LEARNING HOW TO DETECT NEWS BIAS

Seminar I

Supervisor: Dunja Mladenic

Approved by the supervisor: _____________________
(podpis/ signature)

Study programme:
Information and Communication Technologies
Doctoral degree

Ljubljana, 2015
Abstract

The purpose of this paper is to review recent research work related to the field of automatic detection of news bias. First, this paper briefly discusses the notion of bias in both cognitive and media studies. It then introduces news bias in connection with studying media systems as a whole. Second, it gives an overview of the related work on detecting news bias in the fields of NLP, machine learning and data mining. It addresses the issues of identifying the following: geographical news analysis, readability measure, newswire citation analysis, coverage similarity and sentiment mining in news. Last, past work is critically evaluated specifying some weaknesses and limitations and recommendations for future research on intersection between computer and social sciences are presented.

Keywords: news bias, cognitive bias, machine learning, data and text mining, sentiment analysis, news values
Table of Contents

Abstract .................................................................................................................................................. 2
Abbreviations ....................................................................................................................................... 4
Chapter 1. Problem Definition .......................................................................................................... 5
  1.1 Definition of bias ......................................................................................................................... 5
  1.2 News bias ..................................................................................................................................... 6
Chapter 2. Text Mining for New Bias Analysis ................................................................................. 8
  2.1 Keyword analysis ......................................................................................................................... 8
  2.2 Article grammatical and lexical differences ............................................................................... 9
  2.3 Readibility ..................................................................................................................................... 10
  2.4 Newswire citation bias ............................................................................................................... 10
  2.5 Geographical analysis of news .................................................................................................. 11
  2.6 Speed of reporting bias .............................................................................................................. 12
  2.7 Similarity in the coverage of events ........................................................................................... 13
  2.8 Topic similarity bias ..................................................................................................................... 14
  2.9 Gender Bias ................................................................................................................................. 14
  2.10 Sentiment analysis .................................................................................................................... 15
Chapter 3. Critical judgement of the existing research .................................................................... 16
Chapter 4. Future work ...................................................................................................................... 17
Reference List ....................................................................................................................................... 19
Abbreviations

**AFP** – Associate France Press

**AP** – Associated Press

**EU** – European Union

**LDA** – Latent Dirichlet Allocation

**NLP** – Natural Language Processing

**SVM** - Support Vector Machines

**SVO** - subject-verb-object

**TF-IDF** - Term Frequency - Inverse Document Frequency
Chapter 1. Problem Definition

News bias is a ubiquitous phenomenon, which can be potentially present in every news-reporting outlet off - and on-line, has already generated various research studies in different fields: from anthropology, social and media studies to computer sciences.

Already a highly evident problem in the social sciences, many of the media investigations focus its attention on detecting news bias on particular issues like election, immigration, wars, or racism. The news articles are mostly selected and analysed manually using a process called “coding” or theoretical frameworks like discourse analysis and content analysis. This analysis, which requires a lot of effort, concentration, attention to detail and a lot of time, is then limited to a small amount of samples selected by hand. Studies of bias in the computational field, however, concentrate on methods for pattern analysis [1] and annotate large amount of text data in order to detect patterns or biases, without the limitations of time, size of the corpus, analytical framework and hypothesis. Thus the problem of detecting news bias is important to both – the media field, as well as computer sciences, which is trying to ease the workflow of many researchers in social sciences and to automate the analysis of large amounts of data.

However, the current state of the art is not close to reaching the goal of automatically detecting news bias. It has heavily relied on statistical machine learning methods including readability analysis, topic comparison, geographical bias etc., without engagement of various media frameworks and theories on news bias. Recent research developments have highlighted the need to work on the intersections of social and computer sciences in order to be able to advance in detecting news bias.

1.1 Definition of bias

Research in cognitive science, psychology and other social studies offer a great amount of work on biases and their affect on a variety of human activities. With an increasingly high number of researches, we notice a great diversity of the meaning and definitions of the concept of bias. Each individual field has its own problematic, which might be reflected in the notion of the bias as well. This Chapter gives a short overview of how the notion of bias is defined in the field of cognitive science.

Bias in cognitive science is generally defined as a deviation from a norm, deviation from some true or objective value. The presence of bias is easy to assert if a normative model describing how it should be is present. On the contrary, if the norm is unknown to us, the bias is not easy to identify [2]. Bias, as defined by Goldstein&Gigerenzer, assumes that our mind’s information processing is limited and thus, we often apply short cuts using less cognitive effort known as heuristics [3]. Heuristics are reproducible errors in human reasoning, which are used to quickly give answers in many cases, but they often give rise to errors or bias. This approximations make us use and work with whatever first comes to mind, and based on these first thoughts we make erroneous judgements and our mental search process is limited [4].

It has been suggested by Ross that any preferences people have towards watching a certain movie or eating a special food are likely to have some bias associated with them, for most part, unconscious. He refers to it as a »glitch« in thinking that we have been exposed to or taught, or other universal truths we have collected along our ways. Unconscious bias comes with evolution, from social stereotypes, attitudes, generalisations and opinions we form about a certain thing or person or a group of people [5].
Kahneman et al. introduced a dual approach to cognition, calling it »heuristics and biases« one being intuitive, natural, fast, without efforts and difficult to control thinking and the second being analytical, slow, rational, controlled [6]. They also suggest that people rely on a limited number of principals that reduce complex tasks and provoke more simple judgemental cognitive operations arguing that they often lead to unconscious biases [6]. Heuristics is the most widely used explanation of bias, as they are bi-product of processing limitations of our brain in terms of time and ability.

Haselton considers bias as a deviation from the standards of logic and accuracy and refers to biases as flaws [7]. He claims that cognitive bias can stem from three main reasons: 1) Heuristics or shortcuts that tend to work in many circumstances and that result from evolutionary constraints, like stereotypes; 2) error management – a tendency to favour selection toward the less costly error, for example, positive illusions and positive perceptions on oneself; 3) artefacts when a task at hand is not the one for which our mind was designed, such as, confirmation bias.

Important to note, when we speak about bias we might refer to a group bias or to a bias of an individual. Some biases influence decision-making process and some involve a judgement. Cognitive biases are thought to be »inappropriate intrusions of subjective opinion into a factual account«, they are understood as ways of thinking that constrain one’s perception and interpretations of the world [8]. Humans are not consistent and fundamentally bias creatures in nature, simply because that is the way our brain works.

1.2 News bias

Despite the fact that journalists are supposed to provide the readers with impartial, objective, unbiased and reliable information, the reality is somehow different. Every news story has a potential to be biased. Every news story has a potential to be influenced by the attitudes, cultural background, political and economic views of the journalists and editors. Every news reporting is laden with values and on occasion, is fundamentally biased. A journalist collects facts, reports them objectively and the news outlet presents it as it is without any hidden bias and slanted language sounds almost like an utopia. In reality, there are many examples of poor journalistic practice that can be seen almost every day: a journalist stating his personal opinion in a news report, adding incorrect facts and figures, applying unequal space to different sides of a controversial issue, citing people of a certain political class or gender. Bias exist also because journalism is a »subjective art« [9]. Journalists are usually the ones taking decisions on what news story to write, which experts to interview, what facts to include and what stories to present on front page. And each stage in the news making process is a stage where a subjective opinion might enter. It is not easy to completely detach oneself from a story that a journalist covers.

News bias cannot be understood without understanding of the context of media industry as a whole. When any news content is subjected to »organizational routines« then it is already a clear sign of selection bias. Media houses reproduce and publish content selectively according to the criteria that usually suits their own agendas and interests. This criteria can be first professional, example of news values, but also economical or what brings them more ranking, popularity and money. Bias, as suggested by McQuail, may only mean selecting products that are both easy to reproduce and is popular with the target audience [10]. The scholar also underlines that the need from journalists to follow the news values when choosing a story and an importance of authoritative sources like government etc. are already obvious factors of
distortion [10]. Another fact that media is generally oriented towards the interest of the audience for information and entertainment can also account for the distortion.

News bias is a complex process that comprises several dimensions to be taken into account; it is interlinked with social, political and economical problems. In media studies, bias often refers to conspiracy or corporate control and the perception of it comes, in part, as an economical problem [11]. Given that bias is rule in media, not an exception, Edler suggests that media houses have to sell their stories to be in business and have no choice but to »package« their stories within ideological and sociocultural framework of the society [12]. Media is a business and journalists operate in an economical environment. It is often, however, also seen as a political problem when looking at the political affiliation of a particular news outlet.

News bias is also regarded as a problem of ideology and media houses present the news that go alone with the ideology of the environment or the country. Ideological bias in news is defined as prejudice in favour of one country and ideology, news are excluded or included based on what is important within the view of the nation, and people tend to assume what is present in the news is true [12]. If a reader is claiming that the news is bias, it might simply mean that he does not share an ideology of a particular news outlet. Outlet is biased if it systematically slants its content towards a particular political party or an ideology [13]. For example, a Russian journalist sees the world he or she is reporting about slightly or largely differently from his Chinese or Italian journalist colleague. A journalist working for the Guardian is a member of his environment sharing common views on the world with his target audience (in this case, the UK): the journalist and his audience share often similar history, culture, religion and general ideology.

The notion of objectivity is often studied together with news bias as it is seen as a deviation from an objective reporting, but also impartiality of reporting, which is defined as representation of different views and interests without taking sides [14]. Objectivity in general is hard to achieve, especially if we know that news bias is an end product of various already subjective opinions people hold, it is a matter of perception, which is often different among individuals [15].
Chapter 2. Text Mining for New Bias Analysis

Text mining or knowledge discovery is a process of automatically extracting information from unstructured text documents [16], by combining techniques from data mining, machine learning, natural language processing (NLP) and information retrieval. Most common text mining tasks involve document classification, summarization, clustering of documents, concept extraction and sentiment analysis. Text mining has a great number of applications to date, including bias and sentiment detection in news and the questions “What makes a news piece bias?” and “How can we identify what source is more slanted than the other and towards what?” can now be addressed automatically, deploying the above-mentioned analytical techniques. Nowadays there are various applications of text-analysis technologies that can support social scientists in the analysis of news patterns and can help in automating tasks that were usually performed manually [1].

One of the first large-scale content analysis of news across languages, by using a number of text mining techniques, detected a clear bias in the choice of stories covered by numerous media outlets based in the European Union [17]. The detected according to the study bias depended on the economic, geographical and cultural relations among the media outlets and the countries. Countries with strong economic ties, for example, are more likely to write about economy.

In this chapter we investigate the following state-of-the-art done in order to identify bias in news reporting:

- Keyword analysis
- Analysis of grammatical differences
- Readability differences
- Geographical bias
- Topic coverage bias
- Speed of reporting bias
- Newswire citation bias
- Similarity in the coverage of events
- Gender bias
- Sentiment analysis

It is necessary to clarify that numerous state-of-the-art techniques are used for detecting various types of biases. Classical supervised learning techniques, for instance Support Vector Machines, naïve Bayes, together with unsupervised and weakly supervised methods can be used. A selection of various features are also commonly used for a classification task when studying bias, including unigrams, N-grams, or part-of-speech. Due to the space limitations not all of them are discussed in detail in this paper.

2.1 Keyword analysis

To get the first impressions of the data under experiments, often a keyword frequency analysis is performed using Term Frequency - Inverse Document Frequency (TF-IDF) weights for every word in the text. Then, the weight of a term $t$ in document $d$ from document collection $D$ is calculated the following way:
The list of computed keywords is often then sorted by the so-called keyness value - a measure that compares relative frequencies of a certain word in the text and compared to a reference corpus [18].

Along with keyword analysis, one can look at the length of headlines – as they are considered to be the most read part of a story and are the basis of how a story would develop, then sentence length and articles length, counting either tokens, words or sentences and in social media, for example, counting the number of tweets mentioning a certain topic.

Frequency based analysis is usually a first step in analysing any news patterns including bias. It might give some first trends and ideas on how our data looks like, but the main disadvantage is that it is often completely de-contextualized in case of keywords and editorial decision dependent in case of length analysis.

### 2.2 Article grammatical and lexical differences

Often the study of grammatical and lexical differences in news articles between the certain publishers with regards to the usage of various parts of speech, like adjectives, adverbs and nouns and how these properties differ when reporting about articles from different categories are studied in connection to news bias. This is to say, specific choice of words and subtle structure of sentences can persuade the reader towards one point of view or another and are sufficient to influence whether people interpret violent acts as patriotism or terrorism [19].

Adjectives have always been directly linked to studying of news bias, the sentiment part of studying news bias. To attract the readers' attention and to get more clicks on articles, journalists (in particular those working for tabloid news publishers) often make use of the descriptive language that involves the use of colourful adjectives. The use of adverbs in news writing is not against the rules, but most of them are not necessary and the overuse of them can be considered bad journalistic writing style. It has been proven that not only lexical choices can lead to slanted reporting, but even simple syntactic structure of sentences, such as the use of active or passive constructions, which allow a reporter to suppress the agent of the news actions [20].

At the first step a part from the keyword analysis, what is often done is simple automatic count of various parts of speech (e.g. lists of the top twenty adjectives or adverbs per publisher on a certain topic are produced). Further studies of bias linked to lexical and grammatical cues on subjectivity [21], [22] and sentiment [23] are later discussed in Chapter 2.8 on sentiment analysis.

Several studies have produced lexical and grammatical analysis of news. When, for example, studying US elections, the ElectionWatch system on news content, used a parser to extract SVO (subject-verb-object) triplets, in order to form a semantic graph to identify the noun phrases with actors and to classify the verbal links between actors. By identifying the most important actors and triplets, they formed a large directed network, which was then analysed for patterns. The study concluded the presence of the so called Actor bias (the role a certain actor played in the news media) a positive value of a subject says that actor is used as an object, and a negative

\[
TF \times IDF_{t,d,D} = TF_{t,d} \cdot IDF_{t,D} \\
TF_{t,d} = \text{freq}_{t,d} \\
IDF_{t,D} = \log \frac{|D|}{n_{t,D}} \\
n_{t,D} : \text{number of documents in } D \text{ with term } t
\]
value says that the actor is used as an object and used the following formula to calculate the Actor bias [24]:

\[ S_a = f_{Subj}(a) \cdot f_{Obj}(a) / f_{Subj}(a) + f_{Obj}(a) \]

### 2.3 Readability

Central to the study of news bias is the readability analysis. Readability is an indicator of how understandable a text is to a particular group of readers [25]. Readability measures have been used extensively to help to evaluate and develop textbooks, business publications, medical literature and nowadays also news articles. There are many readability formulas, and most of them are based on two metrics: the complexity of sentences and the complexity of words. The complexity of sentences is measured by the average number of words per sentence, while the complexity of words is measured in different ways by different measures. Dale-Chall readability formula is one of the most widely used and it is computed as follows [26]:

\[
0.1579 \left( \frac{\text{difficult words}}{\text{words}} \times 100 \right) + 0.0496 \left( \frac{\text{words}}{\text{sentences}} \right)
\]

The first part of the equation computes the percentage of difficult words in the article. The “difficult” words are the ones that do not appear on a specifically designed list of common words that are familiar to most 4th-grade students. This list originally contained 763 terms and was later expanded to 3,000 words [26]. The computed score is then normalized so that it can be used to estimate the grade level needed to understand the text.

Another widely used formula to calculate the average readability is referred to as Flesch/Flesch-Kincaid readability test used for evaluating technical documents and is computed in the following way taking into consideration not only words, sentences but also syllables. In contrast to the Dale-Chall score, Flesch-Kincaid is higher for article that is easier to read [27]:

\[
206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)
\]

In the work of Flaounas et al. [17] readability was measured together with linguistic subjectivity (measure of usage of words that carry sentiment). They calculated the average readability and concluded that sports articles are easier to read and politics are the hardest to read. All measures of readability do not consider and are limited due to the lack of cultural aspect and structure of language since they were only applicable to English news [17].

### 2.4 Newswire citation bias

Another example of automatic detection of news bias is associated with newswire citations i.e. how frequent selected news publishers cite different news agencies or how newswire hyperlinks are useful to predict bias of websites using their linking pattern [28]. Important to note that news agencies (also called press agencies or newswires) have correspondents and journalists in most of the countries and they often do not publish news to the public, but rather serve as suppliers to numerous news publishers around the world. There are many news agencies in the world, but the following three are the main providers of global news: Associated Press (AP), Reuters and Agence France-Presse (AFP). All mass media outlets depend on
agencies for the international news and the publishers are usually obliged to cite where the information is coming from by referring to the press agency and not the actual author itself.

Traditionally, there are two core citation methods: bibliographical coupling, where documents share one or more references and co-citation analysis, where one measures the similarity between documents in terms of the number of documents cite each other [29]. In one of the experiments for the XLike project¹, newswires citation considered reference analysis and newswire bias was understood as a measure of frequency and was interested in three things: (a) how frequently do news publishers cite a news agency, (b) how frequently are individual news agencies cited, and (c) how biased are news publishers in citing a particular agency. [30]. In order to determine if an agency is cited in an article a set of most popular news agencies was identified (list of 30 news agencies from the corresponding Wikipedia page). For each agency a set of display forms in which the agencies can be mentioned was done: Agence France-Presse, can, for example, appear in text as AFP, Agence France-Presse or Agence France Presse. All articles from the set of news publishers were then analysed in order to identify if they mention any of the news agencies. The results of the experiment showed that the USA Today (32%), The Moscow Times (26%) and Fox News (26%) are the news outlets that have the highest percentage of articles that mention at least one news agency. On the other hand, El Mundo (1.8%), Gizmodo (1.9%) and Financial Times (2%) very rarely mention any news agencies. The Financial Times and Gizmodo – a famous business news publisher and a technology blog, both produce specifically tailored content to the readers that is mostly not reported by news agencies. Also the report showed that the most cited news agency across the 30 selected news publishers was the Associated Press – the oldest and the largest news agency in the USA. It was cited in 4.71% of all articles from the tested news publishers. Next by frequency are Reuters from London (1.78%), French Agence France-Press (1.11%), German Deutsche Presse-Agentur (0.54%) and the Russian Interfax (0.32%).

Research in this field has been also shifting towards plagiarism detection or reuse material of the press agencies. Paul Clough measures, for example, text reuse, which helps to not only detect bias, but also detect plagiarism, breaches of intellectual property and control in journalism e.g. how much text is reused from the well-known British newswire AP [31]. One simple approach is to find the amount of words overlap between two or more documents (e.g. finding document similarity). One of the main experiments concerned measuring text reuse in journalistic domain and finding the source most dependent on the AP. Using manually classified corpus and after having automatically gathered potential reuse candidates, the texts were classified into the three categories: verbatim (copy), rewrite and new. Dotplot program [32] was used to find similarities and to present the results, where the texts are split into N-grams (characters or words) and comparisons are made for all N-grams and a black dot is placed on the graph every time a match exists. He concluded that out of 770 newspaper articles, 605 were derived in some way from the AP and 165 written independently [31].

2.5 Geographical analysis of news

News is one of the main ways we learn about the world i.e. different geographical places we live in. News outlets have a certain geographical focus and have been traditionally divided into local (regional or city based), national (circulating throughout the whole country) and international (including international editions of many national or local publishers). Geographical analysis of news is understood as the coverage of events coming from certain countries news outlets report about. The main idea of analysing geographical bias is to uncover unequal geographical distribution and intensity of the coverage of media attention to certain events in certain

¹ http://www.xlike.org
countries. The imbalance of global information flows contributes to the analysis of geographical bias. In spite of the fact that globalization has brought to us the potential access to world’s media and information, international news coverage continues to be vulnerable [33]. When looking at geographical bias one usually asks the following research questions: 1) Is there an evidence that the closer an event occurred to the geographical presence of the news source, the more attention it will get? 2) How does the attention shift over to time? Do stories closer to a news source focus have a longer tail? 3) Is it possible to identify any ties between economical dependencies of one country to another?

In order to study geographical bias, one has to identify the event location or geographical entities for a given article from a particular source, which can be done in several ways. One most common type is to identify in the article a dateline (the brief piece of text at the beginning of the article) that usually mentions a location [30]. Another example is to use Geoparser (algorithm that assigns geographic identifiers, e.g. codes of geographic coordinates expressed as latitude-longitude, to words and phrases) to correctly identifying the geo place of news articles at the country level. Also to understand the geographical relations in terms of publishing among countries, the chi-square test is often used to measure statistical independence between two variables (variables are stories published in a country). One measures the statistical independence of the media content that appeared in each country compared to the media content in the other country [34].

In a critical report about the geography coverage of Boston Marathon Bombing, D'Ignazio et al concluded that crisis reporting is not equal and the world often turns to the high-attention crisis. The closer a crisis occurs to the source’s geographical area, the longer it pays attention to it and a part from geographical proximity, such factors as Geopolitical ties, GDP etc. influence the geographical distribution of news [35]. Another interesting example is the study of understanding news geography of disasters by Kwak and An [36], where they divided the world into seven regions according to the geographical proximity, then mapped more than 10,000 news media into one of the seven regions based on the classification from Alexa² (a web service that provides analytical web traffic data and rankings) and calculated news geography of disasters in the following way: defined the attention of a region, $r_i$, to a country $c_j$, as the number of the disasters occurred in $c_j$ covered by news media of $r_i$. They used the notation, $N_{ri} = \Rightarrow c_{j}$, for representing the attention of $r_i$ to $c_j$. Then, they defined news geography seen by $r_i$ as $N_{ri} = \{N_{ri} = \Rightarrow c_1, N_{ri} = \Rightarrow c_2, N_{ri} = \Rightarrow c_3, \ldots, N_{ri} = \Rightarrow c_K \}$ where K is the number of the countries. As a result of the research study, they concluded that there existed a clear difference of the news geography across regions: every region is overrepresented by the corresponding region, thus, for example, disasters in Latin America are most reported inside the region of Latin America and are not reported in other regions [36].

26 Speed of reporting bias

Detecting news bias in regards to the speed of the news reporting among publishers is another common practice. By speed of reporting is understood how much time in hours passed between the first article about a certain event was published and picked up by other publishers. It measures roughly how quickly a publisher produces an article about an event in comparison to other publishers. Castillo et al. refer to this phenomenon as timeliness and duration [37]. They analysed how different providers cover a story over time and measured two phenomena: how often is a news provider among the first ones, and how long the provider keep broadcasting the story and their experiments proved that CNN and Fox News are among the first ones when reporting breaking news, and the longest duration of articles being online is

² http://www.alexa.com
found in sports providers, followed by business [37].

There is common assertion that news first are produced by news media and then picked up by blogs and circulate in blogosphere for some time. To determine the time lag between news media and blogs Leskovec et al. concluded when analysing more than 1.5 million of news websites and blogs that typically the median in the news occurs first, and then a median of 2,5 hours later than the median of the news, the piece of information occurs among blogs. Moreover, news volume both increases faster and higher, but also decreases quicker than blogs volume. In blogs, it is exactly the opposite. An issue is discussed much longer in blogs, but has a 2,5 hour lag to be picked up [38].

2.7 Similarity in the coverage of events

The hey aspect in understanding news bias is to compute how biased the news publishers are about which events they report about and how much do the events covered by the two publisher overlap. One of the ways to compute it is to use various similarity methods.

There are different similarity measures that we could use to compute the similarity between two sets of events. One option would be to use the Jaccard index, which computes the ratio of the size of the intersecting elements over the size of the union of elements. The downside of using this measure would be that for pairs of publishers, where one would have significantly more events than other, the resulting similarity would therefore inevitably have to be small. A better measure that “normalizes” the differences in publisher’s size would be cosine similarity. In the same way as we can use it to compare two documents, we can use it to compare two vectors of event IDs. Since the cosine similarity normalizes the dot product of the two vectors by their lengths, the importance of the more “active” publishers would be appropriately reduced. There is also a so-called Named Entity based similarity measure, which uses Name Entities detected in an article to measure the similarity [39]. When comparing two articles A and B, two vectors representations a and b of their NEs, where ai is the weight of NE i in document A (analogous for B). The similarity between the documents is then calculated as the cosine similarity of the vectors, given by Formula 1.

\[
Sim(A, B) = \frac{\sum ab}{\sqrt{\sum a^2} \cdot \sqrt{\sum b^2}}
\]

When no NEs are detected in an article, one can revert to the classic “bag of words” approach, using Term Frequency - Inverse Document Frequency (TF-IDF) weights for every word in the text [39].

Many attempts have been made to compare similarity of news outlets among different countries. Leban et al. have attempted to analyse similarity between publishers in choosing the events they report about using the Jaccard index, which computes the ratio of the size of the intersecting elements over the size of the union of elements [30] and revealed that best agreement on what to report about was seen between the European set of publishers, for example, BBC, The Independent, the Daily Mail etc. and explain it thanks to their close geographical location. Larcinese et al. investigated the intensity of coverage of economic issues and whether there is any significant correlation between the endorsement policies of newspapers. They tried to classify articles according to the topic covered, without attempting to classify whether it is negative or positive. They exclusively concentrated on the agenda setting of several newspapers and their approach was based on simply automatic keyword search. They found out that newspapers with Democratic endorsement pattern systematically give more coverage to high unemployment when the president is Republican [40].
Another focus in this area is the detection of the most popular articles across different news sources. The stories are collected and then ranked using Ranking SVM [41].

2.8 Topic similarity bias

Another topic, loosely connected with the previous one of similarity in the coverage of events, is topic or category similarity, referred by some as topic selection bias [30]. Scholars take a closer look at what topics are given more preferences and more coverage than others in the selected sources. Several studies have produced confirmation of topic similarity across various sources. Flaounas [42] in his PhD dissertation measured the so-called topic selection bias. Each outlet was represented as a vector of several dimensions, one for each topic. The value of each dimension is the number of articles of the outlet for a given topic and then it was normalized over the total number of topics. He then measured the Euclidean distance (distance between two points in Euclidean space) of these vectors and used multidimensional scaling to plot them. Outlets that cover the same topics were displayed closer to each other, and those further apart were the ones reporting on different topics.

Leban et al. [30] analysed coverage bias of top level categories taken from DMOZ taxonomy, which is a human built taxonomy with over 1 million categories, for several selected publishers. As expected, one can see that the subject-matter outlets focus on the topics they are dedicated to. For example, technology websites, such as GigaOm, the Next Web or Gizmodo, gave most of their coverage to topics like Computers and Shopping since their main focus is on news related to consumer electronics. In a similar fashion, the economic and business news publishers, such as Business Insider, WSJ Blogs, Wall Street Journal, Economic Times and Boston business Journal, paid very little attention to Sports, Health and Home issues which lie outside of their main scope. Overall, they concluded that most news websites are heavily biased towards topics they (journalists, editors, editorial teams, etc.) and the public they are writing for, are interested in.

Saez-Trumper measured the distribution of coverage of stories in regards to people, by measuring the distribution of number of mentions per person across different media and different languages. The distributions were then compared using the so-called Jensen-Shannon divergence (a popular method for measuring similarity between two distributions) between them for each pair of news sources and concluded that the importance given by people to different issues tends to be correlated with media as a whole rather than with a specific media source each person follows [43].

2.9 Gender Bias

Gender bias, already a highly evident and yet challenging problem in social sciences, refers to studying the mentions of men and women and their relations across different topics in various media channels. Gender bias nowadays studies how media today construct femininity, masculinity and gender relations [44]. There has been many studies in media that reflect the changing environment and representation of women and some claim that media nowadays both promote ideas of feminism and no longer represent women as housewives spending time in the kitchen and cleaning the house [45] and on the other hand there still exist patterns and topics where women are almost unrepresented [46], [47] or even more drastic existence and

---

3 [http://www.dmoz.org/](http://www.dmoz.org/)
persistence of patterns of sexism (very little coverage of women of middle age or elder women, as an example) [44].

Several works have been carried out to examine the question of gender bias automatically [48], [49]. The systems usually automatically count the gender labels: male, female and unsure for each article/cluster/topic and simply calculate the frequency. It has been proven that references to men outnumber references to women by three to one [1]. Their study also concluded that articles about sports and business are among the most gender-biased and on the contrary entertainment is the least gender-biased category [1].

2.10 Sentiment analysis

Research in the field of automatic detection of news bias also has been shifting its attention to sentiment analysis and opinion mining in the news. Most sentiment and opinion mining analysis has been done on very subjective texts like product launch, movie reviews or blogs, where the opinion of the author is expressed freely in a very subjective and biased way. Recently, sentiment analysis has been drawing attention to news articles, where an opinion of a journalist should not be, but often is present.

There are many different approaches and different understanding currently being adopted on bias-related sentiment in news. The first approach is often based on subjectivity lexicons [50] dictionaries of words associated with a positive or negative sentiment score, also referred to as polarity and other biased loaded adjectives [51]. Most researchers on sentiment analysis refer to SentiWordNet4 – publicly available English lexicon that classifies words in negative, positive and neutral. Such lexicons can be used to classify phrases, sentences or documents as subjective or objective, positive or negative. Another approach on sentiment analysis often employs machine learning text classification [52] and other NLP techniques, such as, speech tagging [23], co-reference resolution [53] or Latent Dirichlet allocation – a probabilistic topic-modelling tool based on three-level hierarchical Bayesian model [54] that extracts latent topics (distribution of words) from news feed. LDA together with Antelope – another NLP tool used to measure the semantic framework that incorporated sentiment bias on a diverse corpus of political news [55].

A recent approach to study quote phrases or memes in order to identify bias of a website and to predict republican vs. democratic bias of news websites and political blogs was proposed by Sonia Gupta [28]. She states that there are discriminatory memes that are quoted by similar biased websites, exploiting the quoting pattern of memes. A simple iterative algorithm that computes bias of websites and memes according to their neighbours works well in predicting. She assumes that memes are hubs and websites are authorities, runs the algorithm for both labels independently and then normalizes the scores in each node. The conclusion was that many names of websites are already predictive of their political bias e.g. stopbarackobama.com [28].

Tim Groseclose and Jeffrey Milyo also have measured media bias by estimating and assigning ideological scores for several American media outlets. They computed the number of times a particular media outlet cites various think tanks and policy groups listed in the website called »Where to do research«5 and then compare it with the number of times the members of the American Congress cite the same groups/politicians. Their results revealed a strong liberal bias in most of the online outlets under observations [56].

4 http://sentiwordnet.isti.cnr.it
5 www.wheretodoresearch.com
Chapter 3. Critical judgement of the existing research

Most of the research studies described in Chapter 2 »Text mining for News Bias analysis« have achieved interesting, fulfilling and state-of-the-art results in the field of NLP, machine learning, data and text mining. It has been, however, clear from the overview of the state of the art that NLP research does not really rely on traditional social science research of news bias and their work is still at the beginning stage. There are several issues to be critically judged and addressed.

As mentioned in the beginning, the main aim of this work is to able to build a system that automatically detects bias in all its complicated forms comprising cognitive and social science research similar to that a human would detect and in order to do that one has to start with a clear definition of news bias that comprises several disciplines.

Chapter 2 have analysed the state of the art work and there are several conclusions to be drawn:

1) a better understanding and definition of bias comprising cognitive and social sciences is needed.

2) whereas traditional media studies of news bias have often concentrated themselves on news values [59], [60] and influential factors like political affiliation, competition, pressure from advertisers, the perspective of NLP and machine learning approaches have so far very little to do with it. A better multidimensional framework on studying news bias is necessary.

3) whereas most social studies of news bias have emphasized relationship between the owner of a media house and the media content, the automatic approaches have been almost lacking this aspect.

4) traditionally, media studies of news bias have affirmed that news bias can be both individual or of an organization. Indeed, the cognitive perspective on bias was often taken into account; the automatic approaches are often too fragmented and broad. A clear distinction is needed between an individual bias of a journalist and of the media house.

5) regarding sentiment analysis, news discourse is not only fixed on positive vs. negative, subjective vs. objective. A negative opinion of a certain speaker does not necessarily mean that the whole piece of information is negatively biased. Both sides (positive and negative) as a balanced reporting can and should be present in a news report. In addition, a position of a speaker or journalist can be communicated without explicit expression of opinion or sentiment, it can be, for instance, conveyed through objective sentences that include or exclude facts. A better comprising framework (e.g. taking into consideration discourse analysis or critical discourse analysis) should be applied.

News bias is a very complex problem. It is a challenging task not only for a computer scientist, but also for a human to purely concentrate on detecting news bias to its fullest form. There is a lot of room for working on the intersection between the social and computer sciences.
Chapter 4. Future work

Despite numerous attempts to automatically detect news bias and its various types, there is clearly a need to combine and work on the intersection of media studies and computer sciences, both fields can only benefit from it. Before working on automatic detection of news bias applying some frameworks from media studies, there is a need to simply describe how media system and journalists work, how they select, process and publish the information and then take a step back and look at what news is and what makes a piece of information newsworthy.

In the following PhD study, the author understands news bias as complex and often contradictory set of processes in the following stages of news making: (1) news selection and newsgathering, (2) news writing, (3) editing and publishing. News is constructed realities: nearly all of them are socially constructed [Fowler]. News has a very layered character; the earlier versions of pieces of information are constructed by later versions, decontextualized, embedded and then put into publishing. The production of news, including news selection, writing, editing, is very biased process in nature [57].

In order to understand a complex process of news biases, the first step is to understand what makes a piece of information newsworthy to be published. The audience has been always eager to know how the news is selected and structured. How does a journalist decide what is newsworthy and what is not? What gets published and what gets left out? Not all events that happen in the world everyday get into news, events have to coincide with certain characteristics or criteria determined by media organizations. This criterion is called news values. Events have to be sufficiently interesting to be reported in a newspaper, online, radio or TV, they have to meet certain criteria. They have to be newsworthy.

News values can be defined as certain rules or guidelines for journalists to follow in producing a news story. They are imagined preferences of what the audience expects. They were built over time. They are also considered to be practice-based and ideological factors in understanding news stories and decisions of journalists [58]. Analysis of news values and selection of news is one of the most widely known area of research in journalism and media studies in general because it tries to understand what is include and excluded in news and why. However, there has been very little attention paid to automatically detect the values. There have been several attempts to list news values and selection criteria. Galtung and Ruge [59] made the earliest attempt to propose, still greatly perceived nowadays, a typology of 12 news values:

- **Frequency** – the time distance, time-span of an event and how often it is in the news
- **Threshold/relevance** – the size, impact and/or the intensity of an event
- **Unambiguity** – clarity of the meaning of the event to the public
- **Meaningfulness/Proximity** – the event is of great value and meaningful to the audience if it culturally and geo-politically close to the location of the audience
- **Consonance** – the event should match conventional expectations of the people and be harmonious.
- **Unexpectedness** – the event has to happen unexpectedly and unplanned
- **Continuity** – the event should be continuous and connected over a period of time
- **Composition** – the event should be balanced and complemented by other pieces of information, citations etc. to form a unified news event
- **Reference to elite nations**
- **Reference to elite persons**
- **Personalization** – the event is seen as actions of individuals, it should be personalized
affecting people, it should have a human interest

- **Negativity** – the event should report bad news; when it bleeds, it leads.

These factors are not independent on each other and are interrelated. The theory has not really been challenged despite the existence of several new lists of news values and the fact that it lacks considerations of the new media, most of the media researchers and journalists still regard the list as relevant. Since Galtung and Ruge, there is a greater change in journalism that brought to more self-awareness of news and its reliance on considerations of house style and substance. In other words, news medium manifests itself in various ways, but its effects are still coming from the nature of news values in practice.

Important to mention, there are stories that do not make it to news and if they get to the news they might be distorted or biased not only because of the news values, but also because of journalistic ethics, stylistic factors, space limitation, political affiliation of a media house etc. These dimensions of studying news bias can also be studied at a later stage.


Reference List


