FP7-ICT-2011.4.2
Contract no.: 288342
www.xlike.org

Deliverable D5.3.1

News reporting bias detection prototype

Editor:	Gregor Leban, JSI
Author(s):	Gregor Leban, JSI; Aljaž Košmerlj, JSI; Evgenia Belyaeva, JSI; Blaž Fortuna, JSI;
Deliverable Nature:	P
Dissemination Level: (Confidentiality)[1]	PU
Contractual Delivery Date:	M33
Actual Delivery Date:	1.10.2014
Suggested Readers:	XLike project partners
Version:	1.0
Keywords:	News bias, reporting, news, opinion

[1] Please indicate the dissemination level using one of the following codes:
• **PU =** Public • **PP =** Restricted to other programme participants (including the Commission Services) • **RE =** Restricted to a group specified by the consortium (including the Commission Services) • **CO =** Confidential, only for members of the consortium (including the Commission Services) • **Restreint UE =** Classified with the classification level "Restreint UE" according to Commission Decision 2001/844 and amendments • **Confidentiel UE =** Classified with the mention of the classification level "Confidentiel UE" according to Commission Decision 2001/844 and amendments • **Secret UE =** Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments

Disclaimer

Full Project Title:	XLike – Cross-lingual Knowledge Extraction
Short Project Title:	XLike
Number and Title of Work package:	WP5
Document Title:	D5.1.1 News reporting bias detection prototype
Editor (Name, Affiliation)	Gregor Leban, JSI
Work package Leader (Name, affiliation)	JSI
Estimation of PM spent on the deliverable:	17

Executive Summary

News bias is a ubiquitous phenomenon that has generated various research studies in different fields. Practically any media outlet can be biased, but the public should be aware of it and news bias should be minimized thereby offering more objectivity to the news reporting.

News reporting bias detection prototype is aiming to test old and new techniques that help to detect and foresee a potential bias of news reporting across 30 selected news outlets across the world: 26 news websites reporting in English from the USA, India, Russia, Canada and Great Britain, 2 Spanish and 2 German outlets. The articles that were analyzed in this study were published during the period between 15th of December 2013 and 15th of August 2014.

Experiments to detect bias in news reporting aim at analyzing supposedly objective news websites like BBC, the Independent and the Financial Times and include analysis of length, grammatical and readability differences, as well as geographical, topic-wise, reporting speed and citation biases. We then also make a first attempt to compare the coverage of the news events about Russia and Ukraine in the above-mentioned news sources, followed by similarity analysis between different publishers in choosing the events they report about and detecting the differences between biased and non-biased news sources.

The bias detection prototype confirmed the existence of bias in news reporting. The key findings can be summarized as following: Most types of detected bias are related to the geographic differences/similarities between the analyzed news publishers. European news sources, for example, put more emphasis on describing events occurring in Europe, while US publishers cite more the events occurring in US. Similarly, the news sources that are geographically close are harder to discriminate than those that are further apart. A similar pattern can be observed which is related to the citations of news agencies. US news sources mostly cite Associated Press (a US based news agency), whereas the European publishers prefer to cite the European news agencies. We have also detected a strong bias among the tabloid outlets such as Daily Mail or Stern Magazine to write longer headlines, shorter articles and to use more colorful language making use of numerous adjectives and adverbs. As expected, we also found a larger percentage of adverbs and adjectives on websites like Gizmodo, GigaOM and The Next Web which write about new products and their evaluations. In some news sources we have also noted a bias related to using proper nouns (names of people, locations and organizations). Finally, our report has shown that readability of articles from the analyzed publishers varies significantly. Most hard to understand seem to be the Spanish and German news articles which could be due to a high number of compound words.

Table of Contents

List of Tables

List of Figures

Abbreviations

LDA	Latent Dirichlet Allocation
MDS	Multidimensional scaling

Definitions

Concept	a named entity or a keyword identified with the annotation service.
Event	a group of articles that are identified to discuss the same happening in the world. The events discussed in this deliverable are identified by the Event Registry using a clustering algorithm.
Article	an article is a unit of news that we collect from various news sources. Each article is described by various features, such as the title, body, date and news source. This information is additionally extended with information produced by XLike services, such as detected named entities and keywords, categories, POS tags, etc.

1 Introduction

News bias can be potentially present in every news-reporting outlet off- and on-line. There are many examples of poor journalistic practice that can be seen every day: a journalist stating his personal opinion in a news report, asserting incorrect facts and figures, applying unequal space to different sides of a controversial issue, citing people of a certain class or gender. News are often written and used to manipulate, to advertise, to create fear among the populations or to follow a certain media agenda or an ideology. News presents already a relatively shaped worldview, where beliefs and values cannot be avoided. [HJ06]

In the recent years news bias, already a highly evident problem in the media field and social sciences, has also inspired many scientific research in computer 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 analyzed 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 [AO10] 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 science, which is trying to ease the workflow of many researchers in social science and to automate the analysis of large amounts of data.

Our study was inspired by an interest to explore news bias in 30 selected online websites from all around the world. The corpus of the study encompasses 26 online news websites reporting in English, including for example, CNN Europe, Fox News and Financial Times. Majority of selected news sources are in English language, except for 2 Spanish news sources (ABC and El Mundo) and 2 German (Stern Magazine and Die Welt). Cross media elements like video or pictures, which are a common part of any online media, can also be a source of news bias, but were not included in our analysis. In this report we have chosen to analyze a broad spectrum of biases. We start by analyzing length, grammatical and readability differences, and continue with analyzing geographical, topic, reporting speed and citation bias. We also analyze how similar are the publishers related to the events they report about as well as how well we can distinguish between them based on the article text. We also analyze content from Wall Street Journal and WSJ Blogs to test how well we can differentiate between the two sources where one is expected to be more biased.

It is important to identify news bias and report about it to the general public, which is exposed to news every day and might be making the sense of the world accordingly. For the public to make decisions on politics, economics and other social issues, they must be given objective and neutral information. We must underline, however, that it is beyond the scope of this report to define what is a fair and just reporting or journalism. Our goal is only to identify and highlight the differences that we were able to identify in different news sources.

This report is broken down into six chapters, although some chapters entail several smaller chapters. We begin our report with a definition of news bias identifying the most significant methods of automatically detecting it and then describe our corpus, news sources and the main reasons for choosing them. Next we present our experiments briefly and later on we describe in detail each of the experiments and discuss their main results. Finally we examine possible future work in the field and conclude by reviewing the outcome of the experiments.

2 News bias

"What is a news bias?" and "What makes a news piece bias?" are the questions we are confronted with almost every day. An online Cambridge dictionary defines bias as "a tendency to support or oppose a particular person or thing in an unfair way by allowing personal opinions to influence your judgement"[2]. Media bias is a universal concern. Despite the fact that newspapers and reporters or 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.

Having defined the news bias, it is important to note that there are several types of reporting news bias in media studies, but there is no unique classification. There is usually a common agreement of the media researchers on the following types specifically: bias by omission or exclusion, by news source, labelling of the main participants in the news, bias by position on the webpage or a newspaper and ideology bias.

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 also addressed automatically, deploying various analytical techniques, due to the availability of large amounts of online news in various languages.

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 [FI10]. The detected 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. 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 [AO10].

In our experiments we also wanted to focus only on the examples of news bias that can be detected algorithmically. The experiments that we performed and will be described in next sections include the following types of bias:

- Analysis of article length differences
- Analysis of grammatical differences
- Readability differences
- Geographical bias
- Topic coverage bias
- Speed of reporting bias
- Newswire citation bias
- Similarity in the coverage of events
- Content similarity

Research in the field of automatic detection of news bias also shifts 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 of news articles, where an opinion of a journalist should not be present, is getting more attention. Two examples of such work include a lexicon- based approach to analyse sentiment of quotes in different newspapers [BA10] and news bias analysis of finding over- and under-stated facts of a particular news outlet [PS09].

[2] . http://dictionary.cambridge.org/dictionary/british/bias

3 Data description

This report conducts a detailed analysis of reporting news bias based on a corpus of 1,289,402 articles from a selected set of news publishers. We examined a period of nine months, between 15th of December 2013 and 15th of August 2014. The time range was selected based on the availability of the data. The articles are written by 30 news publishers, which are considered as one of the most influential daily news websites across the world. We have internally divided 30 selected websites into four main categories:

- International news websites reporting in English, but having headquarters in different parts of the world: Australia, USA, Canada, Great Britain, Russia and India (Sydney Morning Herald, CNN Europe, Fox News, The Guardian, USA Today, The Hindu, Huffington Post, ABC News, Washington Post, Time, Globe and Mail, Daily Mail, BBC, Daily Telegraph, the Moscow Times, DNA India, The Independent)
- Six business websites (Boston Business Journal, Financial Times, Business Insider, Economic Times, Wall Street Journal and Wall Street Journal Blogs)
- Foreign news outlets reporting in German (Stern Magazine, Die Welt) and Spanish (ABC.es, El Mundo) – the two languages represented the most in our system
- Three websites concentrating on technology (GigaOM, the Next Web, Gizmodo).

In our experiments we use color to differentiate between different categories of news sources in the graphs. We use the following color schema: red color for international English news website, light green for the business websites, cyan color for foreign news publishers reporting in Spanish and German, and blue color for technology-related news publishers.

We chose to focus on the above-mentioned media outlets for various reasons. They are all well-known to most readers in their home countries and abroad, easily accessible, and play an important role in shaping the public opinion in the reported country. For example, die Welt in Germany is the most popular online website. Similarly, the Moscow Times is the most read news website reporting in English in Russia.

The 26 websites reporting in English were selected for our experiments with the main idea to cover different geographical continents in order to be able to identify any geographical news bias.

The data that was analysed in these experiments was obtained from the Event Registry [LG214, D432]. The core available information for each article included the title and the body of the article, the news source, date of publication and the list of annotated concepts. Since Event Registry identifies which event is mentioned in each article we were also able to obtain for each article the unique ID of the event as well as the event information. One of the properties, computed for each article by Event Registry, is also the article category. Since it is used in several experiments we describe more details in Section 0.

Data statistics per publisher

The Table 1 shows for each new source the number of collected and analysed articles as well as the number of events that were identified from them. The number of events is always lower since it is common that different articles from the same news source will report about the same event.

Publishers	Number of articles	Number of events
Boston Business Journal	183,361	51,360
Daily Mail	130,723	68,053
BBC	75,500	49,576
Economic Times	64,620	30,655
USA Today	60,724	31,410

The Guardian	59,262	41,725
Wall Street Journal	54,687	24,305
Globe and Mail	52,076	25,217
Huffington Post	49,042	30,551
Washington Post	47,807	29,144
Die Welt	47,349	28,501
Daily Telegraph	45,640	32,061
The Independent	41,719	29,377
CNN Europe	41,319	15,025
ABC.es	37,147	27,695
DNA India	32,169	20,506
FOX News	29,229	19,951
Stern Magazine	27,570	10,945
ABC News	26,584	22,563
Sydney Morning Herald	26,174	14,717
El Mundo	25,958	21,440
Business Insider	25,082	14,024
The Hindu	24,141	15,497
Time	21,922	10,697
The Moscow Times	13,933	5,687
WSJ Blogs	13,630	6,387
Gizmodo	10,591	6,008
GigaOM	8,046	4,307
The Next Web	7,207	3,999
Financial Times	6,190	4,202

Table 1 Number of articles and events per publisher

Article categorization

Articles that are written by news publishers are about various topics – from sports events, gossip and entertainment to various social issues. There is no known categorization taxonomy that would be "The taxonomy" to use for categorizing news articles. Many news publishers do have their own taxonomies but they are kept private. In order to categorize the articles we decided to use the DMoz taxonomy[3], which is a human built taxonomy with over 1 million categories. It is used to categorize web pages but is a good enough fit also for categorizing news articles. These are the top level categories in the taxonomy together with their main topics/keywords:

- **Arts**. Movies, television, music, entertainment, theatre, dance.
- **Business**. Banking, investments, financial services, transport, real estate, employment.
- **Computers**. Internet, software, hardware, programming, robotics, AI, hacking, mobile computing.
- **Games**. Video games, board games, gambling, card games, puzzles.

[3] http://www.dmoz.org/

- **Health**. Alternative health, medicine, conditions and diseases, pharmacy, addictions, weight loss, nutrition.
- **Home**. Cooking, personal finance, family, home improvement, gardening, homemaking.
- **Recreation**. Humor, collecting, food, guns, autos, travel, boating, climbing, parties.
- **Education**. Museums, libraries, exhibitions, dictionaries, maps, bibliography.
- **Science**. Technology, math, physics, biology, astronomy, social sciences, earth sciences.
- **Shopping**. Gifts, toys, food, visual arts, vehicles, crafts, auctions, clothing, music.
- **Society**. Government, military, issues, religion and spirituality, law, crime, holidays, history.
- **Sports**. Events, motorsports, volleyball, softball, football, basketball, hockey, cricket, water sports.

Using the manually assigned web-pages for each of the categories we were able to build a classifier that is able to classify each article into one or more categories. When computing a category for an article we only consider the categories up to three levels deep. This corresponds to about 5.000 possible categories.

In several of the following news bias experiments we use the article categorization to compute some property separately for articles in different categories. Due to large number of possible categories we mostly compute separation on the top level of the taxonomy only. Since the taxonomy is hierarchical, the computation naturally also includes articles that are assigned to any of the sub-categories. When, for example, we are analysing articles from BBC and computing a property related to articles in the Business category, we also include in computation the articles assigned to Banking, Investments and other subcategories inside Business.

4 Detection of news bias

In this section we present different experiments that were carried. For each experiment we explain how the experiment is related to news bias and describe our main findings.

A common way how the results are displayed are in a figure displaying box plots[4]. For each news publisher the box plot displays different properties of the data distribution. As it is usual with the box plots, the left and right edge of the boxes represent the first and third quartile of the, whereas the bold line inside the box represents the second quartile (the median value). The whiskers (vertical lines outside of the rectangles) can represent several possible alternative values. For our purposes we decided to represent the 9th and 91st percentile.

Along with the graphical display of results we also included results in the numerical form. Due to a large number of tables we have put them in the Appendix 1. Along with the first, second and third quartile (Q1, Q2, Q3) we also display in the tables the average values and the standard deviation.

Analysis of length differences

In this experiment we focus on analysing the length differences between the selected publishers. We computed how the publishers differ in regards to article lengths, sentence lengths and headline lengths.

Figure 1 shows results for article lengths for different news sources. The lengths are expressed as a number of tokens[5] which roughly corresponds to the number of words in the article. CNN Europe with an average number of 726 tokens in articles, together with Financial News (748 tokens) and The USA Today (758 tokens) are the web outlets publishing the longest articles. Interesting to note that mostly American publishers do not limit their journalists on writing space. On the contrary, Boston Business Journal (228 tokens), Stern Magazine (366 tokens) and BBC (348 tokens) are the online sources with shorter length of article on average. Shorter articles from Boston Business Journal can be due to a different, business-like, very factual content that does not require a lot of space. BBC, as one of the most trustworthy, objective and quality news outlet in the world, has always been characterized to have short, objective, and factual and to the point news pieces, which was proven by our experiment. BBC journalists write tight, which means they know how to convey as much information as possible in few words. Longer, detailed articles from BBC are often only about very important national or international events.

[4] http://en.wikipedia.org/wiki/Box_plot

[5] https://en.wikipedia.org/wiki/Tokenization

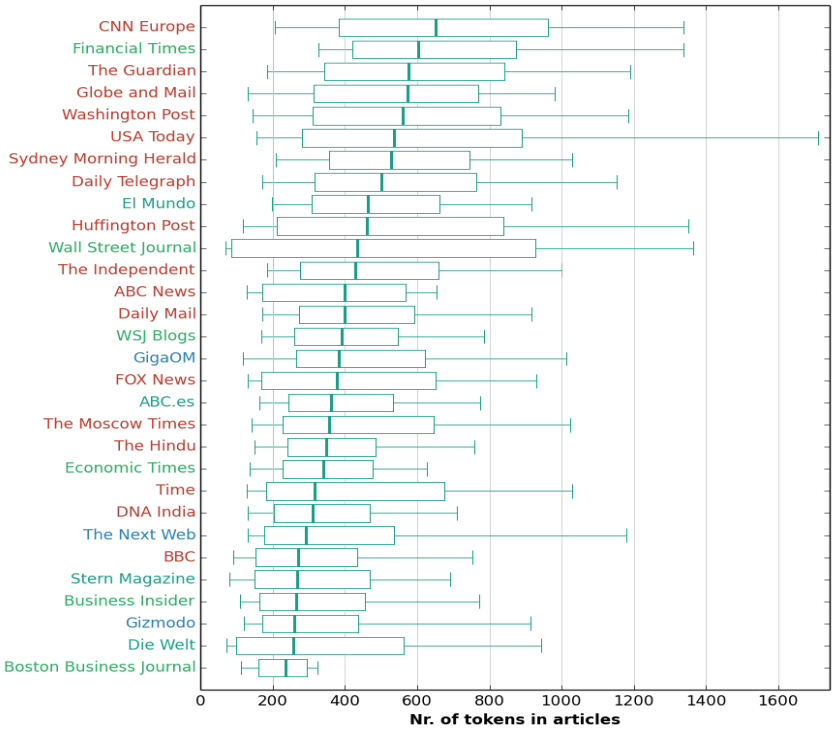

Figure 1 Number of tokens in articles

The Figure 2 shows the lengths of articles from the Huffington Post separated by top level categories. The following categories seem to have longer articles, which can be interpreted as of more importance to the editorial team of Huffington Post: Health, followed by Education, Home and Society. These categories are issues that affect reader's lives and are considered to be soft news and can be typical for popular and tabloid news outlets. Less attention is given to articles related to Recreation, Shopping and Arts.

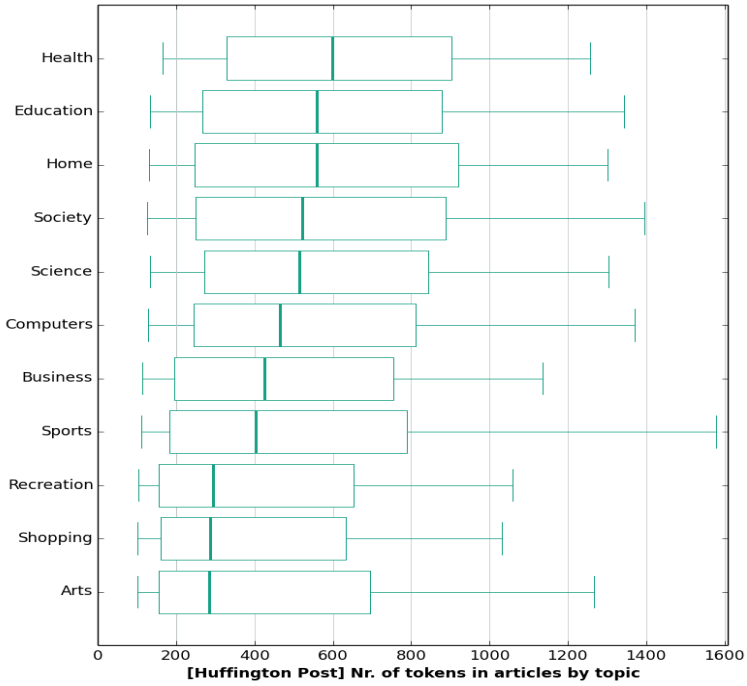

Figure 2 Number of tokens in articles by topic in Huffington Post

Another example showing the article lengths by category is shown in Figure 3 and displays the Indian news publisher The Hindu. Contrary to the results of Huffington Post, the editorial team of the Hindu gives

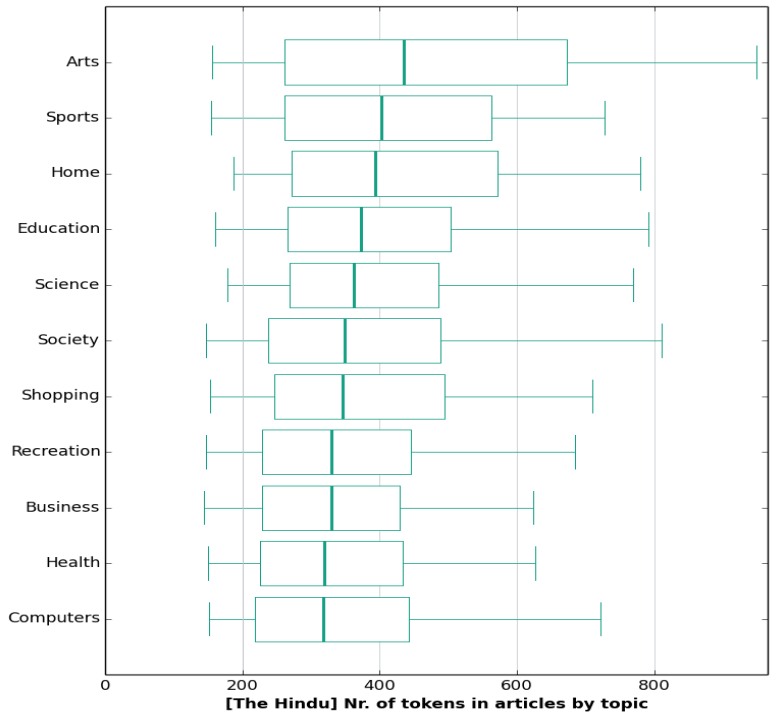

Figure 3 Number of tokens by topic in The Hindu

preference in terms of article length to the Arts, Sports and Home categories. Shorter articles are written when writing about Computers, Health and Business.

For each news source we also made a comparison of the number of tokens in article headlines. The headlines are the most read part of news articles and are the basis for how the stories will develop. They can summarize the content of the whole article, steer the attention and thus convey some hidden bias. Figure 4 shows the results of this experiment. The longest headlines are in the Daily Mail, The Independent and The Next Web. The shortest are written by BBC, CNN and The Hindu.

To catch the reader's attention a headline should be appropriate for the type of news publisher. Public service news publishers like BBC and CNN tend to write shorter and simpler headlines with more formal and specific language when reporting hard news focusing on politics, economics, disasters, accidents, etc. Professional journalists are aware of the fact that readers are typically short of time; they want to be able to scan the headline to get the information they need. On the other hand, private news outlets like Daily Mail or Huffington Post do not use telegraphic headlines, but rather long, descriptive, ambiguous and even emotional titles and do not necessary pay attention to the space dedicated to them. Headlines of a tabloid newspaper like Daily Mail use very emphatic and sensational words intended for their segment of readers.

An example below is taken from the Daily Mail (the news source with the longest headlines) and the BBC (with shorter headlines) announcing the same event to the readers on the 6th of August 2014.

Daily Mail, 6 August 2014:

Shameless Costa Concordia captain gives a lecture on "panic control" as his manslaughter trial over the death of 32 people aboard his liner continues.[6]

BBC, 6 August 2014:

Costa Concordia captain's lecture sparks outrage.[7]

[6] http://www.dailymail.co.uk/news/article-2718251/Shameless-Costa-Concordia-captain-gives-lecture-panic-control-manslaughter-trial-death-32-people-aboard-liner-continues.html
[7] http://www.bbc.com/news/world-europe-28679906

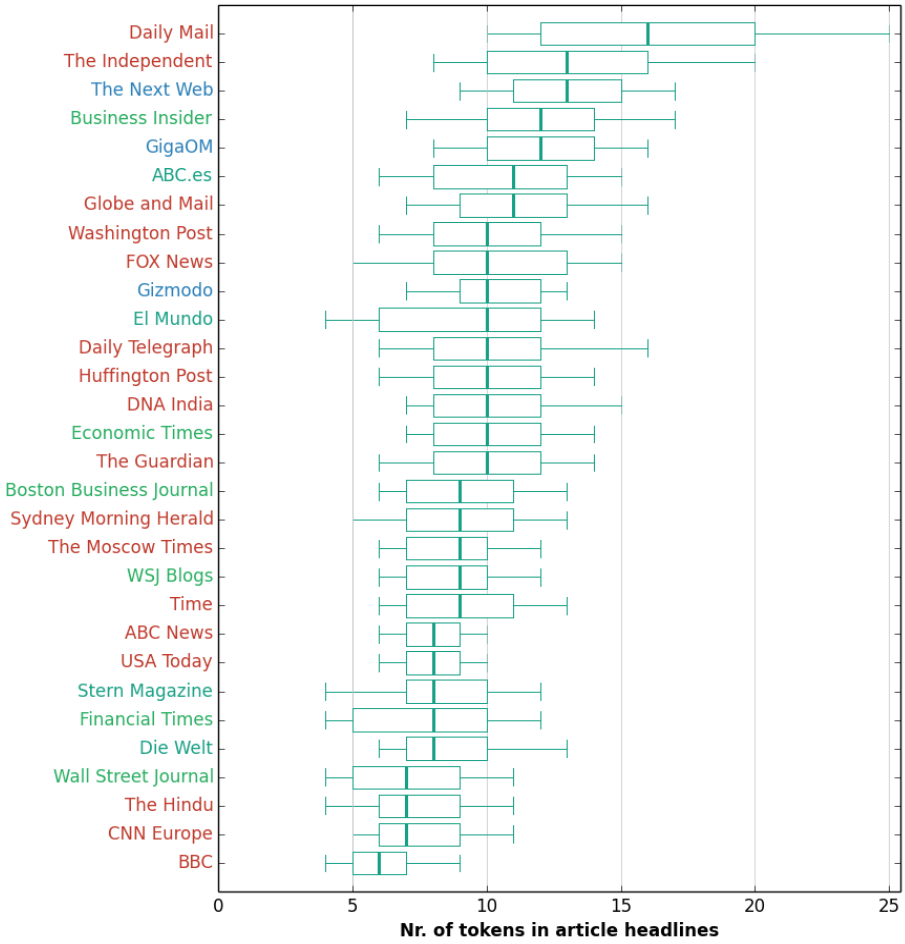

Figure 4 Number of tokens in article headlines according to the selected publishers

Analysis of grammatical differences

In this experiment we focus on grammatical differences between the selected 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.

Adjectives have always been directly linked to studying of 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. A professional journalist, instead, knows from a journalistic school that any news story should be based on the so-called essential 5 W's: Who, What, When, Where, and Why or How and that he should make use of strong verbs, but fewer adjectives. The use of adjectives is not a problem in an opinion piece, but is not desired for an objective news piece.

As seen from the Figure 5 the number of adjectives (in percentage of article tokens) used in articles is not very high, partially proving the fact that we are analysing news websites and not opinion pieces, blogs or comments. The two German news publishers Stern Magazine and Die Welt show a tendency to use more adjectives, whereas the DNA India and USA Today tend to use less adjectives. German outlet Stern Magazine has a reputation of a popular, tabloid newspaper and the reference to more adjectives is not surprising. The websites like Gizmodo, GigaOM and The Next Web also show a relatively high percentage of adjectives. This is also expected, since they write about products and their evaluations and are therefore forced to use more adjectives and adverbs as seen in the Figure 6 as well.

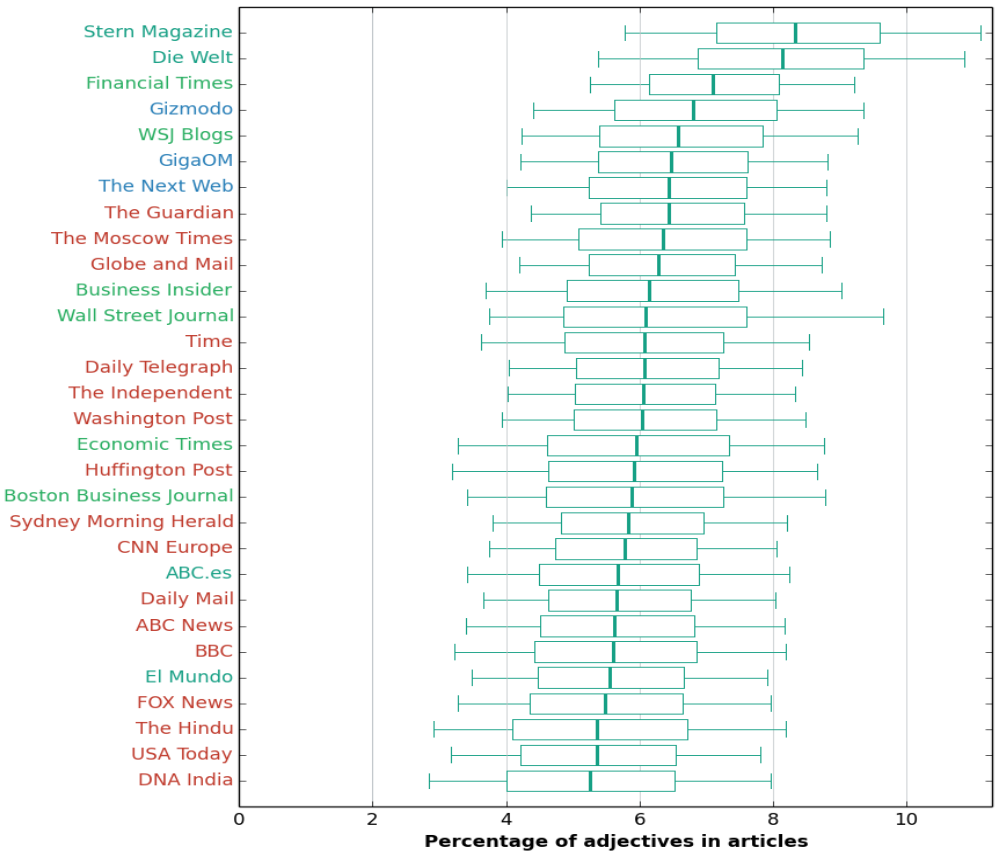

Figure 5 Percentage of adjective in articles per publisher

Figure 6 shows the results for the number of adverbs (again in percentage of article tokens) used in the articles per individual news source. Here too, we see on top a very similar set of publishers as for the number of 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. The abuse of adverbs can simply annoy the reader and should be used in moderation. For example, two business outlets Boston Business Journal and Wall Street Journal, which focus their attention on marketing and financial news, tend to use fewer adverbs because they do not want to clutter serious pieces of information and want to keep high journalistic standard for a very specific public.

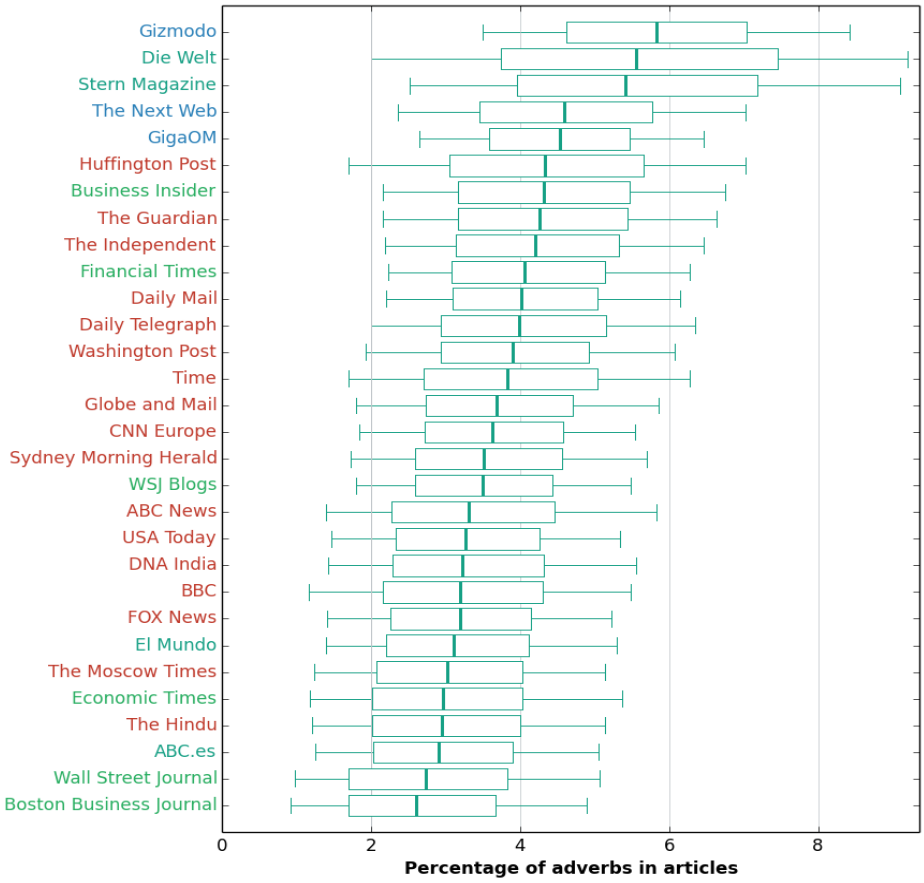

Figure 6 Percentage of adverbs in articles per publisher

A type of part-of-speech that we also tested are proper nouns, which indicate names of people, places and things. The use of proper nouns is a vital part of journalism and storytelling. Journalists should tell the readers what is new, interesting and specific about a certain piece of information and this cannot be done without mentioning people, organizations and locations. Reference to people in journalistic writing is a great way to bring an expert's opinion on a certain topic in a news article and to quote that person. Proper nouns also give validity, accuracy and clarity to the information provided in the article and they often help to hold the readers' attention. They also provide a sense of reality to the readers, as if they are directly present at the event.

From the results in the Figure 7 we can underline that USA Today, Boston Business Journal and Wall Street Journal are among the publishers with the highest number of proper nouns in the articles.

Contrary to the previous results on the higher percentage of the usage of adjectives and adverbs, the two German news outlets die Welt and Stern Magazine use less proper nouns in comparison to other selected sources. The lack of proper nouns in news reporting definitely does not add credit to die Welt or Stern Magazine as a valid source of information. Their news stories are more static, rather than alive without the mentions of people, places and organizations. This might be once again an example of bias and non- effective, poor journalism.

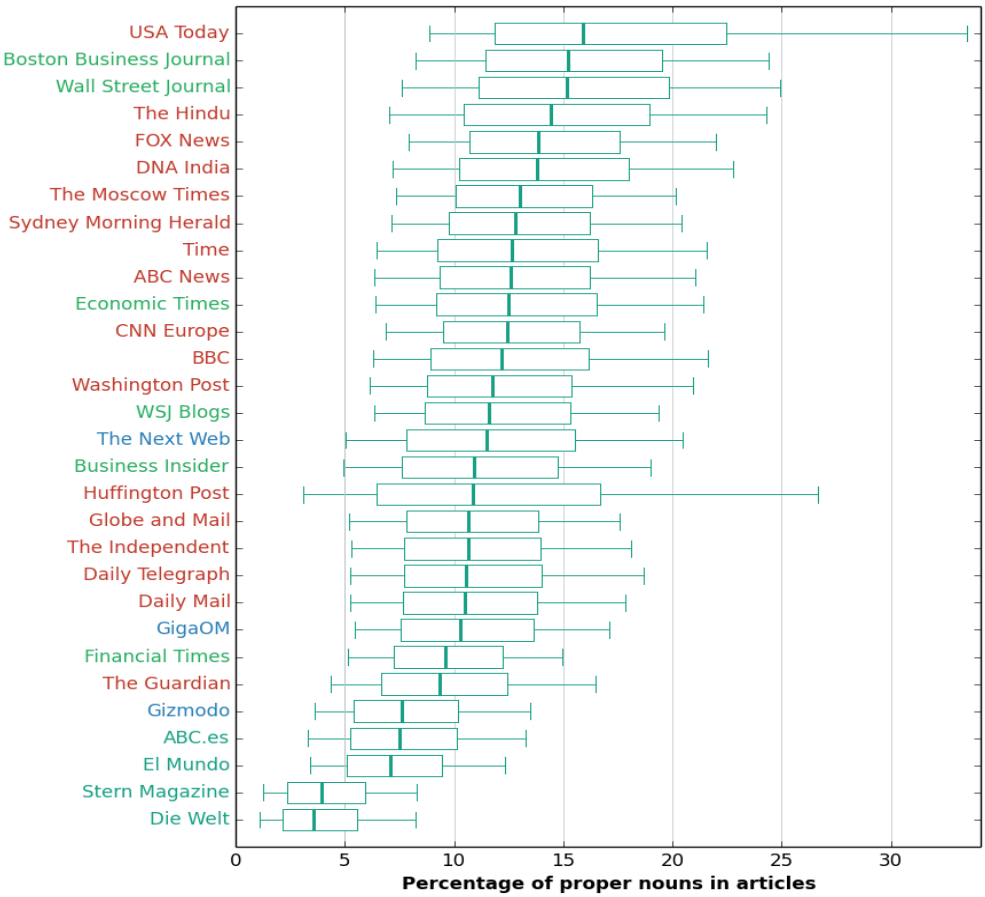

Figure 7 Percentage of proper nouns in articles per publisher

As evident from the following Figure 8, percentage of verbs used by the news sources in their news articles remains relatively high, especially in the three British news sources BBC, Daily Mail and The Independent. Similar to the previous results, the two German news outlets die Welt and Stern Magazine refer to verbs less when reporting news. Young journalists especially at the beginning of their careers are often corrected and encouraged by the editors to use more verbs that make their news writing powerful and simply enhance their style and their professional journalistic writings. Verbs evoke visual images in the public's mind and are great examples for writing excellent and effective description. News stories, especially online news stories, often do not have enough space for long, descriptive passages, but with just a few key verbs they can convey the main meaning and make the readers aware of the problem described in an article. Verbs often provide a story with a sense of movement and action.

An interesting observation to mention is, for example, that BBC uses lower number of verbs when writing articles about Arts (15,6%) and Sports events (15,4%), whereas the highest usage of verbs belonged to the categories Health (18%), Home (17,6%) and Society (17,5%). More or less the same observations were noticed for The Independent: highest usage of verbs for the categories Health (17%), Society (16,6%) and Computers (16,2%) and lowest usage of verbs when writing about Arts (15,1%), Shopping (15,5%) and Sports (15,6%).

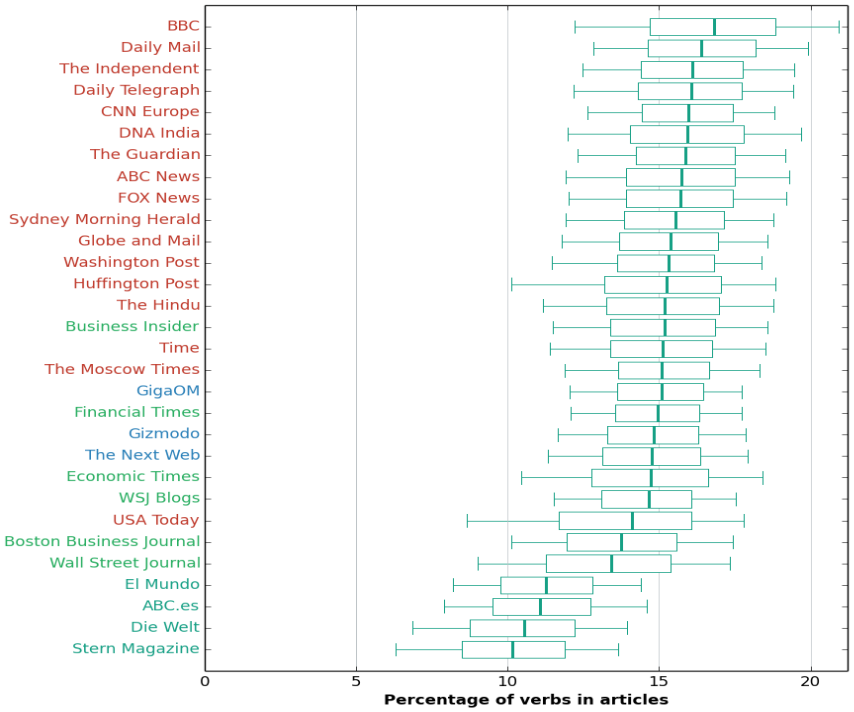

Figure 8 Percentage of verbs in articles

Analysis of readability differences

In this experiment we wanted to evaluate how the publishers differ in regards to the readability of their articles. Readability is an indicator of how understandable a text is to a particular group of readers [KS11]. Such measures have been used extensively to help to evaluate and develop textbooks, business publications, medical literature, etc. 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. For our purposes we have chosen to use two of the oldest but at the same time the most accurate readability measures:

Dale-Chall readability formula.

The measure is computed as follows:

$$0{,}1579\left(\frac{difficult\ words}{words}\times 100\right)+0{,}0496\left(\frac{words}{sentences}\right)$$

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 [CJ95]. Since the list of non-difficult words is available only for English language we had to come up with a similar list also for German and Spanish in order to evaluate articles in those languages. We computed these words by analysing 10,000 German and Spanish articles and identifying the most common 3,000 words. This set of words will not provide us with a precise score but it is the best approximation we were able to make.

The computed score is normalized so that it can be used to estimate the grade level needed to understand the text. The mapping between the Dale-Chall score and the grade level can be seen in table below.

Dale-Chall score	Grade level
4.9 and below	Grade 4 and below
5.0 to 5.9	Grades 5-6
6.0 to 6.9	Grades 7-8
7.0 to 7.9	Grades 9-10
8.0 to 8.9	Grades 11-12
9.0 to 9.9	Grades 13-15 (College)
10 and above	Grades 16 and above (College graduate)

Table 2 Dale-Chall scores and the associated grade levels

Flesch-Kincaid Reading Ease

This measure is a standard way for many US agencies for evaluating technical documents. The measure is computed as follows [KJ75]:

$$206{,}835 - 1{,}015\left(\frac{total\ words}{total\ sentences}\right) - 84{,}6\left(\frac{total\ syllables}{total\ words}\right)$$

In contrast with the Dale-Chall score, the Flesch-Kincaid Reading ease score is higher for text that is easier to understand.

Looking at the results in the Figure 9, we see that certain news websites like ABC.es (9,9 % on average), El Mundo (9,8 % on average), Stern Magazine (9,9 %) and die Welt (9,7 %) have high readability scores, meaning that word complexity according to the Dale and Chall formula is higher. This can be the result of us using a list of 3,000 common words that are not comparable to the list of words manually chosen for English language. Alternatively, it could also be due to the fact that the German and Spanish language have a higher number of compound words, which makes the quotient for “difficult words” larger. Among the English articles, the most difficult to understand are articles from the Indian news publishers and three of the financial news publishers. It is interesting to note that within each publisher there is quite large variability in the readability of the articles. It seems, for example that more than 10% of the articles from the Wall Street Journal have a score below 8 (grades 9-10), whereas about 20% of them have a score above 10 (college graduate).

The results of the Flesch-Kincaid Reading Ease experiment are shown in the Figure 10. The text with a score between 90 and 100 is considered very easy, 80-89 easy, 70-79 fairly easy, 60-69 standard, 50-59 fairly difficult, 30-49 difficult and 0-29 very difficult. Gizmodo (score 72.2 on average), Daily Mail (70.9) and BBC (68.2) scored high on the ease of readability, suggesting that their news articles are easy to read. The Spanish and German news publishers, however, again achieved scores that make them very difficult to understand. Their average scores were: 12.3 for ABC.es, 13.3 for El Mundo, 26.7 for Die Welt and 31.6 for the Stern Magazine.

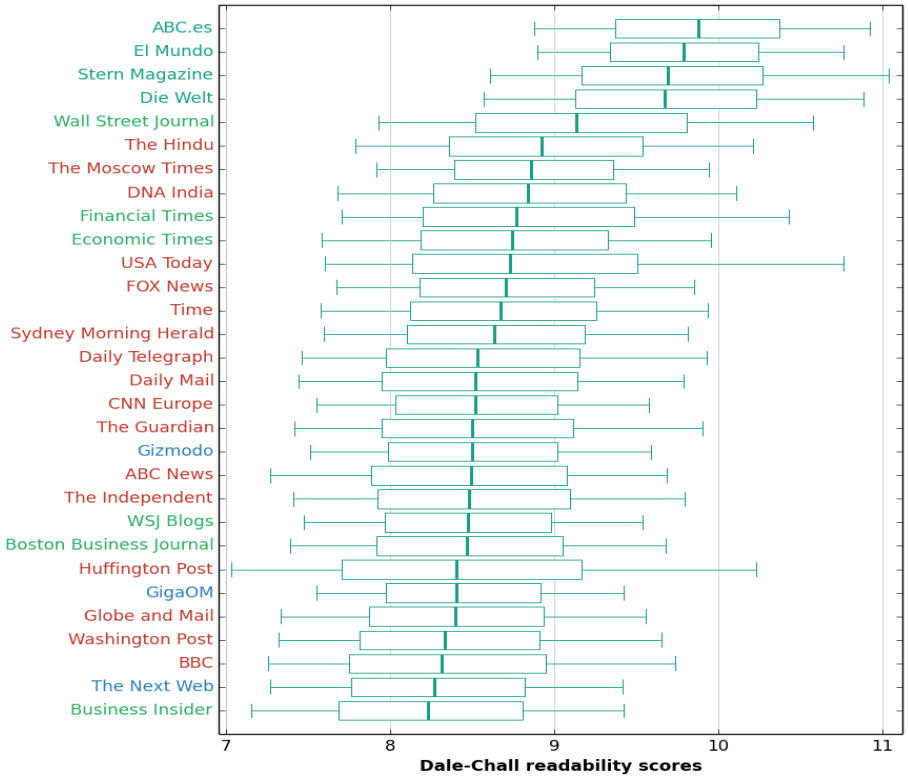

Figure 9 Dale-Chall readability scores

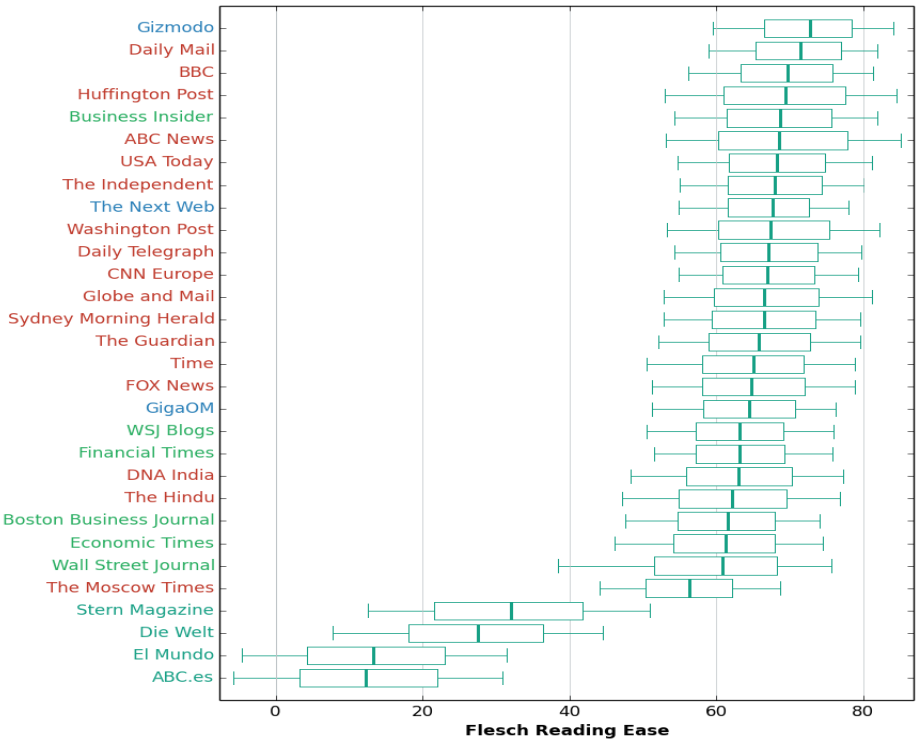

Figure 10 Results for the Flesch-Kincaid Reading Ease experiment

Bias in newswire citations

In this part of our experiment, we focus on how frequently different news agencies are cited by our 30 selected news publishers. News agencies (also called press agencies) have correspondents and journalists in most of the countries. They however do not publish news to the public, but rather serve as suppliers to various 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 obliged to cite where the information is coming from by referring to the press agency and not the actual author itself.

In order to determine if an agency is cited in an article we first had to identify a set of most popular news agencies. We obtained the list of 30 news agencies from the corresponding Wikipedia page[8]. For each agency we have created a set of display forms in which the agencies can be mentioned. 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.

In this experiment we were 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.

The results of the first experiment are summarized in Figure 11 and show that the USA Today (32%), The Moscow Times (26%) and Fox News (26%) are the news agencies 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. It is interesting to see that both Spanish news sources are among the lowest ones in the ranking. This could be due to an editorial decision not to mention the news agencies (an example of a poor journalistic practice) or a consequence of writing extensively about topics that are not reported by the agencies.

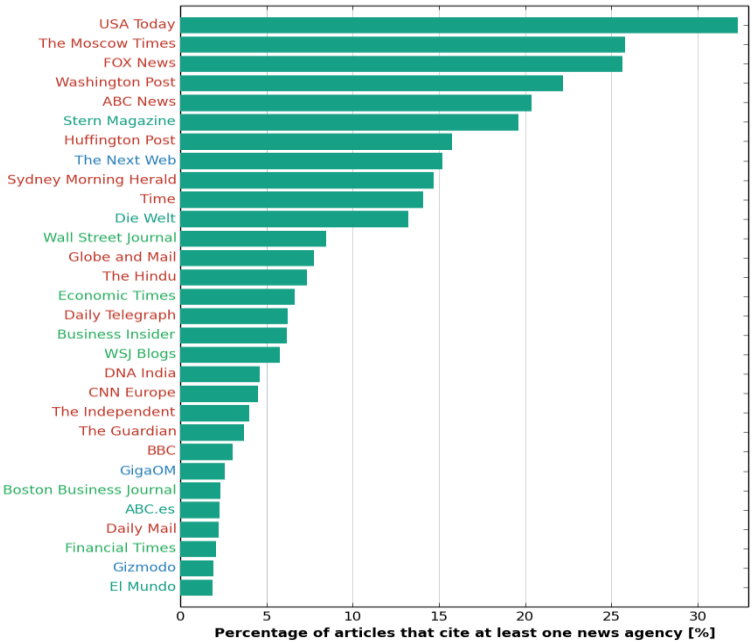

Figure 11 Percentage of articles per publisher that cite at least one news agency

Our experiment (see Figure 12) also 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%

[8] http://en.wikipedia.org/wiki/News_agency

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%).

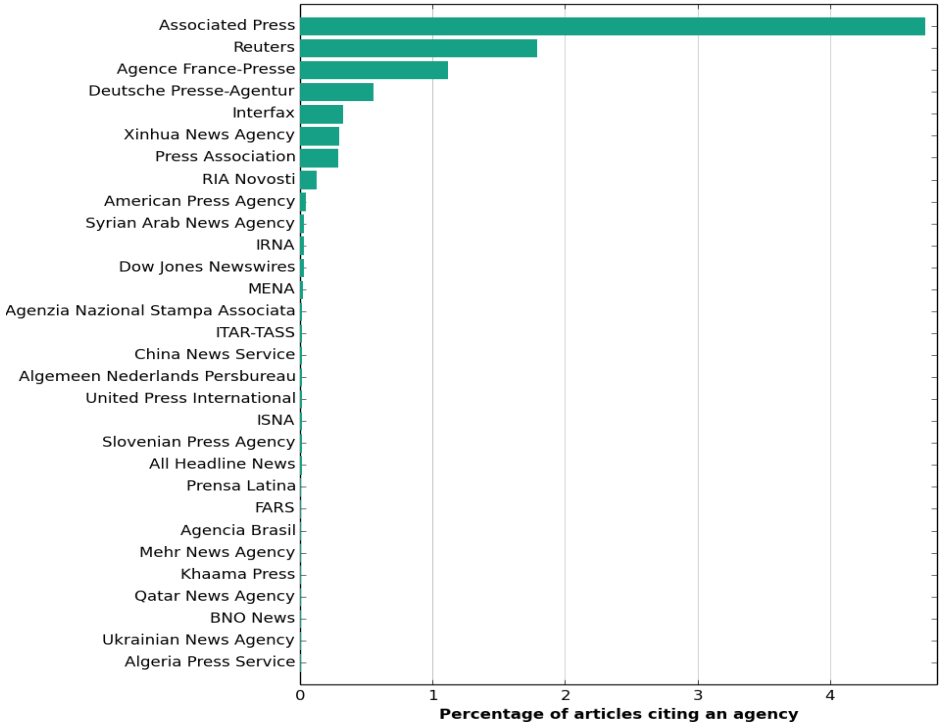

Figure 12 Percentage of articles citing a news agency

In the next figure (Figure 13) we show the bias that the news publishers have when citing the news sources. The figure is a sieve diagram (FM00) displaying a dependency between the two variables. The diagram is built by showing splitting the data (the whole square) according to the distributions of the two variables. The split is done under the assumption that the variables are independent. The width of the AP rectangles, for example amount to 55% of total width of rectangles because 55% of all citations by the displayed publishers cite AP. Similarly, the height of the Washington Post rectangles amount to 30% of total height since citations by this publisher account to 30% of all citations (among the publishers displayed in the diagram). If the independence assumption would be correct, the area of the rectangle for Washington Post and AP should correspond to the ratio of times Washington Post cited AP, compared to all citations. Since pairs of variables are rarely completely independent we use the color of the rectangles to highlight the degree of dependence.

In order to compute how dependent individual pairs of values of two variables are we use the Standardized Pearson residuals. It can be computed as:

$$p = \frac{O - E}{\sqrt{E}}$$

Where O represents the observed number of cases and E the expected number of cases. Following our previous example, the O would be the actual number of times the Washington Post cited AP. The expected number of cases E would be computed as P(AP) * P(Washington Post) * N, where P(AP) is the ratio of times AP is cited, P(Washington Post) is the ratio of citations by Washington Post and N is the total number of citations.

If computed residual value is close to 0 this indicates small deviation from the independence. If the residual is positive this indicates that there are more cases with this combination of values than expected under the independence assumption. The opposite is true if the number is negative. Positive residuals are usually displayed with different shades of blue (depending on the residual value) and negative residuals with red color.

Based on the diagram in Figure 13 we can conclude that news publishers are very biased when citing news wires. First, the American news publishers (Washington Post, Fox News and Time) show a very strong tendency to cite Associated Press (an American news wire). Time is the only publisher that is citing a lot the British news agency Reuters. Reuters is also cited a lot by other non-US publishers. AFP seems to be commonly cited by the Stern Magazine, BBC and the Sydney Morning Herald. Deutsche Presse-Agentur is not very popular and is cited mostly in the Stern Magazine – the only German publisher in the diagram. As expected, Interfax is cited the most by the Moscow Times, but also by the Stern Magazine. Similarly, the Press Association, which is a UK based news wire, is cited mostly by BBC, The Guardian and the CNN Europe.

The general conclusion can be that news publishers are citing those news agencies that are geographically close. This is not totally unexpected since news generated by news wires are often relevant only for a particular geographic area and not worldwide. Non-US publishers are however more diverse in their set of agencies they cite. The best example of this is the Stern Magazine which seems to uniformly cite in their articles four different news agencies.

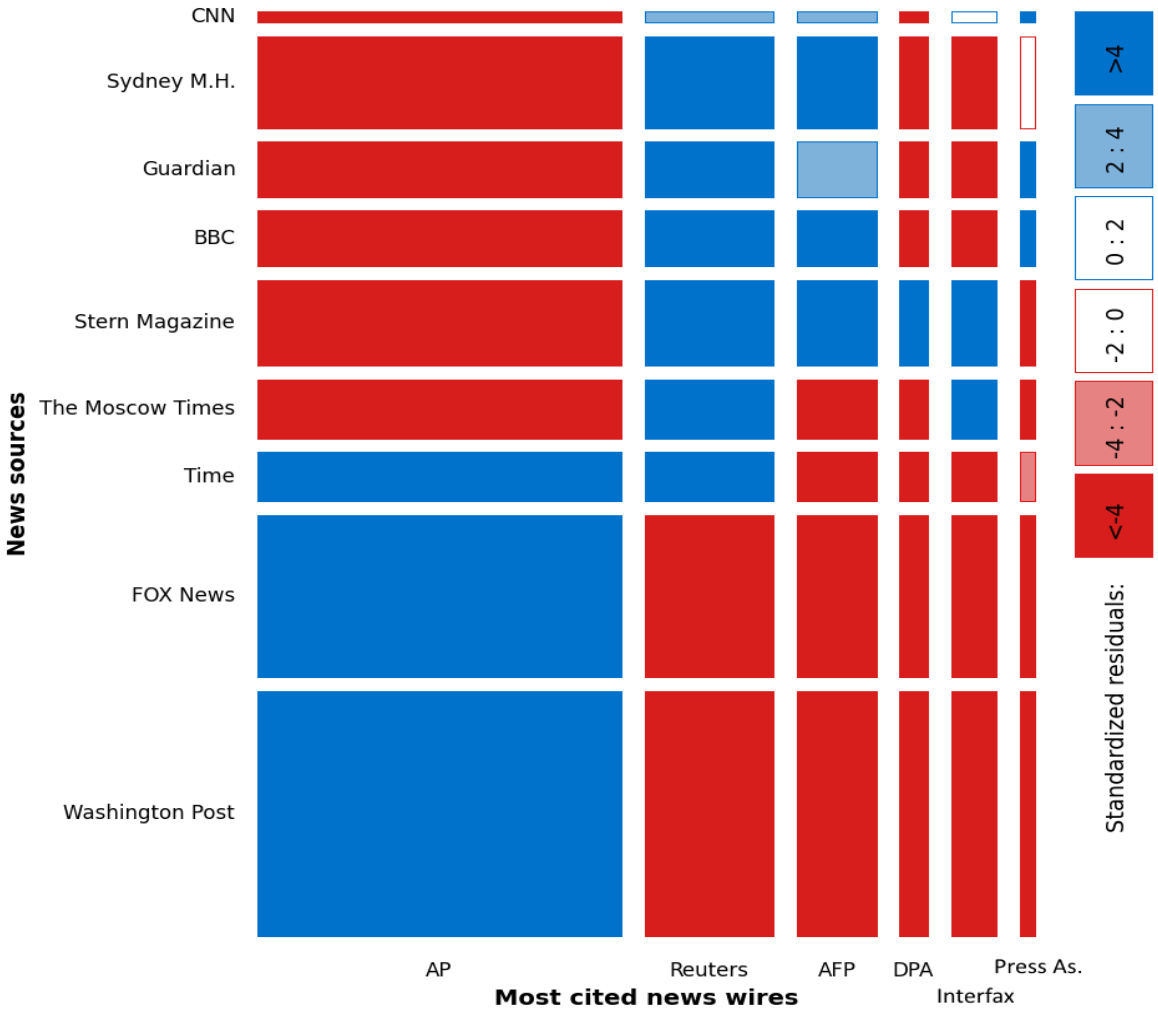

Figure 13 Sieve diagram showing the correlations between news wires and news publishers

Geographical bias

In this experiment we have focused on the geographical analysis of the news articles published by several publishers. The countries on the following figures are coloured with respect to the standardized residual computed for that country and publisher so the colors of countries correspond to colors used in sieve diagrams and heat maps in other sections of the document. We have limited our focus and color range to

residual values in range [-4, 4] and all residual values outside that range are colored as min/max values within that range. White color (or rather light grey) denotes residual value 0 and is also used for countries with less than 500 articles over all publishers. This limit was used to remove spurious residuals that occur with small number of articles.

Identifying event location for a given article is done in the following way. If we are able to identify in the article a dateline (the brief piece of text at the beginning of the article) that mentions a location, then we assume that is the location of the event, described in the article. In large number of articles, the dateline is, however, not present. In those cases we check to which event does the article belong in the Event Registry and use the event's location. Location of the events in the Event Registry is determined by a classification algorithm that takes into account all articles belonging to the event. In some cases, the location of the event cannot be determined reliably. Articles that belong to such events were not used in the analysis.

News is one of the main ways we learn about the world and different geographical places. News outlets 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). Nowadays more and more news outlets are rather subject-matter oriented and cover one or a few particular topics (financial matters, sports, entertainment). The outlets selected for analysis in this deliverable are either national, such as die Welt or the Hindu, or subject-matter, such as Gizmodo and Economic Times.

Not all events that happen in the world get into news. Events have to coincide with certain characteristics in order for news creators and consumers to find them newsworthy. These news criteria or news values have been a matter of detailed study and several complex news value factors have been proposed [GJ65] like Frequency, Threshold, Unambiguity, Meaningfulness etc. Since these factors are quite abstract and hard to analyze automatically we limit ourselves to proximity for our analysis in this section. Proximity stands for the geographical closeness of an event to the readers – stories that happen near to us and near to our homes have more significance and newsworthiness – as opposed to closeness in terms of values, interests and expectations of the news audience (which is of course also very influential but is not a matter of study in this section). For example, 5 women raped and killed in a small Indian town, the event will definitely be considered as news for most of the Indian newspapers and online publishers. If 5 women suffered the same fate in a small town in a Western country, the news would probably pass without significant notice in Indian media.

The overall results of our experiment confirmed very strong geographic preferences of most news publishers, with some exceptions. In the following Figure 15 we see the results of geographical scope and distribution in national American outlet USA Today demonstrating news coming and happening in the USA as almost the only focus of attention. Some attention is given to Ukraine due to the American interest and role in the conflict between Russia and Ukraine and some Northern African countries. In Asia, North Korea stands out most likely due to coverage of the countries state and actions under the Kim regime.

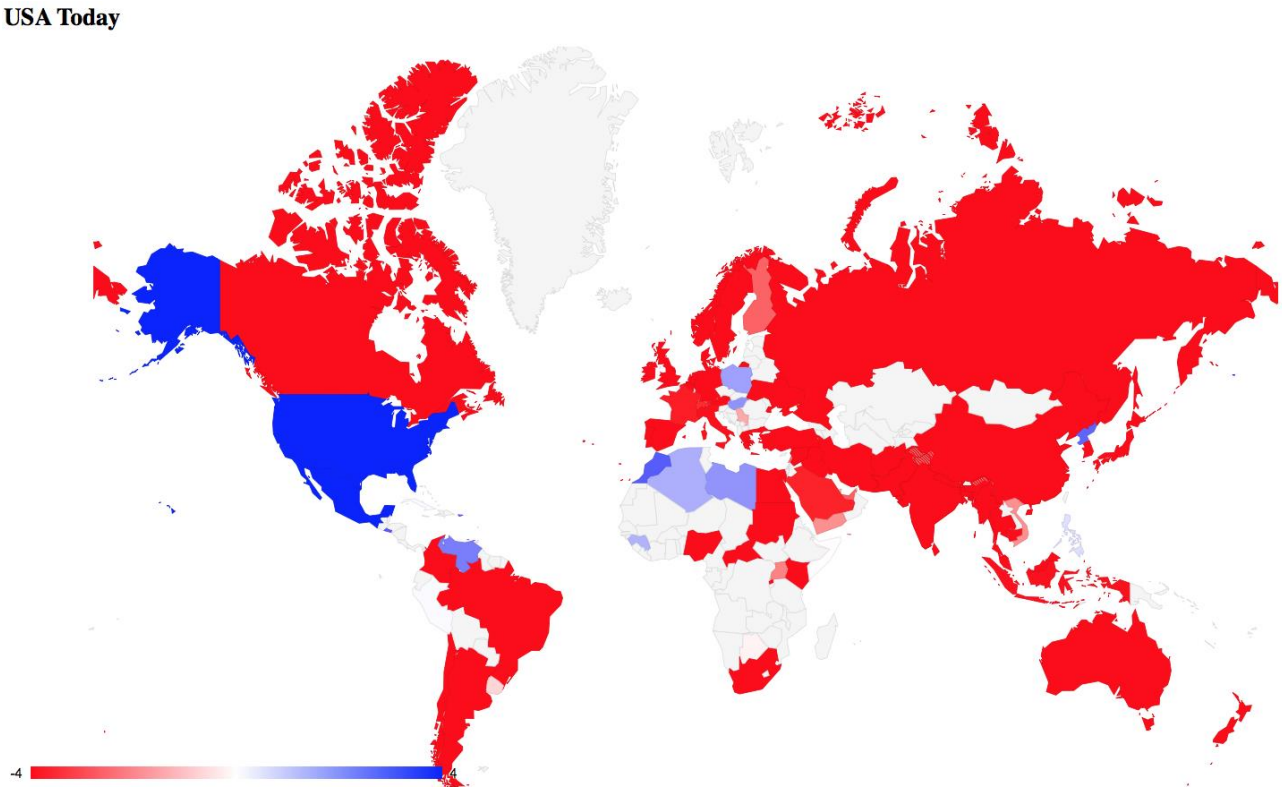

Figure 14 Geographical distribution of news articles in USA Today

In Figure 15 and Figure 16 we see the results for two European publishers: Spanish and German national publishers ABC.es and die Welt. Note that ABC.es, apart from their key focus on Spain, places a lot of attention on neighbouring countries most likely having similar problems (such as Italy, Greece, Malta, Marocco, etc.) and to the Spanish speaking countries in South America – an interesting example of the language proximity.

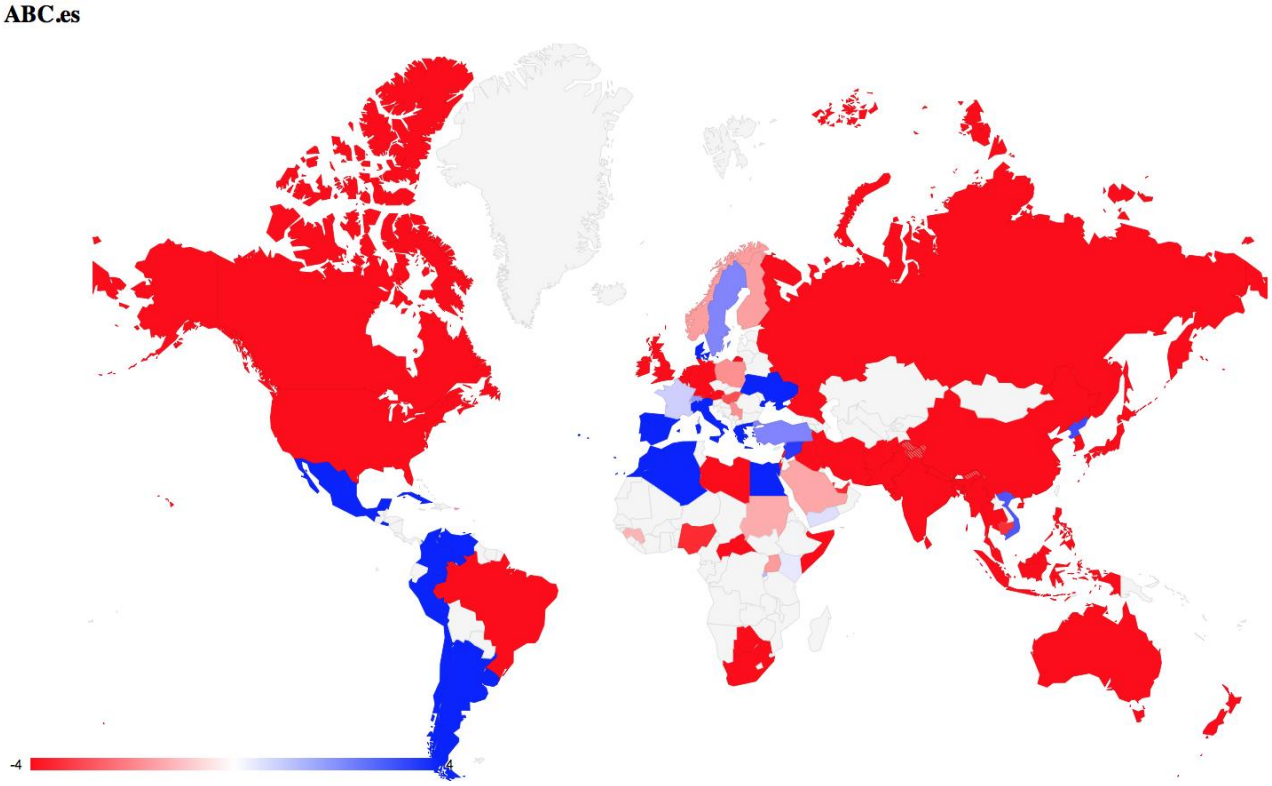

Figure 15 Geographical distribution of news articles in ABC.es

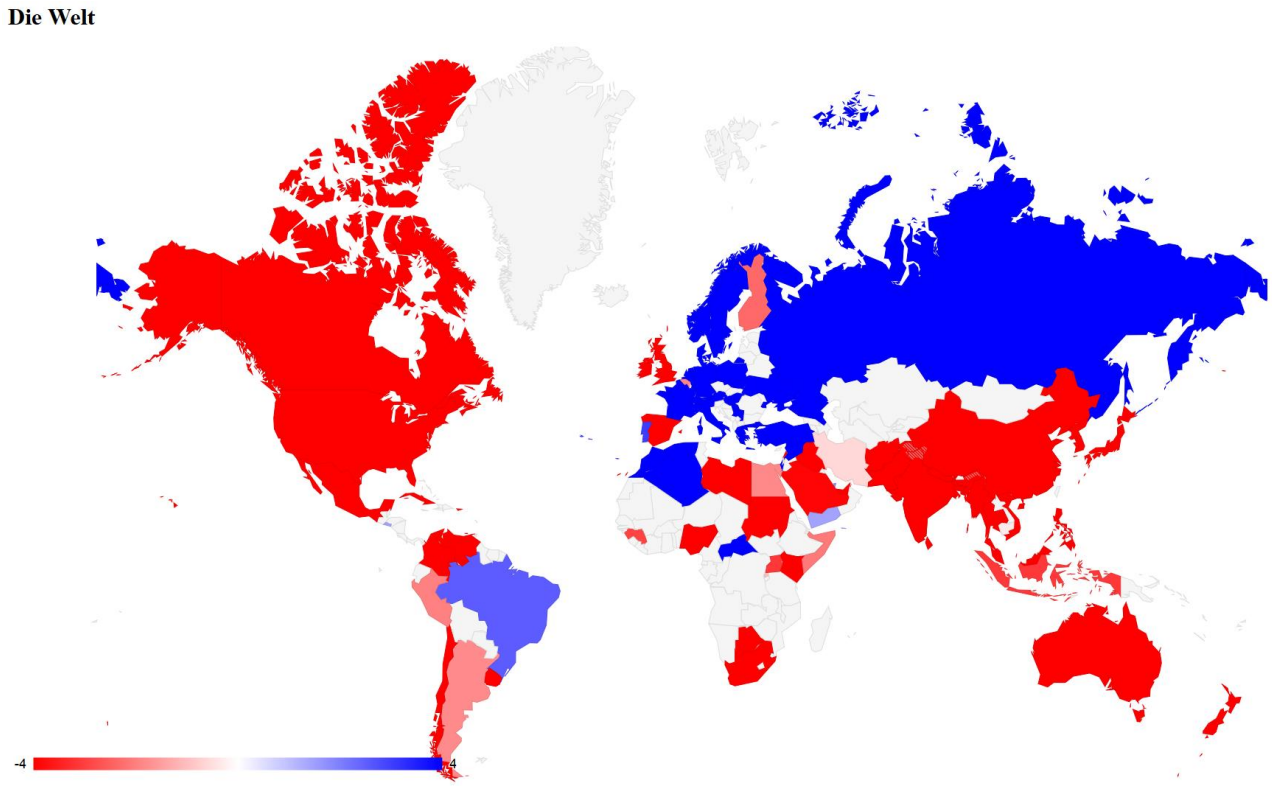

Figure 16 Geographical distribution of news articles in die Welt

The national German publisher die Welt, unlike ABC.es, has a very strong EU preference but surprisingly low interest in Spain, Finland and the UK. Noteworthy is the focus on Russia possibly due to the historical, political and economic ties. Larger countries like Russia especially tend to receive bigger amount of attention when they are part of a conflict such as the recent Crimean crisis.

The next two Figures, Figure 17 and Figure 18, show results for two subject matter oriented news outlets. First is The Economic Times which is an Indian English speaking publisher with main focus on India, and neighbouring countries China and Japan. The second is Financial Times, a UK based international English language publisher. The Economic Times is the second most widely read English business publishers after World Street Journal[9] but with such a narrow geographical focus, most likely because it is published from 12 big cities in India and their main content is on the Indian economy.

[9] http://en.wikipedia.org/wiki/The_Economic_Times

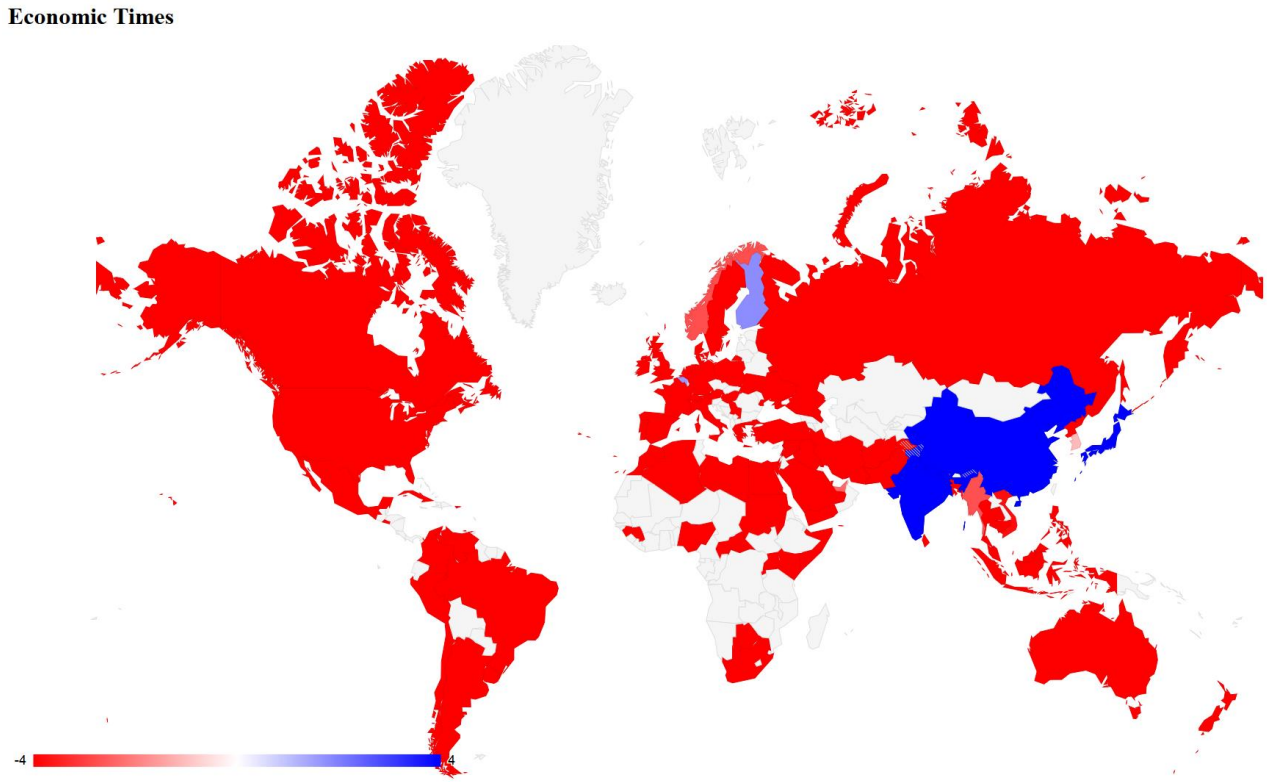

Figure 17 Geographical distribution of news articles in Economic Times

It is interesting to note that geographical scope in Financial Times is so much wider and more scattered throughout the world in comparison to Economic Times or many other publishers under our analysis. Financial Times, traditionally with readers only in London, became an international business outlet with a specific focus on business and economics all around the world, as seen from the Figure 18. Their focus does not fall on the USA or Latin America since they do not need to compete with an American business outlet the Wall Street Journal, which already has a great coverage there due to their geographical proximity. Also amongst the more neglected (very little or no content) are many of the African and Asian countries probably due to a different world financial agenda or simple lack of access to the information and low interest from the public for some countries.

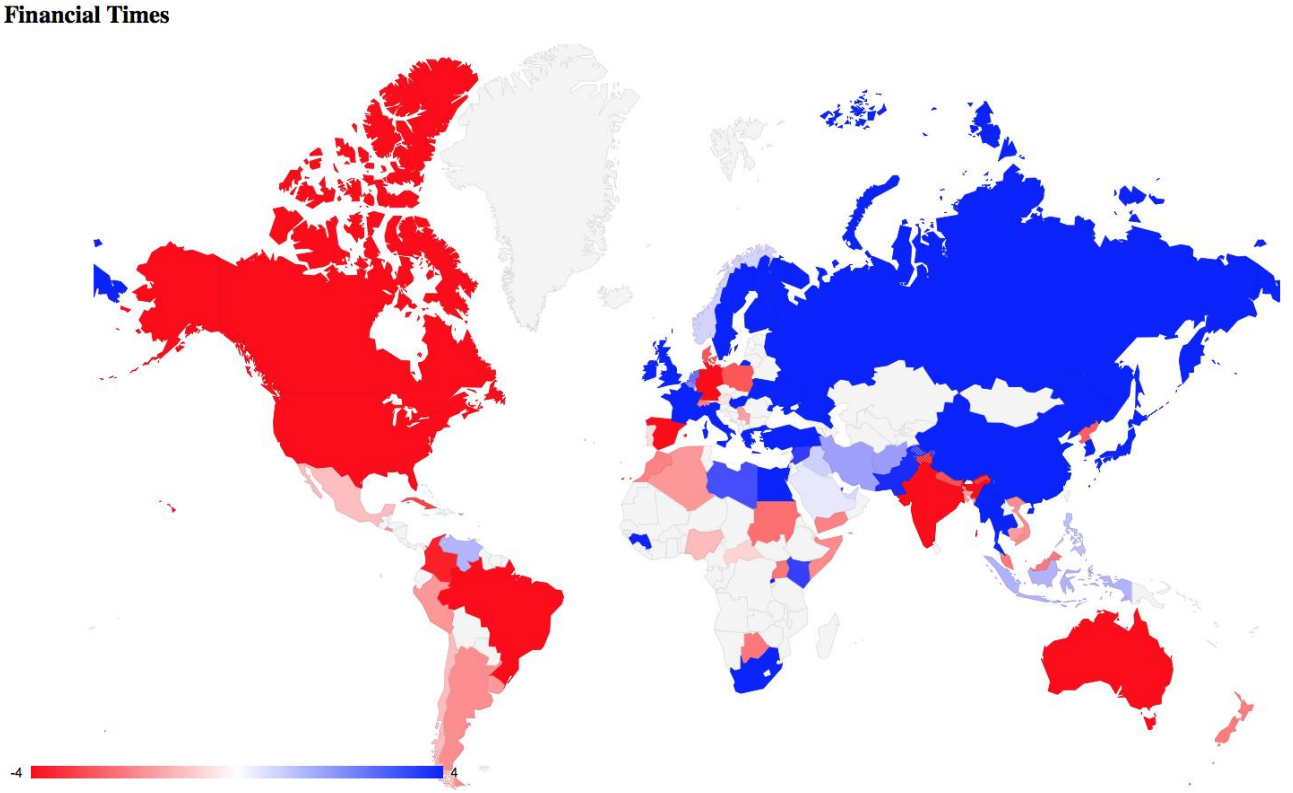

Figure 18 Geographical distribution of news articles in Financial Times

Topic (category) coverage bias

In this part of the experiment we had a closer look at what topics are given more preferences and more coverage in the selected sources. The topics used in our analysis correspond to DMoz categories described in section 0. Besides the distribution over top level categories for all publishers we have also analysed bias inside sub-categories for some of the publishers. Due to a very large number of sub-categories in the DMoz taxonomy we have limited ourselves to second level sub-categories only and removed from the charts all categories that have less than 500 articles to avoid spurious residuals.

Charts in this section are heat maps[10] of residuals. They are essentially the same visualization as sieve diagrams (such as the one in the Figure 13) but here the area of the box conveys no information and is the same for all variable pairs. The color blue represents a stronger tendency to cover a topic, whereas the colour red stands for less coverage.

Figure 19 shows coverage bias bias of top level categories for all publishers. As expected, we 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, give 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, pay very little attention to Sports, Health and Home issues which lie outside of their main scope. Overall, Figure 19 only confirms 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.

[10] https://en.wikipedia.org/wiki/Heat_map

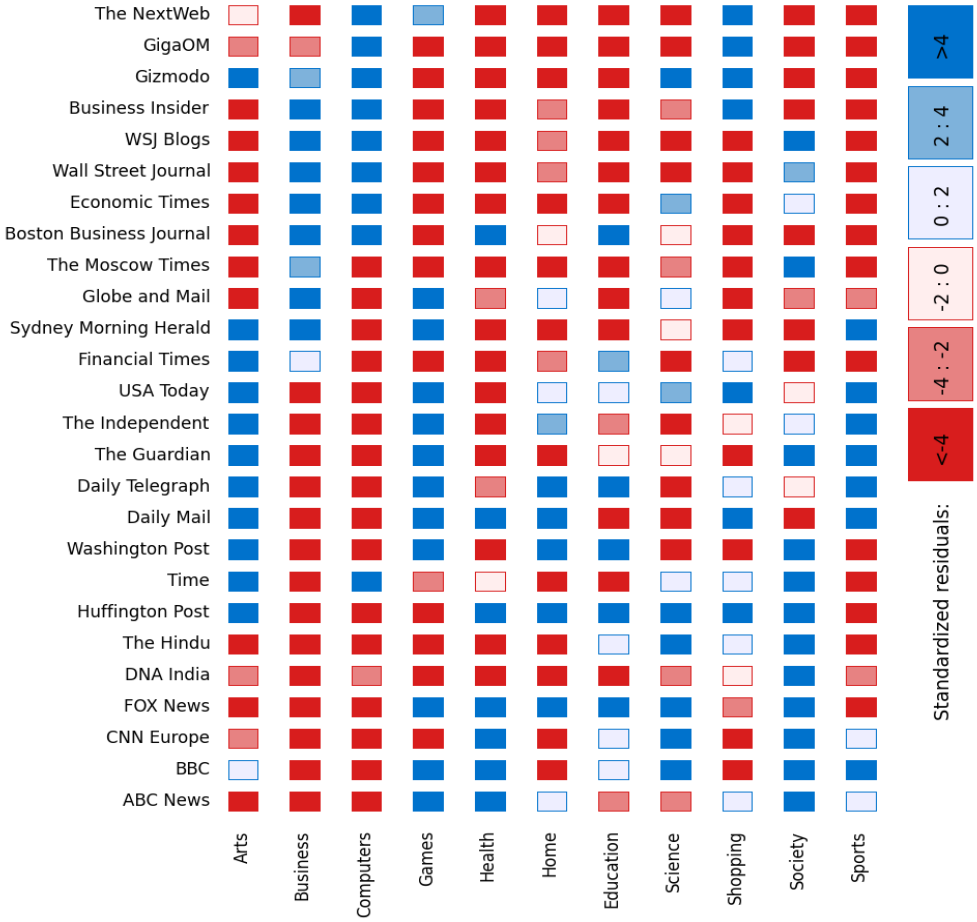

Figure 19 Coverage bias of top level categories for all publishers

In the Figure 20 we take a closer look at what sub-topics of the general topic Society the selected publishers are particularly interested in. Interesting to point out is that, for example, the Moscow Times reports a lot about governments, history, relationships and ethnicity – a traditionally big topic in Russia with focus immigration from former Soviet Republics on the rise. As one of the few newspapers reporting from Russia in English that is owned by a big Finnish Corporation and is often very critical towards the Russian government, The Moscow Times gives some attention to the topic of gay and lesbians, which usually has no coverage whatsoever in the Russian language news providers. Big American publishers like Fox News, Time and European edition of CNN, give their preferences in covering issues connected with military – a topic of concern to most Americans. The Indian news outlet the Hindu shows a biased tendency towards topic like ethnicity and religion and spirituality. Overall, this figure demonstrated and confirmed the key and often-traditional topics of concern to a particular country or continent.

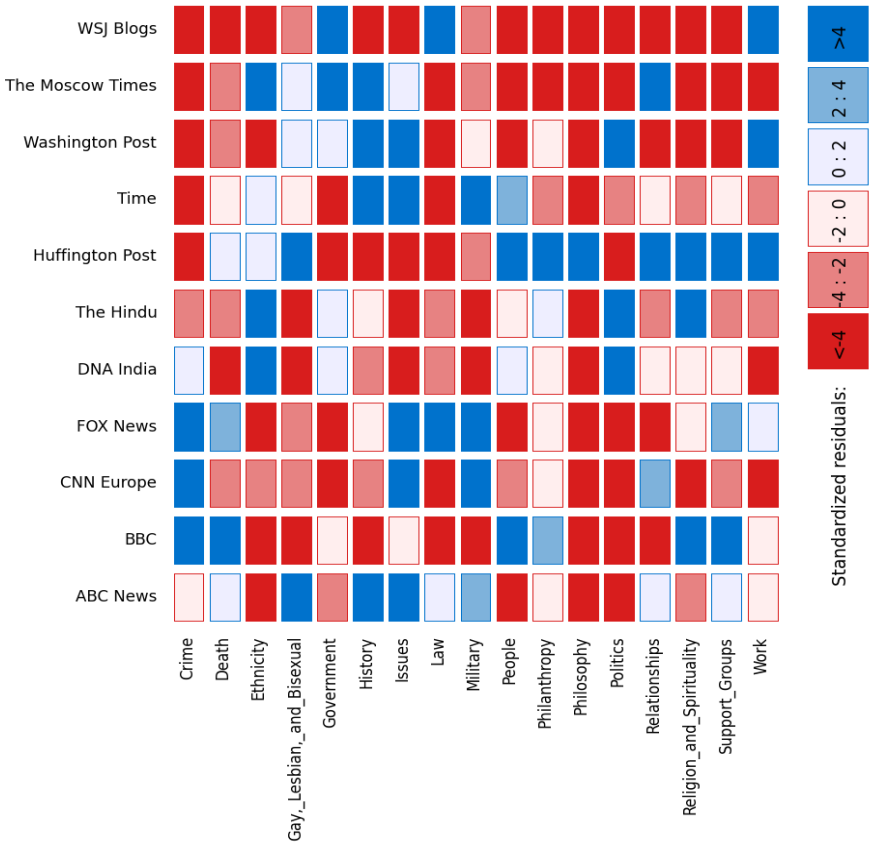

Figure 20 Coverage bias for Society sub-categories

In Figure 21 we look at sub topics of the category Computers in the five selected websites, including the technology related outlets we included in our analysis. We see that unsurprisingly The Next Web focuses on internet and software, whereas GigaOM and Gizmodo cover more software and hardware related areas of consumer technologies respectively. Business Insider and WSJ Blogs are not primarily technology outlets and their focus is broader and less specialized.

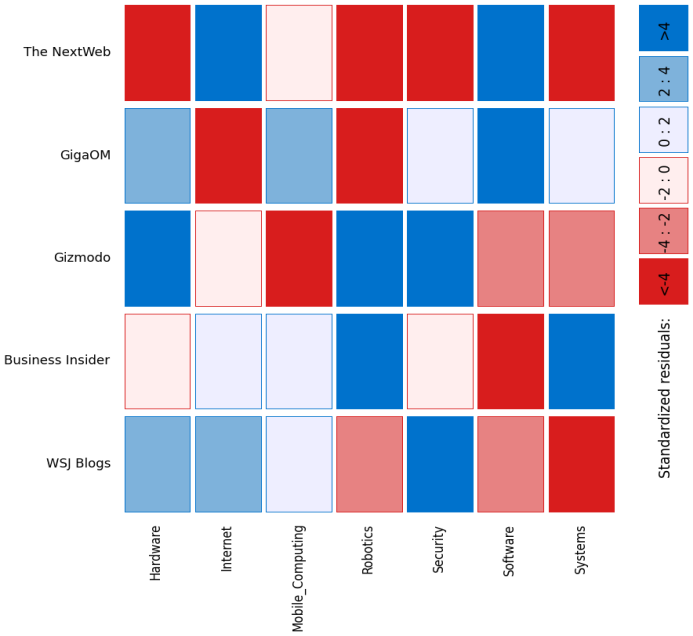

Figure 21 Coverage bias for Computers sub-categories

Figure 22 shows bias within the Health category. Boston Business Journal as the only primarily financial outlet stands out by putting strong emphasis on topics that deal with either education in medicine or healthcare for senior citizens (Senior_Health, Nursing, Pharmacy, Home_Health) perhaps due to the fact that these topics are commonly a matter of funding issues and debate. We can also see some tendency towards more "advice column" type topics (Addictions, Conditions_and_Diseases, Reproductive_Health, Weight_Loss) in more tabloid news outlets like Daily Mail and Huffington Post and a strong bias against alternative medicine in more serious outlets like Time and CNN.

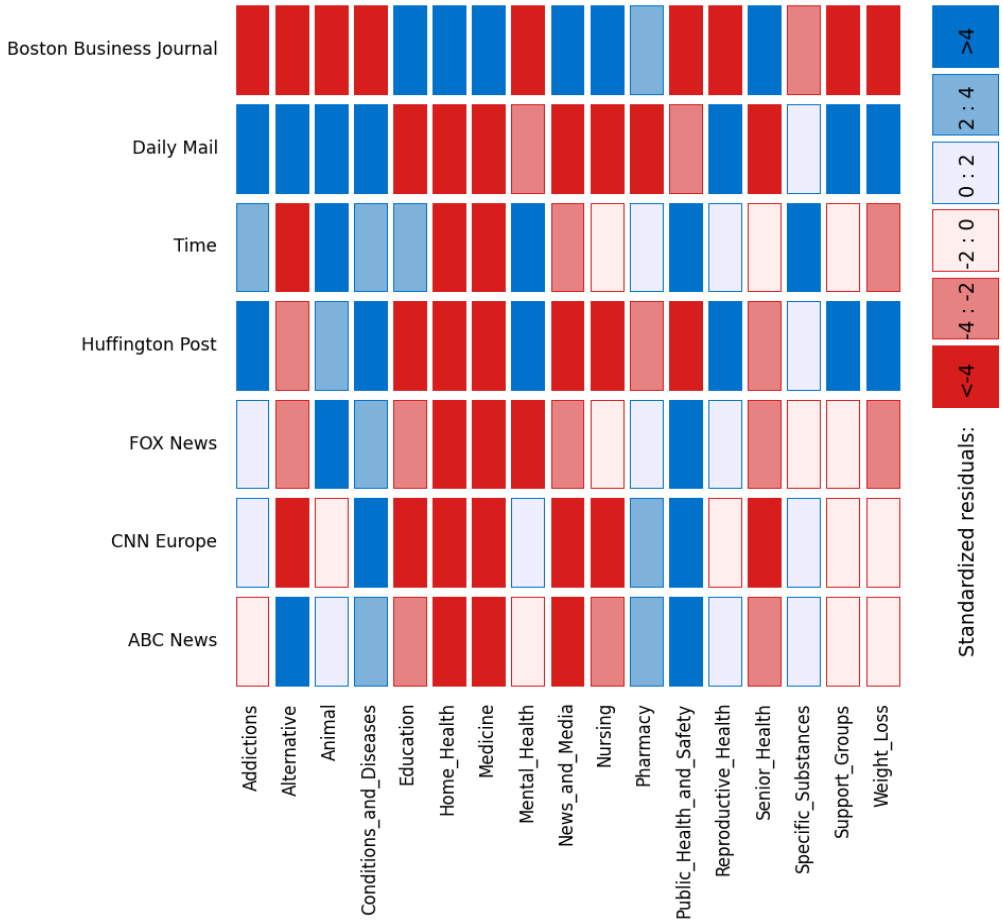

Figure 22 Coverage bias for Health sub-categories

Finally, Figure 23 shows bias for sports sub-topics. Here the publishers are biased towards coverage of sports that are typical/traditional for the region where the publisher is from. USA Today thus has strong bias towards typically America sports like baseball, basketball and bowling. Cricket and soccer are primarily covered in publishers from the UK (Daily Mail, Daily Telegraph, The Guardian, The Independent) or those from countries with strong historic British influence (Sydney Morning Herald). Note that for The Guardian and Sydney Morning Herald the biased sub-category is Football and not Soccer. This is most likely an artefact of the category classifier as the referenced sport is almost certainly the same (as opposed to American football). As the nomenclature in the case of football/soccer is often unclear (and a matter of quite subjective international debate) it is not particularly surprising that the classifier might have trouble in this case.

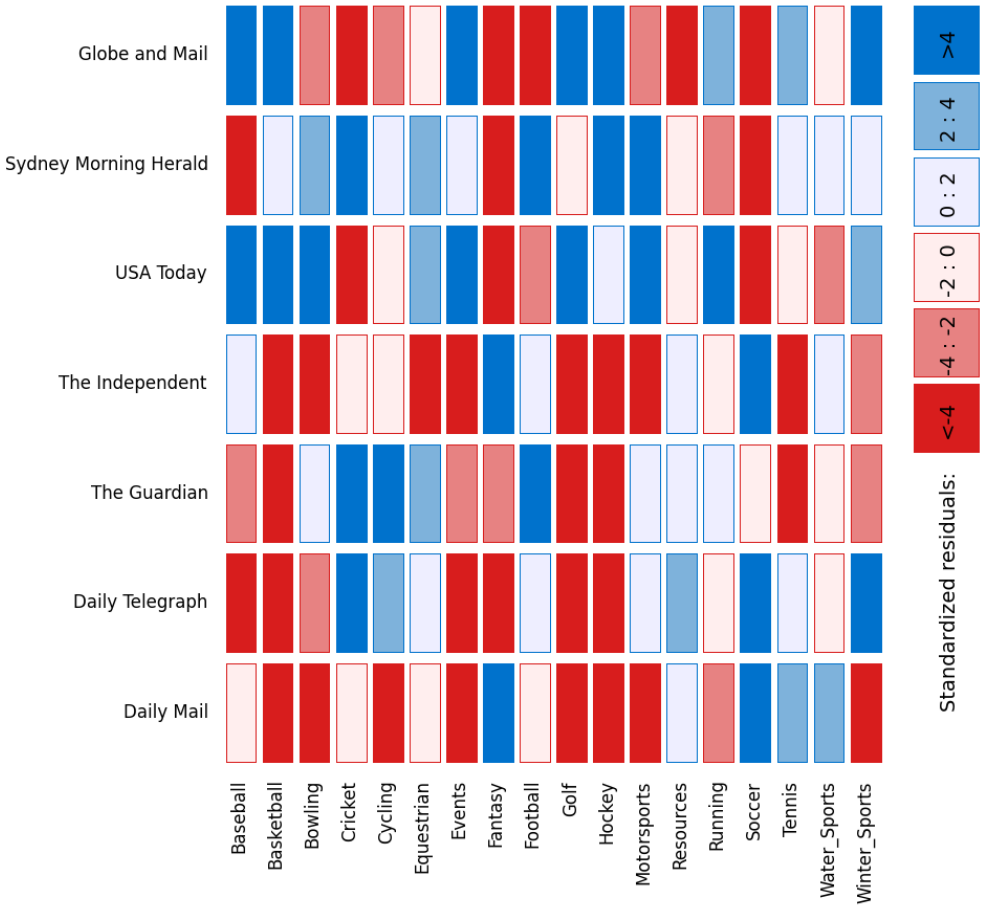

Figure 23 Coverage bias for Sports sub-categories

Bias in the speed of reporting

This step of the experiment was dedicated to detecting news bias in the speed of the news reporting among publishers. Figure 24 shows how much time (in hours) passed between the first article about an event and the first article of this publisher. It measures roughly how quickly a publisher produces an article about an event in comparison to other publishers. Important to note that we have taken into account articles with similarity measure to the event cluster higher or equal to 0.6, meaning that less reliable articles or, more accurately, those articles whose clusters may not have been determined correctly are removed from the analysis.

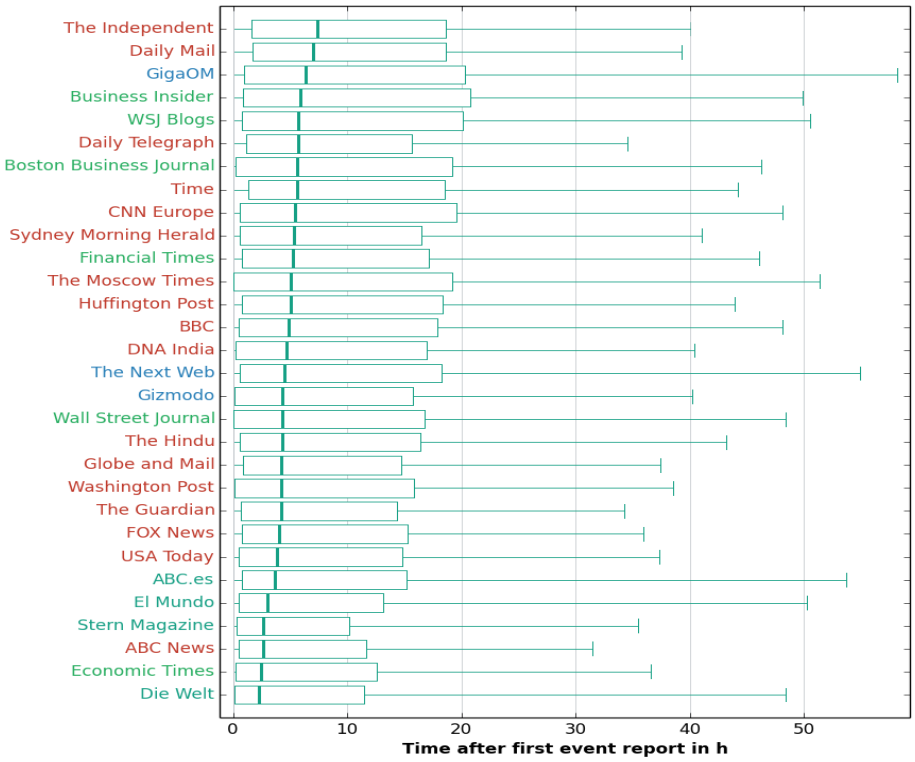

Figure 24 Time after the first event report in hours

As seen from the figure, the highest speed in reporting among the English news outlets belongs to the Australian broadcaster ABC News, followed by USA Today, The Guardian, FOX News, Washington Post and Daily Telegraph. Interesting to note here is that a quality broadcaster BBC is slower in publishing articles as to compared with other two British publishers the Guardian and the Independent. A possible explanation to this fact is that quality newspapers have a tendency in being fast when publishing a short breaking news, but take longer time when producing an international story containing an expert's opinion and some background information. Otherwise the chart does not show any other obvious cases of bias in speed of reporting as median values for most of the publishers are within a two hour difference.
It is also necessary to add that, as visually noticeable on the graph, the non-English publishers like El Mundo or die Welt are biased towards better scores (faster in producing stories) because our system has fewer non-English articles in the corpus and the inter-lingual comparison function is right now imperfect. Non-English clusters (events) are typically smaller and articles within them have do not have many competitors to compete with.

Predicting article news source

In the following set of experiments we wanted to test how specific a news publisher is in their reporting about news. In other words, is it possible to determine who the author of an article is based on the article text alone?

This experiment can be performed in several ways. We have chosen the following tests:

- **Overall test** in which we compared a set of English publishers on their articles on any topic
- **Ukraine-Russia test** in which we compared a subset of news publishers on their articles related to the incident between Ukraine and Russia
- **Objective vs. biased test** in which we compared how well we can distinguish between an "objective" and "biased" news source

For each of the tests we performed a pair-wise test in which we compared for each pair of publishers how well can we separate between articles from the two news sources. For each pair we performed the following series of steps:

1. From each of the two publishers we randomly select 10.000 news articles. The reason we use only a subset of available articles by a publisher is that we want to have a fair comparison between all pairs of publishers. Since publishers vary significantly in the number of written articles, not using equally-sized groups of articles would make some pairs of publishers more easily separable than others. Separating Boston Business Journal (183,361 articles) and Financial Times (6,190 articles) can be, for example, 97% accurate by simply classifying all articles as being written by Boston Business Journal. By using 10.000 news articles from each publisher, the learning problem is equally hard for all pairs of publishers and the baseline accuracy is always 50%.
2. For each article we extract learning features that can be then used by a learning algorithm to distinguish between the publishers. There are various features that can be extracted from the articles. In our experiments we were interested in features related to the article content. To compute the features we first extract article tokens and remove the stop-words. Additionally we also identify and remove words that we found to be particular (and therefore discriminative) to one publisher. The words we remove contain the names of all tested publishers as well as words such as mailonline (appearing in Globe and Mail), huffpost (Huffington Post), and abcfullscreen (ABC News). The remaining tokens are then used as learning features. Additionally we also perform experiments where only tokens of a particular part-of-speech were used as features.
3. The learning examples (20.000) are then randomly split into two groups. The learning group contains 70% of examples and is used to build a classification model. We first use a chi-squared test to rank the relevance of individual features and select a subset of 1.000 most discriminative features. A linear SVM model is then trained on the learning examples as represented using the selected subset of features.
4. The 30% of examples that were not used for learning are then used to test the accuracy of the trained model. The measure used for evaluating the accuracy of the classifier is classification accuracy.

Once we evaluate all pairs of news sources we essentially have a matrix with the accuracies. The tables for different experiments are displayed in the appendix in Section 0. Since they contain many values we wanted to generate an informative visual display of the information. For this purpose we decided to use Multidimensional scaling (MDS) [BI05]. MDS is an optimization algorithm that takes as input a distance matrix between objects and produces as output a visualization in low dimensional space. The locations of the objects in the visualization (in our case 2D) are determined so that the original distances between objects are best preserved. The objects that are placed in the visualization closer together are therefore more similar than the objects that are further apart. It is important to note that since MDS computes an embedding of higher-dimensional data into a 2-D space, the visualization is only an approximation – the computed distances between objects do not exactly match the values in the original similarity matrix.

In order to use MDS we first needed to generate a distance matrix of news publishers. In the plot we wanted to see close together those news source that are similar to each other and therefore hard to discriminate. The news sources that can be well discriminated, on the other hand, should be plotted further apart. Obtaining a distance matrix D from our matrix of accuracies A can therefore simply be done as $D = 1 - A$.

Overall test

For the first test we selected the set of news publishers that write in English language and produce general articles (we ignored the financial and technology news sources).

The visualization obtained using MDS is in Figure 25. The two news sources that we found to be the most similar and therefore the most difficult to discriminate are The Independent and the Daily Telegraph. Both are UK based news sources that are producing content about similar topics and using similar vocabulary. The accuracy of the model predicting the news source was 66.9%. Additional sources that are also similar to the

mentioned ones are The Guardian (76.1% accuracy), Daily Mail (77.7%) and BBC (84.3%). Another group of publishers that are relatively hard to distinguish include Time, ABC News, Huffington Post and Washington post. The accuracies in this group ranges between 78.1% and 84.2%. Two publishers that are also relatively similar are The Hindu and DNA India – their separation accuracy is 82.8%. Many of the other news publishers are, however, very unique and easy to discriminate. Sydney Morning Herald, for example, has accuracies between and 95.7% and 98.7%. Similarly easy to separate are also The Moscow Times, The Hindu and DNA India. A general (and expected) conclusion that we can also make is that European publishers are more similar to other European publishers, and US publishers are more similar to other US publishers.

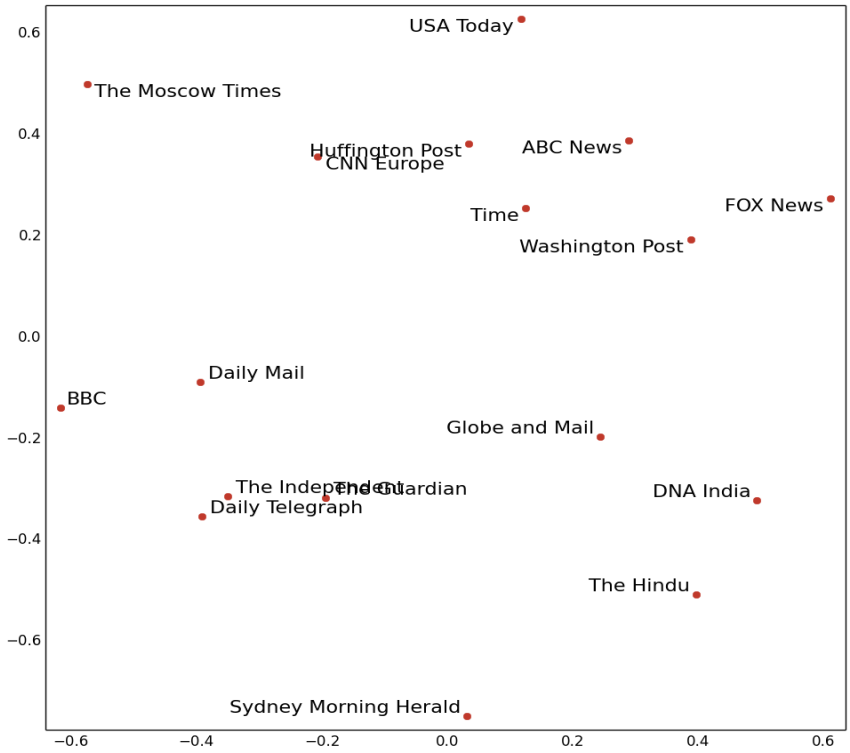

Figure 25 Separability between news publishers based on article text

Since we have use a linear SVM as the learning algorithm we are also able to extract features that are most relevant when separating between two publishers. In the Table 3 we present some example keywords that were found to be relevant for separating each news source from the others. As we can see, many of the discriminative words are proper nouns that are related to the geographic location of the news publisher. When discriminating between a US and a European news source, the words that get a high score are those that have a different spelling, such as color/colour, center/centre, favorite/favourite.

We were also interested in how are words of particular type important for discrimination between news sources. For this purpose we repeated the same test multiple times, each time using only words of particular class. In all experiments, the achieved accuracy was lower, but the changes varied significantly depending on the word class used. By using adjectives only, the classification accuracy was on average lower for 12.0%, using adverbs for 18.1%, nouns 0.7%, proper nouns 2.5% and using verbs alone 8.8%. As we can see, the nouns seem to be the most informative in discriminating between the sources - using nouns alone there is almost no change in the model's ability to separate between the sources. The least discriminable are adjectives and adverbs which could be due to their low frequency and uniqueness.

News source	Keywords
Time	selfie, uber, rapper, minister, researchers, Hillary, sex, party, relationships
Washingon post	Maryland, redskins, Georgetown, Montgomery, reelection, capital, Arlington, officials, capitol
FOX News	Jerusalem, Christians, minister, dynasty, officials, prosecutors
Daily mail	Mr, uk, centre, behavior, football, Scotland, British, premier, Cameron
BBC	Correspondents, Belfast, council, Scotland, UK, Wales, Ireland
CNN	Editor, commentary, opinions, affiliate, according, London, official, Pakistan
Sydney Morning Herald	Australian, Australia, Melbourne, Fairfax, Canberra, rugby, Queensland, Brisbane,
Moscow times	Russian, Russia, Interfax, soviet, Barack, Petersburg, Siberia, Sergei, Kremlin, Duma, Yury, Oleg
DNA India	Mumbai, indian, singh, Bollywood, Kapoor, cricket, crore, Bangalore, lakh, Pune, Kumar, Tata, Gandhi

Table 3 Some of the relevant keywords that discriminate a publisher from the others

Ukraine-Russia test

In contrast to the previous experiment, where we analyzed all types of articles written by the news publishers, we wanted in this experiment to see how publishers differ when they report about similar content. For this purpose we needed to identify a topic that spanned over a longer period of time and was reported by each publisher frequently enough in order to have enough testing articles. The reporting topic that we decided to use was the tense relationship between Ukraine and Russia that started in the beginning of 2014.

To identify the relevant articles we used the following criteria: Articles have to

1. mention Russia and Ukraine
2. have to be published in the time period between Jan 15th, 2014 and June 1st, 2014, and
3. have to be categorized into the Business or Society category (or any of their subcategories). This criterion was used to remove the articles related to the Olympic games and other irrelevant topics.

Given the selected criteria we found that not all news publishers produced a large enough set of articles in order to perform a reliable test of source separability. For this purpose we decided to include in the test only publishers that wrote at least 500 articles that match the mentioned criteria. To make the learning problem equally hard for all pairs of news publishers we again had to create equally sized sets of articles from each publisher. Since the lowest ranking publisher wrote 500 articles we therefore used for each publisher 500 randomly sampled articles matching the criteria. The remaining of the testing procedure is the same as described at the beginning of Section 0 with the exception that we only selected for each publisher top 100 features (words) due to a smaller sample of documents.

The MDS graph representing how well we can separate between the tested set of publishers can be seen in Figure 26Figure 26 (the corresponding data is shown in a table in the Appendix). As expected, the US publishers (Washington Post, USA Today and Huffington Post) are among the hardest to discriminate which is why they are located close together. Other news sources are in general well separable from each other. The only exception are surprisingly The Moscow Times and DNA India. The accuracy of separability between the two publishers is 74.5%, which is relatively low compared to other pairs of news publishers.

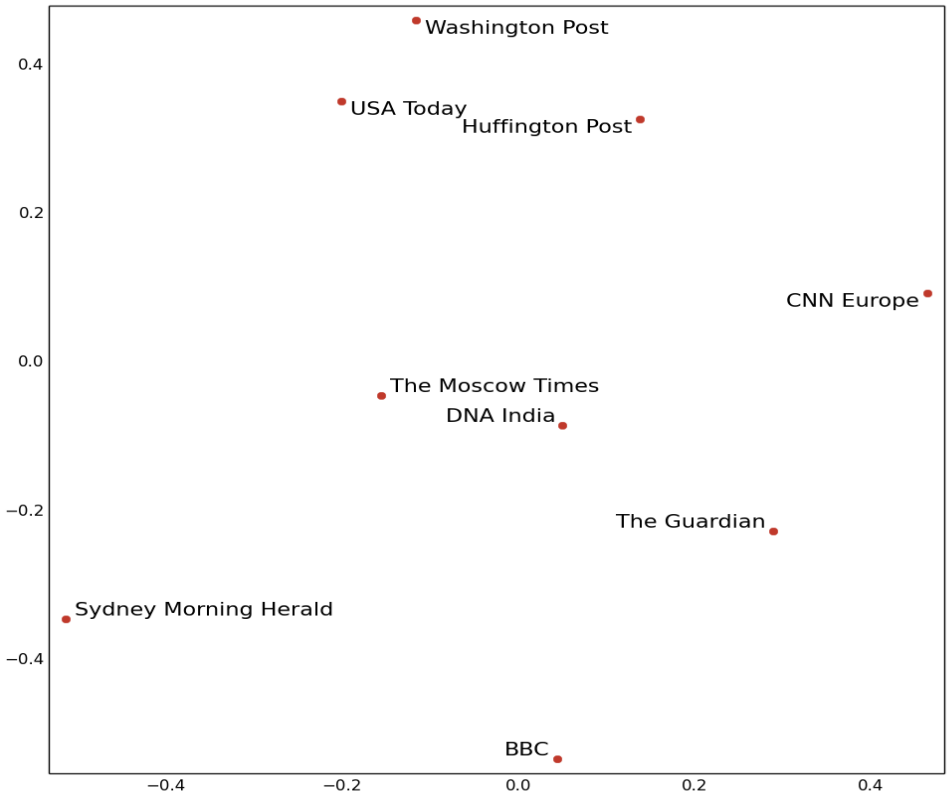

Figure 26 Separability between publishers on topics related to Ukraine-Russia incident

Since we were interested in understanding what are the differences in reporting by the Moscow Times and other news publishers we have extracted some top keywords that separate the Russian publisher from the other publishers. Some of the top keywords are listed in Table 1Table 4.

News source	Keywords
Non-Russian sources	troops, allies, invasion, aggression, excellency, insurgents, investigation, defense, ethnic, seized
The Moscow Times	overthrow, authorities, western, election, khodorkovsky, detained, annexation, pressure, seizure, fanning, imposed, domestic

Table 4 Some relevant keywords for discriminating between the Moscow Times and other publishers

Objective vs. biased test

As the last experiment in this series we also wanted to see how much difference there is in the objective reporting compared to reporting that is supposedly biased. In this experiment we only took the articles from the Washington Post. The Washington Post has their articles organized into 11 categories, such as World, Local, Politics and Business. Two of the categories are also Blogs and Opinions. Based on their names we can assume that the content published in these two categories is less objective and expresses the author's opinions and beliefs.

To test this hypothesis we performed an experiment where we put articles into two groups – one group contained articles from the "biased" categories and the second group contained articles from other categories. To make the classification problem as hard as possible we randomly selected 10.000 articles from each of the two groups of articles. The rest of the experiments was performed as it is described at the beginning of Section 0.

The SVM model built using the top 1.000 most discriminative keywords was able to achieve 81.4% classification accuracy in separating between the two groups of articles. The accuracy is not low, but it is also not as high as we might expect given that blogs and opinions should be using a vocabulary that is easily discernible. There results suggest that articles in these two categories are still written by professional journalists who are trained in writing in a factual and objective way. Putting the articles in these two categories might be just the journal's way of providing to their readers an explicit disclaimer that the article might be expressing the author's point of view and not only factual information.

The top keywords discriminating between the two groups are shown in Table 5. We can see that the objective categories use more formal words (e.g. spokesman, officials, insight), whereas the opinions and blogs use more adjectives and adverbs. To see how particular part-of-speech tags contribute to the separability of articles we have also performed the same experiment, but only using words of particular type. Using adjectives only, we were able to achieve 71.7% accuracy, using adverbs 73.9%, using nouns 78.9%, and using verbs 76.8%. We can see adjectives and adverbs are not used frequently enough to be used as the only type of data to discriminate between articles. Using verbs and especially nouns we can, however, achieve accuracy that is close to the one obtained with all word types.

News source categories	Keywords
Objective categories	Copyright, redistributed, officials, rewritten, associated (press), insight, said, material, survivors, marketplace, broadcast, spokesman, stocks, shares, manners, humanoid, suspect
Opinions and Blogs	Today, likely, actually, via, reliable, editorial, getty, yesterday, inequality, redskins, regime, beer, regarding, nationals, specials, fans, layman, actual, shouldn't, perhaps, hey, probably, gonna, thanks, basically, digs

Table 5 Some keywords for discriminating between objective and biased articles from Washington Post

Similarity between publishers in choosing the events they report about

In this experiment we wanted to compute how biased the news publishers are about which events they report about. In our previous experiments we have already shown their bias with respect to geographic or topic coverage of events. Here we want to compute for each pair of publishers a single numerical score that represents how much do the events covered by the two publisher overlap.

To obtain information about which events have been covered by each of the tested news publishers we have again used information from the Event Registry. For each article written by the news publisher we can obtain from the Event Registry the ID of the associated event. Our task is then to compute each pair of news publishers how similar their corresponding sets of event IDs are.

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.

Since the tested publishers vary significantly in the number of articles they publish we decided to use the cosine similarity when comparing them. We have also decided to focus only on the set of publishers reporting in English, which produce general articles (we ignored the financial and technology news sources). The MDS plot showing the results is shown in Figure 27 (the corresponding table is in Section 0). Based on the data we can see that the best agreement in the reported events is between the European set of news publishers (BBC, Daily Telegraph, The Independent, The Guardian and Daily Mail). Their cosine similarity is between 0.24 and

0.3. This is expected due to their close geographic location. Equally high agreement can also be seen between the events covered by The Hindu and DNA India (0.27). The similarity between the US publishers is, however, significantly lower and is between 0.07 and 0.17. The low similarity is most likely due to the size of the US and the large number of events that publishers have to choose from. This reasoning does not, however, agree with the fact that the two Indian publishers have very high agreement despite India's size.

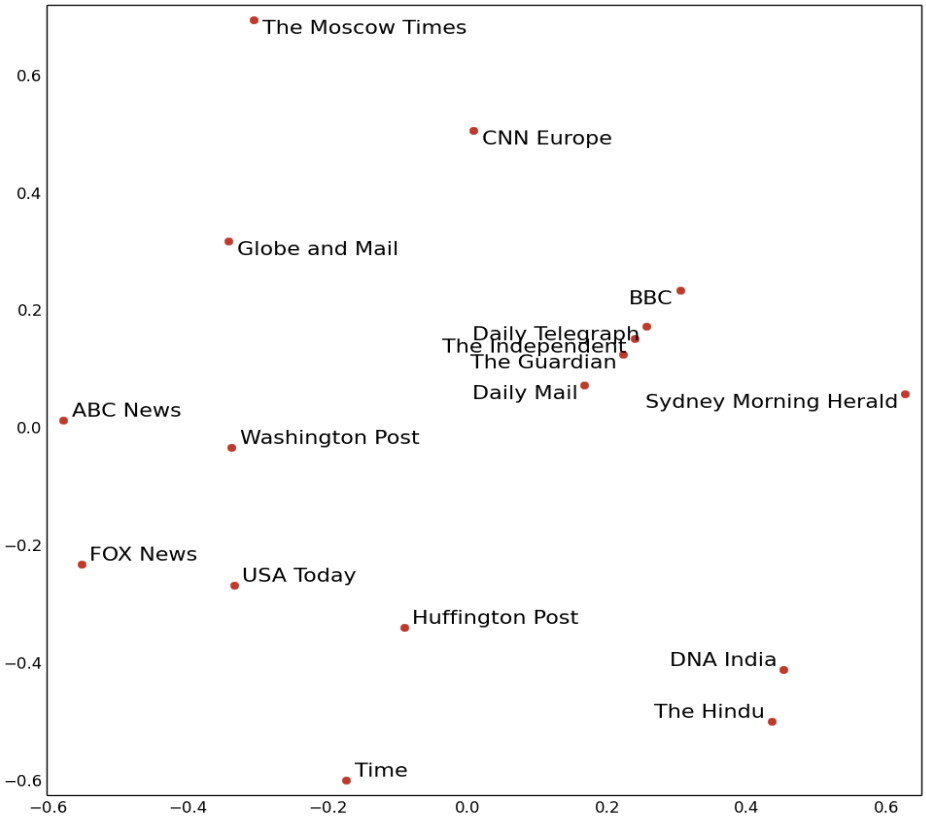

Figure 27 Similarity between publishers in choosing the events

5 Future work

Although we have with our experiments tested bias in several ways there are naturally a large number of ways in which bias can be present and expressed in the news. A big area of research that we have not touched includes opinion mining and sentiment analysis. The reason we have decided to skip it from our tests is that our current methodology does not achieve a high enough accuracy that we would be willing to make conclusions based on the results. Current systems for sentiment analysis are usually domain specific (such as, for example, movie or hotel reviews) and rely on assigning a particular sentiment to individual words. When analyzing general news, however, the sentiment of a word can be highly dependent on the context in which the word appears.

Additional type of bias that would be interesting to analyze would also be gender bias. In this case we would analyze who gets cited more in the news – males or females. It would be especially interesting to see if there are any deviations from the distribution if we also take into account the event category.

6 Conclusion

News is the main source of information for the majority of people. The language of news is a vital, integrated and often taken for granted part of our lives. Ordinary readers often do not give a second thought of how news language is constructed and whether reporting is objective or biased. It is in the public's interest to be aware of the problems of news bias. It is vital to examine it so that the people can learn to recognize it and to prevent its influence on their understanding of the information described in the article.

There are several ways in both social and computer sciences to detect potential news bias. In this report we have provided a detailed review of various experiments carried out to detect news bias across 30 pre-selected news publishers across the world: 26 news websites reporting in English from the USA, India, Russia, Canada and Great Britain as well as 2 Spanish and 2 German outlets. The articles we examined were collected between 15th of December 2013 and 15th of August 2014. Specifically, 11 main experiments were carried out, ranging from simple length and grammatical analysis to bias related to types of covered events. For each of the experiments we presented the relevant details of the experiment as well as the interpretation of the obtained results.

The collected results clearly confirm the existence of various types of bias across different news sources. Most types of bias are connected to the geographic differences/similarities between the analyzed news publishers. European news sources, for example, put more emphasis on describing events occurring in Europe, while US publishers cite more the events occurring in US. Similarly, the news sources that are geographically close are harder to discriminate than those that are further apart. A similar pattern can be observed which is related to the citations of news agencies. US news sources mostly cite Associated Press (a US based news agency), whereas the European publishers prefer to cite the European news agencies. We have also detected a strong bias among the tabloid outlets such as Daily Mail or Stern Magazine to write longer headlines, shorter articles and to use more colorful language making use of numerous adjectives and adverbs. The choice of words for a headline and the length of a headline depend on the style of a news outlet and the journalists working for it. German outlet Stern Magazine has a reputation of a popular tabloid newspaper and the reference to more adjectives is not surprising. As expected, we also found a larger percentage of adverbs and adjectives on websites like Gizmodo, GigaOM and The Next Web which write about new products and their evaluations. In some news sources we have also noted a bias related to using proper nouns (names of people, locations and organizations). The lack of proper nouns in news reporting in die Welt and Stern Magazine, for example, does not add credit to them as a valid source of information. Finally, our report has shown that readability of articles from the analyzed publishers varies significantly. Most hard to understand seem to be the Spanish and German news articles which could be due to a high number of compound words.

References

[AO10] Ali, O., Flaounas, I., De Bie, T., et al. (2010) "Automating News Content Analysis: An Application to Gender Bias and Readability" JMLR: Workshop and conference Proceedings 11.
[BA10] Balahur, A., Steinberg, R. et al. "Sentiment Analysis in the News" (2010). Proceedings of the 7th International Conference on Language Resources and Evaluations, Malta, pp. 2216-2220.
[BI05] Borg, I., Groenen, P. (2005). *Modern Multidimensional Scaling: theory and applications* (2nd ed.). New York: Springer-Verlag. pp. 207–212. ISBN 0-387-94845-7.
[CJ95] Chall, J.S., Dale, E., "readability Revisited: the new Dale-Chall readability formula (1995). Brookline Books, Cambridge, MA.
[D432] Deliverable D4.3.2 Final event extraction prototype, XLike project.
[FI10] Flaounas, I., Turchi, M., et al. (2010) "The Structure of the EU Mediasphere" PLoS ONE. Vol.5, Issue 12.
[FM00] Friendly, Michael. *Visualizing categorical data*. SAS Institute, 2000.
[GJ65] Galtung, J., & Ruge, M. H. (1965). The structure of foreign news the presentation of the Congo, Cuba and Cyprus Crises in four Norwegian newspapers. *Journal of peace research*, *2*(1), 64-90.
[HJ06] Harrison, J. (2006). News. Routledge, New York, p.99.
[KS11] Karmakar, S. (2011) "Syntactic and Semantic Analysis and Visualization of Unstructured English Text" Computer Science Dissertations. Paper 61.
[KJ75] Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S. (1975) "Derivation of new readability formulas (Automated Readibility Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel Research Branch Report 8-75, Millington, TN.
[LG14] Leban, G., Fortuna B., Brank J., Grobelnik M., Event Registry – Learning About World Events From News, WWW 2014, pp. 107-111.
[PS09] Park, S., Kang, S. (2009) "Newscube: delivering multiple aspects of news to mitigate media bias." Proceedings of the 27th international conference on Human factors in computing systems.

Annex 1

Length Analysis

Token count

Label	Average	St. dev.	Q1	Q2	Q3
CNN Europe	725.9	468.7	383	652	963
Financial Times	748.9	538.3	422	603	873
The Guardian	660.4	499.6	342	577	842
Globe and Mail	570.5	341.8	314	573	768
Washington Post	638.4	509.1	311	561	832
USA Today	758.6	786.2	281	535	889
Sydney Morning Herald	604.7	432.3	356	527	744
Daily Telegraph	612.9	553.5	317	501	763
El Mundo	525.5	319.0	308	464	662
Huffington Post	630.0	633.5	212	460	840
Wall Street Journal	584.1	587.9	86	435	928
The Independent	520.1	357.8	276	428	658
ABC News	418.8	326.0	172	398	569
Daily Mail	493.9	384.6	273	398	593
WSJ Blogs	457.6	345.4	260	392	547
GigaOM	492.3	381.9	265	384	622
FOX News	474.2	415.1	169	378	652
ABC.es	424.8	279.0	244	361	533
The Moscow Times	525.6	587.9	229	357	646
The Hindu	410.8	277.3	241	349	484
Economic Times	369.5	201.6	229	339	477
Time	474.7	421.9	182	317	676
DNA India	378.3	283.0	203	311	468
The Next Web	486.7	566.7	178	291	537
BBC	348.3	290.7	152	271	434
Stern Magazine	367.0	392.8	150	269	470
Business Insider	373.2	399.6	163	266	456
Gizmodo	409.5	479.9	171	259	437
Die Welt	396.2	397.1	100	256	564
Boston Business Journal	228.3	81.5	160	236	295

Token count in headlines

Label	Average	St. dev.	Q1	Q2	Q3
Daily Mail	16.73	5.81	12	16	20
The Independent	13.65	4.83	10	13	16
The Next Web	13.12	3.26	11	13	15
Business Insider	12.10	3.72	10	12	14
GigaOM	11.81	2.89	10	12	14

ABC.es	10.89	3.39	8	11	13
Globe and Mail	11.20	3.13	9	11	13
Washington Post	10.18	3.26	8	10	12
FOX News	10.24	4.00	8	10	13
Gizmodo	10.26	2.16	9	10	12
El Mundo	9.41	3.79	6	10	12
Daily Telegraph	10.23	3.54	8	10	12
Huffington Post	10.19	3.18	8	10	12
DNA India	10.53	3.14	8	10	12
Economic Times	10.18	2.74	8	10	12
The Guardian	10.12	2.85	8	10	12
Boston Business Journal	9.38	2.82	7	9	11
Sydney Morning Herald	8.88	3.13	7	9	11
The Moscow Times	8.74	2.60	7	9	10
WSJ Blogs	8.84	2.34	7	9	10
Time	9.19	2.56	7	9	11
ABC News	7.99	2.00	7	8	9
USA Today	8.11	1.75	7	8	9
Stern Magazine	8.13	2.72	7	8	10
Financial Times	7.66	2.89	5	8	10
Die Welt	8.64	2.68	7	8	10
Wall Street Journal	7.32	2.72	5	7	9
The Hindu	7.43	2.41	6	7	9
CNN Europe	7.66	2.55	6	7	9
BBC	6.23	1.84	5	6	7

Grammatical analysis

Percentage of adjective in articles per publisher

Label	Average	St. dev.	Q1	Q2	Q3
Stern Magazine	8.44	2.13	7.14	8.33	9.59
Die Welt	8.14	2.16	6.87	8.13	9.35
Financial Times	7.15	1.52	6.14	7.10	8.09
Gizmodo	6.87	1.93	5.62	6.81	8.05
WSJ Blogs	6.67	1.90	5.39	6.58	7.85
GigaOM	6.51	1.76	5.38	6.46	7.62
The Next Web	6.44	1.81	5.24	6.44	7.60
The Guardian	6.51	1.75	5.41	6.44	7.56
The Moscow Times	6.38	1.89	5.09	6.35	7.60
Globe and Mail	6.40	1.79	5.24	6.29	7.43
Business Insider	6.28	2.05	4.91	6.15	7.47
Wall Street Journal	6.42	2.37	4.86	6.10	7.60
Time	6.09	1.90	4.88	6.07	7.26
Daily Telegraph	6.16	1.71	5.06	6.07	7.18
The Independent	6.12	1.66	5.03	6.05	7.13

Washington Post	6.14	1.78	5.01	6.04	7.16
Economic Times	6.01	2.06	4.62	5.95	7.33
Huffington Post	5.93	2.11	4.63	5.92	7.24
Boston Business Journal	6.00	2.04	4.61	5.88	7.26
Sydney Morning Herald	5.93	1.69	4.83	5.83	6.96
CNN Europe	5.83	1.64	4.73	5.78	6.86
ABC.es	5.76	1.86	4.50	5.67	6.88
Daily Mail	5.77	1.69	4.63	5.65	6.77
ABC News	5.72	1.83	4.52	5.63	6.82
BBC	5.67	1.89	4.43	5.60	6.85
El Mundo	5.63	1.69	4.48	5.56	6.67
FOX News	5.56	1.80	4.35	5.48	6.65
The Hindu	5.48	1.98	4.10	5.36	6.72
USA Today	5.43	1.77	4.22	5.36	6.54
DNA India	5.34	1.94	4.00	5.26	6.53

Percentage of adverbs in articles per publisher

Label	Average	St. dev.	Q1	Q2	Q3
Gizmodo	5.91	1.88	4.62	5.83	7.04
Die Welt	5.62	2.65	3.74	5.56	7.45
Stern Magazine	5.63	2.61	3.97	5.42	7.18
The Next Web	4.67	1.77	3.46	4.60	5.78
GigaOM	4.55	1.47	3.59	4.53	5.47
Huffington Post	4.38	1.99	3.05	4.34	5.66
Business Insider	4.39	1.75	3.18	4.32	5.47
The Guardian	4.34	1.70	3.18	4.27	5.44
The Independent	4.27	1.61	3.14	4.20	5.33
Financial Times	4.17	1.51	3.08	4.06	5.14
Daily Mail	4.11	1.49	3.09	4.02	5.04
Daily Telegraph	4.09	1.64	2.94	3.99	5.16
Washington Post	3.96	1.58	2.94	3.90	4.93
Time	3.93	1.72	2.71	3.83	5.04
Globe and Mail	3.77	1.53	2.74	3.69	4.71
CNN Europe	3.67	1.40	2.72	3.63	4.58
Sydney Morning Herald	3.62	1.49	2.59	3.52	4.56
WSJ Blogs	3.58	1.40	2.60	3.49	4.44
ABC News	3.47	1.73	2.27	3.31	4.46
USA Today	3.36	1.44	2.34	3.27	4.26
DNA India	3.38	1.56	2.29	3.23	4.33
BBC	3.28	1.61	2.16	3.20	4.31
FOX News	3.26	1.46	2.26	3.20	4.15
El Mundo	3.24	1.48	2.21	3.11	4.12
The Moscow Times	3.11	1.48	2.07	3.03	4.03
Economic Times	3.15	1.64	2.01	2.96	4.03
The Hindu	3.08	1.49	2.02	2.96	4.01

ABC.es	3.05	1.46	2.03	2.91	3.90
Wall Street Journal	2.86	1.65	1.70	2.74	3.84
Boston Business Journal	2.77	1.51	1.71	2.60	3.67

Percentage of proper nouns in articles per publisher

Label	Average	St. dev.	Q1	Q2	Q3
USA Today	18.47	9.60	11.86	15.88	22.44
Boston Business Journal	15.84	6.26	11.44	15.22	19.53
Wall Street Journal	15.79	6.69	11.15	15.18	19.82
The Hindu	15.25	7.07	10.43	14.43	18.97
FOX News	14.54	5.58	10.71	13.87	17.55
DNA India	14.55	6.27	10.22	13.82	17.98
The Moscow Times	13.82	6.34	10.09	13.02	16.31
Sydney Morning Herald	13.56	6.17	9.77	12.80	16.23
Time	13.48	6.17	9.26	12.65	16.58
ABC News	13.30	6.24	9.35	12.57	16.24
Economic Times	13.34	5.95	9.20	12.50	16.53
CNN Europe	13.01	5.37	9.50	12.42	15.75
BBC	13.36	6.88	8.90	12.17	16.18
Washington Post	12.95	6.91	8.77	11.74	15.38
WSJ Blogs	12.31	5.14	8.63	11.57	15.33
The Next Web	12.29	6.31	7.81	11.47	15.54
Business Insider	11.58	5.65	7.63	10.89	14.73
Huffington Post	13.83	12.56	6.44	10.89	16.67
Globe and Mail	11.15	5.00	7.80	10.68	13.84
The Independent	11.32	5.42	7.74	10.64	13.94
Daily Telegraph	11.59	6.37	7.69	10.57	14.02
Daily Mail	11.12	5.12	7.66	10.49	13.80
GigaOM	10.93	4.74	7.58	10.27	13.64
Financial Times	9.93	3.94	7.26	9.58	12.20
The Guardian	10.16	5.68	6.68	9.34	12.44
Gizmodo	8.18	4.09	5.39	7.63	10.17
ABC.es	8.02	4.15	5.25	7.48	10.10
El Mundo	7.58	3.64	5.10	7.11	9.46
Stern Magazine	4.44	3.02	2.36	3.94	5.93
Die Welt	4.18	3.05	2.15	3.57	5.54

Percentage of verbs in articles

Label	Average	St. dev.	Q1	Q2	Q3
BBC	16.71	3.33	14.69	16.83	18.85
Daily Mail	16.38	2.70	14.63	16.41	18.18
The Independent	16.02	2.69	14.40	16.08	17.76
Daily Telegraph	15.89	2.87	14.29	16.05	17.72
CNN Europe	15.83	2.46	14.43	15.96	17.45

DNA India	15.88	2.94	14.02	15.93	17.80
The Guardian	15.77	2.74	14.22	15.88	17.51
ABC News	15.66	2.92	13.92	15.73	17.52
FOX News	15.65	2.78	13.91	15.72	17.43
Sydney Morning Herald	15.42	2.70	13.83	15.55	17.16
Globe and Mail	15.28	2.60	13.68	15.39	16.95
Washington Post	15.04	2.89	13.61	15.31	16.82
Huffington Post	14.76	3.69	13.18	15.25	17.05
The Hindu	15.05	2.94	13.27	15.18	16.98
Business Insider	15.11	2.73	13.39	15.17	16.86
Time	15.02	2.72	13.37	15.13	16.75
The Moscow Times	15.00	2.85	13.64	15.09	16.67
GigaOM	15.00	2.22	13.62	15.08	16.45
Financial Times	14.92	2.16	13.55	14.96	16.33
Gizmodo	14.79	2.39	13.29	14.83	16.31
The Next Web	14.70	2.58	13.14	14.77	16.36
Economic Times	14.59	3.01	12.76	14.74	16.61
WSJ Blogs	14.57	2.31	13.10	14.65	16.06
USA Today	13.70	3.38	11.70	14.12	16.05
Boston Business Journal	13.78	2.76	11.96	13.76	15.58
Wall Street Journal	13.33	3.17	11.27	13.43	15.39
El Mundo	11.29	2.34	9.77	11.27	12.79
ABC.es	11.16	2.53	9.50	11.08	12.74
Die Welt	10.48	2.68	8.75	10.54	12.23
Stern Magazine	10.01	3.03	8.48	10.17	11.89

Readability analysis

Dale-Chall scores

Label	Average	St. dev.	Q1	Q2	Q3
ABC.es	9.91	0.87	9.37	9.88	10.37
El Mundo	9.82	0.77	9.34	9.78	10.24
Stern Magazine	9.88	1.38	9.17	9.69	10.27
Die Welt	9.70	0.91	9.12	9.67	10.23
Wall Street Journal	9.21	1.05	8.52	9.14	9.80
The Hindu	8.98	0.95	8.36	8.92	9.54
The Moscow Times	8.92	0.87	8.39	8.86	9.35
DNA India	8.88	0.95	8.26	8.84	9.43
Financial Times	8.91	1.02	8.20	8.77	9.49
Economic Times	8.78	0.95	8.19	8.74	9.32
USA Today	8.96	1.28	8.14	8.73	9.51
FOX News	8.74	0.85	8.18	8.70	9.24
Time	8.73	0.93	8.12	8.67	9.26
Sydney Morning Herald	8.71	1.29	8.10	8.63	9.19
Daily Telegraph	8.66	1.23	7.98	8.53	9.15

Daily Mail	8.58	0.94	7.95	8.52	9.14
CNN Europe	8.55	0.82	8.03	8.52	9.02
The Guardian	8.62	1.21	7.95	8.50	9.12
Gizmodo	8.52	0.79	7.99	8.50	9.02
ABC News	8.50	0.98	7.88	8.49	9.07
The Independent	8.56	0.96	7.92	8.48	9.10
WSJ Blogs	8.48	0.79	7.97	8.47	8.98
Boston Business Journal	8.51	0.89	7.92	8.46	9.05
Huffington Post	8.58	1.40	7.70	8.40	9.17
GigaOM	8.45	0.71	7.97	8.40	8.91
Globe and Mail	8.43	0.88	7.87	8.40	8.94
Washington Post	8.45	1.09	7.81	8.33	8.91
BBC	8.42	1.01	7.75	8.31	8.95
The Next Web	8.31	0.84	7.76	8.27	8.82
Business Insider	8.27	0.89	7.68	8.23	8.81

Flesch-Kincaid Reading Ease scores

Label	Average	St. dev.	Q1	Q2	Q3
Gizmodo	72.18	9.13	66.46	72.69	78.47
Daily Mail	70.86	9.86	65.33	71.48	76.96
BBC	68.23	12.60	63.30	69.70	75.75
Huffington Post	68.75	13.15	61.00	69.44	77.56
Business Insider	68.26	10.61	61.48	68.72	75.63
ABC News	68.79	12.26	60.28	68.48	77.83
USA Today	67.89	11.78	61.63	68.31	74.85
The Independent	67.75	9.94	61.52	68.02	74.34
The Next Web	66.93	8.67	61.58	67.61	72.67
Washington Post	66.99	14.17	60.19	67.38	75.40
Daily Telegraph	66.66	15.05	60.60	67.03	73.77
CNN Europe	66.93	9.68	60.83	66.87	73.38
Globe and Mail	66.60	11.13	59.65	66.57	73.96
Sydney Morning Herald	66.03	18.54	59.39	66.49	73.49
The Guardian	65.48	14.82	58.90	65.78	72.71
Time	64.86	11.98	58.03	65.11	71.94
FOX News	64.84	10.30	58.00	64.76	71.97
GigaOM	64.16	9.48	58.21	64.50	70.75
WSJ Blogs	63.13	9.70	57.17	63.22	69.10
Financial Times	63.37	9.14	57.19	63.15	69.27
DNA India	62.88	11.53	55.90	63.04	70.25
The Hindu	62.04	11.39	54.84	62.15	69.58
Boston Business Journal	61.07	10.60	54.76	61.51	67.97
Economic Times	60.71	11.34	54.10	61.33	68.01
Wall Street Journal	58.71	14.69	51.53	60.76	68.26
The Moscow Times	56.32	9.79	50.36	56.32	62.11
Stern Magazine	31.60	16.93	21.62	32.07	41.83

Die Welt	26.72	14.24	18.04	27.48	36.43
El Mundo	13.34	14.52	4.26	13.36	23.02
ABC.es	12.35	14.03	3.33	12.28	22.02

Newswire citations analysis

Percentage of articles that cite at least one agency

News Publisher	Percentage
USA Today	32.29
The Moscow Times	25.73
FOX News	25.61
Washington Post	22.15
ABC News	20.33
Stern Magazine	19.56
Huffington Post	15.70
The Next Web	15.14
Sydney Morning Herald	14.63
Time	14.02
Die Welt	13.18
Wall Street Journal	8.41
Globe and Mail	7.71
The Hindu	7.28
Economic Times	6.60
Daily Telegraph	6.18
Business Insider	6.11
WSJ Blogs	5.72
DNA India	4.54
CNN Europe	4.46
The Independent	3.97
The Guardian	3.64
BBC	3.00
GigaOM	2.54
Boston Business Journal	2.31
ABC.es	2.25
Daily Mail	2.17
Financial Times	2.01
Gizmodo	1.89
El Mundo	1.85

Percent of all articles that cite a news agency

News Agency	Percent of articles that cite it
Associated Press	4.713
Reuters	1.782
Agence France-Presse	1.109
Deutsche Presse-Agentur	0.549
Interfax	0.32
Xinhua News Agency	0.285
Press Association	0.281
RIA Novosti	0.12
American Press Agency	0.037
Syrian Arab News Agency	0.023
IRNA	0.022
Dow Jones Newswires	0.021
MENA	0.018
Agenzia Nazional Stampa Associata	0.011
ITAR-TASS	0.01
China News Service	0.01
Algemeen Nederlands Persbureau	0.01
United Press International	0.008
ISNA	0.005
Slovenian Press Agency	0.004
All Headline News	0.004
Prensa Latina	0.003
FARS	0.002
Agencia Brasil	0.002
Mehr News Agency	0.001
Khaama Press	0.001
Qatar News Agency	0.001
BNO News	0.001
Ukrainian News Agency	0
Algeria Press Service	0

Number of cited agencies per news source

	AP	Reuters	AFP	DPA	Interfax	Press Association
Washington Post	7,666	615	276	4	208	89
FOX News	5,062	592	46	16	56	85
Time	1,152	442	117	1	26	18
The Moscow Times	178	714	32	1	1,199	0
Stern Magazine	33	832	737	1,170	314	14
BBC	652	576	595	7	82	84

Guardian	695	685	277	13	53	261
Sydney M.H.	641	1,146	1,382	36	40	71
CNN	106	112	68	0	34	82

Topic (category) coverage bias

Coverage bias of top level categories for all publishers

	Arts	Business	Computers	Games	Health	Home	Education	Science	Shopping	Society	Sports
ABC News	1754	4019	625	1610	1655	502	476	1298	1211	11100	2186
BBC	6074	10643	1726	3150	4609	628	1528	4408	2136	25459	9653
CNN Europe	3260	3883	946	725	2797	515	928	2470	1260	20506	3322
FOX News	1428	3809	711	1181	1898	611	824	1725	1155	14365	1040
DNA India	2550	4930	1471	514	1176	291	486	1568	1413	14950	2418
The Hindu	1501	4767	833	398	1041	224	533	2056	1078	10157	1392
Huffington Post	6766	5256	1353	871	5662	1536	1264	2880	3488	18251	1820
Time	2849	3056	1385	656	1166	303	342	1224	1055	9548	659
Washington Post	4381	5573	1503	3271	1978	968	2026	2111	1604	20084	3477
Daily Mail	16424	15702	2569	6483	10979	3960	1790	6522	11121	34414	21435
Daily Telegraph	4440	8143	1543	2031	2326	1069	1222	2198	2085	13598	5979
The Guardian	9234	8857	1741	2586	2490	827	1244	3168	2261	19566	7244
The Independent	5419	5638	1017	2373	1855	810	832	1874	1866	12926	7082
USA Today	6564	9320	2435	3674	2913	1097	1285	3332	2901	17619	7206
Financial Times	1730	1627	199	70	95	78	171	227	297	1573	161
Sydney Morning Herald	2488	6898	854	1594	1002	359	356	1334	731	7098	3180
Globe and Mail	3282	15029	2071	2742	2667	978	927	2870	1907	15509	4067
The Moscow Times	808	3730	360	125	227	70	134	679	229	6801	858
Boston Business Journal	7600	97735	11642	2242	12625	3285	6189	9938	4348	25215	5104
Economic Times	1230	30078	3905	268	1077	811	520	3649	2605	19783	744
Wall Street Journal	3820	18561	3035	995	1941	908	881	2533	2197	16836	2447
WSJ Blogs	490	4333	1858	266	228	209	147	411	319	5774	421
Business Insider	1880	7989	4513	639	500	399	335	1306	1603	5856	679
Gizmodo	1257	2794	1844	245	341	73	114	1408	1366	931	384
GigaOM	655	1993	3982	86	63	56	26	312	712	643	69
The NextWeb	626	1413	3642	320	71	50	65	200	501	532	98

Coverage bias for Society sub-categories

	Crime	Death	Ethnicity	Gay, Lesbian, and Bisexual	Government	History	Issues	Law	Military	People	Philanthropy	Philosophy	Politics	Relationships	Religion and Spirituality	Support Groups	Work
ABC News	297	69	267	253	316	1021	3395	836	147	343	111	21	558	189	633	55	92
BBC	1238	254	799	210	793	1409	7032	1610	139	1188	318	25	1290	240	2822	243	197
CNN Europe	725	79	716	283	298	1444	8132	1379	336	715	215	36	1060	388	763	82	86

FOX News	562	92	297	211	222	1083	4566	1447	381	382	146	29	667	144	958	89	133
DNA India	451	35	1017	64	491	1027	3222	976	43	611	128	43	2100	211	975	68	49
The Hindu	243	32	586	27	366	740	2659	651	35	396	113	16	1191	122	829	26	61
Huffington Post	345	89	681	909	170	1135	3574	888	158	1343	274	735	765	571	1674	134	229
Time	222	38	398	148	159	1071	3350	407	166	470	68	27	615	150	586	40	49
Washington Post	339	83	658	365	698	2126	6440	1421	230	621	189	64	2065	266	978	47	249
The Moscow Times	116	22	575	123	683	704	2054	326	43	51	35	5	298	227	323	2	26
WSJ Blogs	79	6	49	79	980	138	969	1788	34	67	20	7	145	13	120	0	179

Coverage bias for Computers sub-categories

	Hardware	Internet	Mobile Computing	Robotics	Security	Software	Systems
WSJ Blogs	117	833	144	38	180	369	304
Business Insider	190	1794	305	213	210	740	1167
Gizmodo	159	645	56	181	144	320	303
GigaOM	189	1004	284	55	193	1024	753
The NextWeb	42	1692	208	21	77	945	608

Coverage bias for Health sub-categories

	Addictions	Alternative	Animal	Conditions and Diseases	Education	Home Health	Medicine	Mental Health	News and Media	Nursing	Pharmacy	Public Health and Safety	Reproductive Health	Senior Health	Specific Substances	Support Groups	Weight Loss
ABC News	58	156	113	411	14	34	186	72	4	26	97	201	54	9	71	53	59
CNN Europe	98	34	132	861	20	59	282	160	7	25	166	363	66	16	94	82	96
FOX News	77	52	149	462	20	55	218	51	16	37	104	301	63	12	50	61	56
Huffington Post	285	166	343	1466	61	98	320	820	42	64	191	87	225	70	190	273	281
Time	60	15	102	301	35	13	74	100	10	19	59	103	38	12	74	35	24
Daily Mail	486	679	880	3129	41	132	1040	495	13	181	370	441	393	128	360	489	657
Boston Business Journal	144	330	214	996	568	1452	4406	126	494	472	626	214	99	397	365	198	206

Coverage bias for Sports sub-categories

	Baseball	Basketball	Bowling	Cricket	Cycling	Equestrian	Events	Fantasy	Football	Golf	Hockey	Motorsports	Resources	Running	Soccer	Tennis	Water Sports	Winter Sports
Daily Mail	938	106	492	852	569	757	745	1293	3066	612	171	1280	407	235	8125	456	228	325
Daily Telegraph	200	27	176	384	253	227	266	218	918	119	76	480	130	70	2101	137	43	212
The Guardian	275	28	270	532	344	333	338	316	1311	188	97	573	153	104	2211	92	59	104
The Independent	343	22	138	273	229	141	255	450	1063	135	35	363	124	86	2957	84	64	113
USA Today	388	397	652	11	220	282	649	225	910	871	196	782	103	131	498	128	45	174
Sydney Morning Herald	89	48	156	325	110	161	195	77	644	118	126	303	47	27	596	69	33	72
Globe and Mail	267	99	101	6	109	126	637	130	232	285	808	258	19	77	604	99	29	143

Bias in the speed of reporting

Time after the first event report in hours

Label	Average	St. dev.	Q1	Q2	Q3
Die Welt	27.030	107.710	0.117	2.283	11.533
BBC	26.873	114.552	0.550	4.867	17.933
El Mundo	26.119	97.468	0.550	2.983	13.167
ABC.es	25.549	81.904	0.767	3.717	15.200
Stern Magazine	20.938	92.336	0.333	2.650	10.183
GigaOM	19.984	40.927	1.000	6.417	20.367
The Next Web	19.605	54.621	0.567	4.550	18.233
The Moscow Times	19.263	51.201	0.000	5.100	19.183
Business Insider	18.117	36.706	0.883	5.883	20.800
WSJ Blogs	17.974	45.210	0.767	5.733	20.183
Wall Street Journal	17.902	52.941	0.017	4.300	16.783
Financial Times	17.600	42.198	0.783	5.267	17.183
CNN Europe	16.555	31.497	0.650	5.417	19.583
Time	16.160	32.129	1.383	5.617	18.583
Boston Business Journal	16.045	29.713	0.267	5.633	19.250
The Hindu	15.755	44.542	0.583	4.283	16.417
The Independent	15.726	40.373	1.650	7.383	18.683
Daily Mail	15.297	30.490	1.733	7.067	18.650
Huffington Post	15.276	29.852	0.833	5.067	18.400
Gizmodo	15.250	37.192	0.117	4.300	15.783
DNA India	15.006	36.058	0.250	4.667	16.950
Sydney Morning Herald	14.803	31.763	0.650	5.367	16.550
Globe and Mail	13.701	35.099	0.850	4.250	14.767
Daily Telegraph	13.687	39.925	1.150	5.683	15.700
Washington Post	13.572	30.247	0.133	4.200	15.817
Economic Times	13.163	40.542	0.217	2.500	12.583
FOX News	13.119	32.530	0.750	4.033	15.283
The Guardian	12.936	33.954	0.700	4.200	14.367
USA Today	12.889	29.325	0.550	3.850	14.800
ABC News	11.124	32.475	0.550	2.633	11.633

Predicting article news source

Overall text

Classification accuracy in separating between pairs of publishers using article tokens

	USA Today	Huff. Post	Globe and Mail	ABC News	Time	CNN Europe	BBC	Daily Mail	The Guardian	Daily Telegraph	Wash. Post	The Indep.	FOX News	Sydney Morning Herald	The Moscow Times	DNA India	The Hindu
USA Today		87.6	93.4	84.0	85.4	89.9	97.4	95.6	94.9	96.5	86.0	96.0	91.8	98.4	97.7	96.1	97.1
Huffington Post	87.6		91.9	84.2	78.1	87.0	96.3	93.1	91.5	93.8	82.7	92.6	90.3	96.5	97.4	93.8	96.0
Globe and Mail	93.4	91.9		90.5	90.9	91.2	94.7	92.5	91.9	93.1	92.5	91.9	93.9	97.7	96.9	93.4	93.5
ABC News	84.0	84.2	90.5		80.5	84.5	96.1	92.7	92.5	95.1	83.7	94.9	86.5	97.5	95.9	94.0	94.8
Time	85.4	78.1	90.9	80.5		84.3	95.3	93.2	91.6	94.4	82.7	93.0	88.0	97.3	96.4	93.0	95.5
CNN Europe	89.9	87.0	91.2	84.5	84.3		94.8	91.4	91.7	94.5	86.6	92.4	92.5	97.9	96.3	93.5	95.1
BBC	97.4	96.3	94.7	96.1	95.3	94.8		87.6	83.3	82.4	96.9	84.3	97.5	97.8	98.3	93.1	94.0
Daily Mail	95.6	93.1	92.5	92.7	93.2	91.4	87.6		85.2	79.4	94.6	77.7	95.4	97.1	98.1	91.8	94.3
The Guardian	94.9	91.5	91.9	92.5	91.6	91.7	83.3	85.2		76.5	93.5	76.1	95.1	95.7	97.7	90.7	93.3
Daily Telegraph	96.5	93.8	93.1	95.1	94.4	94.5	82.4	79.4	76.5		95.9	66.9	96.9	96.7	97.6	91.6	93.4
Washington Post	86.0	82.7	92.5	83.7	82.7	86.6	96.9	94.6	93.5	95.9		94.9	90.6	98.1	96.8	95.2	96.1
The Independent	96.0	92.6	91.9	94.9	93.0	92.4	84.3	77.7	76.1	66.9	94.9		96.5	96.4	98.1	91.1	93.7
FOX News	91.8	90.3	93.9	86.5	88.0	92.5	97.5	95.4	95.1	96.9	90.6	96.5		96.3	97.8	96.2	96.9
Sydney Morning Herald	98.4	96.5	97.7	97.5	97.3	97.9	97.8	97.1	95.7	96.7	98.1	96.4	96.3		98.7	97.1	97.8
The Moscow Times	97.7	97.4	96.9	95.9	96.4	96.3	98.3	98.1	97.7	97.6	96.8	98.1	97.8	98.7		97.4	98.3
DNA India	96.1	93.8	93.4	94.0	93.0	93.5	93.1	91.8	90.7	91.6	95.2	91.1	96.2	97.1	97.4		82.8
The Hindu	97.1	96.0	93.5	94.8	95.5	95.1	94.0	94.3	93.3	93.4	96.1	93.7	96.9	97.8	98.3	82.8	

Ukraine-Russia test

Classification accuracy in separating between the selected set of news sources on topics related to the Ukraine-Russia incident

	USA Today	Huffington Post	Washington Post	CNN Europe	BBC	The Guardian	Sydney Morning Herald	The Moscow Times	DNA India
USA Today		83.7	70.1	85.1	91.7	84.5	92.9	84.2	89.7
Huffington Post	83.7		76.2	86.9	90.9	85.8	93.5	85.1	79.7
Washington Post	70.1	76.2		88.1	96.1	91.9	95.0	87.0	87.0
CNN Europe	85.1	86.9	88.1		90.4	84.8	91.7	88.2	86.0
BBC	91.7	90.9	96.1	90.4		83.3	88.4	85.6	88.8
The Guardian	84.5	85.8	91.9	84.8	83.3		91.5	87.4	77.6
Sydney Morning Herald	92.9	93.5	95.0	91.7	88.4	91.5		95.1	88.0
The Moscow Times	84.2	85.1	87.0	88.2	85.6	87.4	95.1		74.5
DNA India	89.7	79.7	87.0	86.0	88.8	77.6	88.0	74.5	

Similarity between publishers in events they report about

The table below shows the cosine similarity between pairs of publishers. The similarities are computed on the events that the publishers report about.

	USA Today	Huffington Post	Globe and Mail	ABC News	Time	CNN Europe	BBC	Daily Mail	The Guardian	Daily Telegraph	Washington Post	The Independent	FOX News	Sydney Morning Herald	The Moscow Times	DNA India	The Hindu
USA Today	0.00	0.13	0.11	0.14	0.10	0.11	0.08	0.12	0.11	0.08	0.16	0.09	0.13	0.07	0.04	0.05	0.04
Huffington Post	0.13	0.00	0.09	0.09	0.11	0.10	0.08	0.16	0.12	0.09	0.14	0.10	0.11	0.06	0.04	0.06	0.05
Globe and Mail	0.11	0.09	0.00	0.08	0.07	0.07	0.09	0.11	0.11	0.10	0.10	0.10	0.07	0.08	0.06	0.07	0.05
ABC News	0.14	0.09	0.08	0.00	0.07	0.09	0.07	0.10	0.09	0.07	0.15	0.07	0.17	0.05	0.04	0.04	0.04
Time	0.10	0.11	0.07	0.07	0.00	0.09	0.07	0.09	0.08	0.07	0.10	0.08	0.09	0.05	0.04	0.04	0.03
CNN Europe	0.11	0.10	0.07	0.09	0.09	0.00	0.10	0.11	0.11	0.10	0.12	0.10	0.09	0.07	0.04	0.06	0.05
BBC	0.08	0.08	0.09	0.07	0.07	0.10	0.00	0.26	0.26	0.26	0.09	0.24	0.06	0.10	0.06	0.10	0.08
Daily Mail	0.12	0.16	0.11	0.10	0.09	0.11	0.26	0.00	0.27	0.28	0.10	0.29	0.12	0.12	0.04	0.10	0.07
The Guardian	0.11	0.12	0.11	0.09	0.08	0.11	0.26	0.27	0.00	0.29	0.12	0.29	0.08	0.15	0.06	0.10	0.08
Daily Telegraph	0.08	0.09	0.10	0.07	0.07	0.10	0.26	0.28	0.29	0.00	0.09	0.30	0.06	0.11	0.05	0.09	0.07
Washington Post	0.16	0.14	0.10	0.15	0.10	0.12	0.09	0.10	0.12	0.09	0.00	0.09	0.14	0.07	0.06	0.05	0.06
The Independent	0.09	0.10	0.10	0.07	0.08	0.10	0.24	0.29	0.29	0.30	0.09	0.00	0.06	0.10	0.05	0.09	0.07
FOX News	0.13	0.11	0.07	0.17	0.09	0.09	0.06	0.12	0.08	0.06	0.14	0.06	0.00	0.04	0.03	0.04	0.04
Sydney Morning Herald	0.07	0.06	0.08	0.05	0.05	0.07	0.10	0.12	0.15	0.11	0.07	0.10	0.04	0.00	0.04	0.07	0.05
The Moscow Times	0.04	0.04	0.06	0.04	0.04	0.04	0.06	0.04	0.06	0.05	0.06	0.05	0.03	0.04	0.00	0.04	0.03
DNA India	0.05	0.06	0.07	0.04	0.04	0.06	0.10	0.10	0.10	0.09	0.05	0.09	0.04	0.07	0.04	0.00	0.27
The Hindu	0.04	0.05	0.05	0.04	0.03	0.05	0.08	0.07	0.08	0.07	0.06	0.07	0.04	0.05	0.03	0.27	0.00