.

[182] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and users from user votes. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2011.

[183] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2018.