A Computational Analysis of News Media Bias: A South African Case Study

Laurenz A Cornelissen
Computational Social Science Group
Centre for AI Research
Department of Information Science
Stellenbosch University
aldu cornelissen@sun.ac.za

Lucia I Daly
Computational Social Science Group
Department of Information Science
Stellenbosch University
lucia.daly@icloud.com

Qhama Sinandile
Computational Social Science Group
Department of Information Science
Stellenbosch University
qhamasinandile@gmai.com

Heinrich de Lange
Computational Social Science Group
Department of Information Science
Stellenbosch University
18311792@sun.ac.za

Richard J Barnett
Computational Social Science Group
Centre for AI Research
Department of Information Science
Stellenbosch University
barnettrj@acm.org

ABSTRACT

News media in South Africa is assumed to be unbiased and objective in their reporting of the news. Indeed, editors are required to uphold an objective and balanced view with no favour to external political or corporate interests. This assumption of objectivity is tested on a large scale by computationally analysing 30 000 articles published by five media houses: News24, SABC, EWN, ENCA, and IOL. Using topic modelling, 38 topics are extracted from the corpus, and sentiment is computed for each topic. The study highlights various cases of both over and under-reporting by media houses on particular topics. We also identify various tonality biases by media houses.

CCS CONCEPTS

• **Applied computing** → **Publishing**; **Document analysis**; *Annotation*; Sociology; • **Information systems** → *Data cleaning*; • **Social and professional topics** → Political speech.

KEYWORDS

news, media bias, NLP, topic modelling, south-african politics

ACM Reference Format:
Laurenz A Cornelissen, Lucia I Daly, Qhama Sinandile, Heinrich de Lange, and Richard J Barnett. 2019. A Computational Analysis of News Media Bias: A South African Case Study. In *Conference of the South African Institute of Computer Scientists and Information Technologists 2019 (SAICSIT '19), September 17–18, 2019, Skukuza, South Africa.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3351108.3351134

SAICSIT '19, September 17–18, 2019, Skukuza, South Africa

ACM ISBN 978-1-4503-7265-7/19/09…$15.00
https://doi.org/10.1145/3351108.3351134

1 INTRODUCTION

Origgi [28] reminds us of a key knowledge paradox: 'the greater the amount of information that circulates, the more people rely on so-called reputational devices to evaluate it'. Origgi's motivation for the new reputational age as the successor of the information age is difficult to dismiss if considering the recent political black swans in the United States (the election of Donald Trump) and Europe (the United Kingdom voting to leave the European Union). With the hyper-connectivity brought on by new information communication technologies, people are more reliant on reputational devices to aid not just their decision-making but broader social epistemological deduction. One such reputation device is news media [13]. Traditional news media are professionalised organisations that render a service to society by finding, digesting and broadcasting important information to the majority of people. Consider the definition of *social epistemology* in the journalistic context by Godler et al. [13]: 'social epistemology studies the acquisition of beliefs from the testimony of others.' The general public trusts traditional news media to interpret and communicate information fairly.

Post-apartheid South Africa has a strong tradition of non-partisanship, particularly when compared to more partisan traditions, for instance in the United States. The voluntary governing body, the South African Press Council [see 35], highlights the non-partisanship of news coverage. Nevertheless, South African news media is still regarded with distrust according to the annual trust barometer published by Edelman [10]. Globally, media is the least trusted institution. However, in South Africa the government is the least trusted of all institutions, and is also the least trusted when compared to all other governments—despite a seven-point increase from 2018.

In 2013, South Africans witnessed the first conspicuous partisan news media house since the first democratic elections in 1994. At that time, a controversial family launched a media house which was strongly aligned with the ruling party. This consisted of both broadcast (ANN7) and print media (The New Age) outlets. Both outlets were sold in 2018 and were re-branded, but are no longer producing content. This illustrates the non-partisan tradition in

South African media since despite being aligned with the majority party in the country, it was not well received by the public.

It would be naïve to imply that South African news media is free of biased reporting or political agenda setting. However, it is less obvious than, for instance, media in the United States, where it is generally acceptable to openly align to a political party. The objective of this study is, therefore, to investigate possible latent biases in the coverage of salient topics in South African politics by the largest media houses in South Africa. Of particular interest is finding differences in reporting frequency on certain topics, additionally differences in sentiment towards particular topics. The next section is a review of previous research related to the topics of this study, while also highlighting prominent alternative options for future research.

2 PREVIOUS WORK

Investigating news media bias is not a new problem. The question has been asked in earnest since 1976 when Hofstetter investigated partisanship influence over network television news coverage [15]. The question is again gaining popularity, arguably due to two events. The 2016 US presidential election and the British referendum to leave the European Union. The difference this time is that the focus is not on television, radio, or print, but on electronic news media, particularly social media. This project focuses on electronic news media, specifically in the context of South Africa. The domination of internet mediated news enables researchers to computationally analyse news reporting on a much larger scale than previously. As an example, Faris et al. [11] published a report on a large scale analysis of news content on the 2016 US presidential elections. They clearly illustrate the dominance of reporting on news related to the Trump campaign. This study attempts a first such computational investigation on South African news articles.

There are many nuances in analysing partisanship in news coverage. A clear definition of 'bias' is required before proceeding. The next section settles on a practical working definition to analyse partisanship in South African media.

2.1 Media Bias

Instead of untangling the concept of bias through an exhaustive exposition, it is more prudent to estimate a working definition which will enable the systematic uncovering of bias among media houses. A helpful definition of bias is a predisposition to or disproportionate favour for, or against, an object. Where an object can take the form of things such as people, groups, events, languages, or ideas. Bias is not inherently negative, indeed, it can come in the form of a helpful heuristic, or simply a healthy bias towards avoiding the flu.

To narrow the conception of bias to the context of media bias, Eberl et al. [9] highlights two key forms of media bias, *visibility bias* and *tonality bias*. Visibility bias is the frequency of reporting on a particular topic which is either systematically low or high. In the case of Eberl et al. [9] the visibility bias was particularly studied in terms of coverage of political agents, such as prominent party members. Tonality bias is the systematic positive or negative slant in reporting on particular topics or concepts. Previous research on media bias, particularly in South Africa used manually gathered and annotated news content [see 37]. In contrast, a more computationally effective means is proposed in this paper.

Before concluding, it is important to highlight the distinction between bias and accuracy. For example, an article may have a negative tone when reporting on crime, death, and other objectively negative topics. This sustained negative tone does not, necessarily, reveal anything about the media house's engagement with the topic, but rather that the topic itself is inherently negative. Moreover, some media houses may have financial articles, while others do not. This will result in apparent under and over-reporting on these topics. However, this apparent bias is simply a sampling error. It is, therefore, important to remain cognisant of the nuances of news topics to navigate such pitfalls in this methodology.

To summarise, bias is not inherently a negative trait, it is an observation of a pattern caused by complex human behaviour. A bias measurement is not a value judgement, but rather an observation of a systematic deviation from an observed norm. Uncovering such a norm, and observing any systematic deviations from it would suggest that bias is therefore in play.

The remainder of this section expands on the computational aspects of this study. First, the literature on the possible approaches to uncovering topics from large corpora is reviewed; before investigating the available approaches to efficiently analyse sentiment on a large number of documents.

2.2 Topic Extraction

To measure visibility bias, three areas require consideration. First, each document by a media house should be labelled with one or more appropriate topics. Second, the number of times that a media house produced a document on a particular topic must be recorded. Finally, the frequencies of topic engagement can be compared between media houses to derive the average engagement on a topic, which should highlight when a media house over or under engages with a topic. The second and third areas are matters of the routine process: counting and comparison using rudimentary statistics. The first, however—labelling an article with an appropriate topic—is a more difficult exercise and will, therefore, receive the bulk of the attention here.

There are multiple ways to assign topics to documents. For instance, a simple way would be to search all documents for the occurrences of a known entity such as a political figure, public event or political party. Recording the frequency of such a topic for each across the corpora of the media houses would give an indication of differences in the engagement, and therefore visibility, of the topic generated by media houses. The problem with such an approach is that it assumes, *a priori*, which topics are important. This approach, therefore, introduces another level of bias by the researcher. An alternative approach is to assume that there is no way of knowing what information is contained in a body of text, and attempt to extract meaningful information from the text itself. The field concerned with this approach is called information retrieval. The idea of automating the extraction of information from documents was introduced as early as 1958 by Luhn [19], who suggested that weights are assigned to terms and sentences based on their frequency and other statistical information. It is more difficult than simply extracting often recurring words in a corpus, because

some words, for instance 'the', appear disproportionately more than others. This is known as Zipf's law [41], and words of this nature do not offer much information. Moreover, there may be latent topics which are only identifiable on a wider scale. For example, consider the word 'party', which changes meaning depending on context, when talking about politics, it would most probably refer to a 'political party', whereas if it is an article on entertainment it might refer to an 'after party'. This requires an approach that would take the wider context of a term into account when attempting to extract information from the term. Simple frequencies would therefore not suffice to extract meaningful topics from the corpus.

For this reason, there are computational methods that aim to extract latent topics from large text corpora. Automated information retrieval from large corpora can be divided into supervised and unsupervised approaches. Early approaches tended to make use of unsupervised techniques. With the availability of more labelled data in recent years, there is an increase in supervised approaches. Supervised methods make use of a manually, or crowd-sourced, labelled data set as input data. The result of which would then offer a training data-set for a supervised machine learning approach. For this study, an unsupervised approach is implemented since the data is not labelled. Innovative unsupervised topic extraction is increasingly popular to automatically, effectively and efficiently mine topics from large quantities of data [4].

In 1998, Papadimitriou et al. [29] developed an early topic approach called *latent semantic indexing* (LSI). The method projects the document term matrix to a lower dimensional vector space.[1] From this, it is possible to uncover key correlations between terms, thus extracting latent meaning in the corpus. Figure 1 illustrates a basic framework for a DTM. Matrices that are similar to the DTM, for instance, a term co-occurrence matrix (TCM), are central to many natural language processing (NLP) tasks since it is a simple and efficient manner to encode text data in a structured way. Hofmann [14] provided yet another improvement through a method called probabilistic latent semantic analysis (PLSA). Arguably, the most popular method is an extension of PLSA, which is called latent Dirichlet allocation (LDA), developed by Blei et al. [5]. Most subsequent advances are mostly elaborations of LDA, for instance, structural topic models (STM) by Roberts et al. [33]. A promising recent advancement in unsupervised topic modelling uses a stochastic block modelling to extract topics [12].

Another approach is to use clustering algorithms to extract topics [40]. These approaches make use of word embeddings, which is a word representation in vector space [22]. Instead of relying on a DTM, these approaches work on the n-gram or skip-gram model, where a sliding window is applied to the text to define the co-occurrence of terms in a corpus. Approaches, like Zhang et al. [40] or Wartena and Brussee [36], then use these representations to apply unsupervised clustering algorithms to extract topics from the text.

The challenge with such clustering methods is that it is difficult to relate a topic back to a specific piece of text, except if a separate model is developed for each document. Since the key objective is to relate a topic to a specific document, topic modelling is used.

[1] A document term matrix (DTM) is a matrix containing a document per row, and term per column, with frequencies of occurrences of a term in a document in the corresponding cell.

$$\mathbf{DTM} = \begin{array}{c} \begin{matrix} T_1 & T_2 & \cdots & T_t \end{matrix} \\ \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1t} \\ d_{21} & d_{22} & \dots & d_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nt} \end{bmatrix} \end{array} \begin{matrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{matrix}$$

d_{ij} = number of occurrences of T_j in D_i

Figure 1: Document term matrix (DTM).

2.3 Sentiment analysis

In a manner similar to how topic extraction can be used to assess visibility bias, tonality bias can be assessed computationally to scale to increasingly larger corpora of text. Instead of manually annotating news articles for the polarity of sentiment, it is possible to automate the measurement of sentiment in text in various ways.

As with topic extraction, there are supervised and unsupervised approaches to sentiment analysis. However, the data is not annotated with sentiment, which leaves only dictionary and pre-trained models as reasonable options. A dictionary approach is simply matching words with a predetermined dictionary of sentiment [*bad* = -1/negative] or [*good* = 1/ positive], and simply aggregating the retrieved sentiment to the required level of text. A pre-trained model approach uses a pre-trained language model provided by CoreNLP [20], Monkeylearn [23], etc. to infer the sentiment of the text. With preliminary attempts, it was found that the dictionary approach offers better results, as opposed to a pre-trained model. This is due to two factors. *First*, all available pre-trained models are trained in particular domains and do not perform well outside of those domains [3]. *Secondly*, although the methods offer polarity for the text, it does not provide the strength of sentiment. Naldi [26] reviewed prominent lexicon approaches to sentiment analysis that are available in R. From these methods, sentimentR from Rinker [31] is found to be superior, due to its incorporation of valence shifting on a sentence level.

The methodology for sentimentR moves beyond simple aggregation across the text from a bag of annotated words and offers a more nuanced calculation of sentiment which take into account phrases and parts of speech to determine a score for the given text. Valence shifters can be categorised as negators (ex. *not*, *yet*), amplifiers (ex. *very*, *much*), de-amplifiers (ex. *little*, *barely*), adversative conjunction (ex. *but*, *however*). Valence shifters can potentially change the polarity of other words. For example, adversative conjunctions such as *not* and *however* can reverse the initial sentiment of the statement. Consider the sentence: *He believed in his people, but he did not agree with their methods*. The first part of the sentence *He believed in his people* is positive, but the adversative conjunction *but* inverts it with the sentiment of the rest of the sentence, which is negative. A negator example would be *not great* which lowers the sentiment of *great* to -0.354 [17].

3 METHODOLOGY

To computationally analyse news articles produced by media houses in South Africa, requires a dataset of the text of articles produced by each media house. Given this text, the sentiment of each article can be computed, i.e., positive or negative. For each article, it is possible to extract topics from the text. These extracted topics can be linked back to each article where it inherits the sentiment of the article.

Five media houses were selected for this study: Eye Witness News (EWN)[2], News24[3], the South African Broadcasting Corporation (SABC)[4], Independent Online (IOL)[5], and eNews Channel Africa (ENCA)[6]. These are the largest media houses in South Africa, all of which have a strong electronic footprint, along with multi-platform outlets like radio, television, and periodicals. The SABC is the only state-owned media house but is governed by the same standards as the others.

There is no standard repository or data archive for these websites, leaving crawling and scraping as the only viable alternatives to collecting the data. The following section outlines how the data was collected for each site.

3.1 Data Collection

There is no effective method for building a scraper that extracts clean text. Text that is scraped from web pages is often filled with non-textual artefacts, for example JavaScript code, *iframes*, and embedded social media elements. This makes the cleaning process very difficult, especially if non-textual artefacts are difficult to find in a large corpus. An iterative approach was used to scrape the text to avoid building a labour intensive cleaner that removes non-textual artefacts.

The URLs for each article was collected using Octoparse [7] which automated the process. The search terms used to collect the appropriate articles comprised of *ANC*, *COPE*, *DA*, *EFF* and *FF*.

Each URL returned an article which was scraped with the `Rvest` package [38] in R. The HTML elements within the raw text of the article were identified for each media house and used to extract the text of the article itself.

The task was performed on a sample of the collection of URLs per media house so as to implement tokenisation and remove stopwords from the text. This technique enabled the extraction of sentences that occurred frequently within each of the articles. The specific names of the HTML containers could then be identified to ignore certain sentences, which resulted in a *blacklist* of non-textual artefacts.

Once the list HTML elements were sufficient, it could be removed in the scraping process to see if the tokenisation on the filtered sentences returned clean text. The scraper was only run on the full list of URLs, once the sample text was clean.

[2] ewn.co.za
[3] www.news24.com
[4] www.sabc.co.za/sabc
[5] www.iol.co.za
[6] www.enca.com
[7] www.octoparse.com

3.2 Text Cleaning and Pre-processing

After the data for each media house was processed separately, all the articles were added to a single database for further text cleaning. This involved removing all non-ASCII characters, after which all contractions are converted (ex. *isn't* into *is not*), replacing URLs, since they do not offer any information for topic modelling or sentiment analysis. It was found that there are many cases where a full-stops and commas were not followed by a space, which could cause problems with processing functions. All extra white spaces remaining after removing large HTML sections were also removed. All apostrophes following nouns, such as "Africa's", since this would negatively impact topic modelling.[8]

In order to find a systematic manner to investigate the structure of the data, each article was tokenised into sentences and calculated the number of words and characters per sentence. A scatter-plot of these two variables reveals the strong linear relationship between word count and character length (see Figure 2). Any noticeable deviation from this linear relationship would highlight potential unwanted artefacts in the text. A threshold was set at 710 character length and 112-word count, any sentence above both these thresholds will be removed from the corpus. A linear model was applied to the two variables to extract the residuals. These residuals will highlight words that are unnaturally long (usually URLs or script tags), or unnaturally short. Observations were only retained if it had a residual value within six times the interquartile range (IQR) of the median residual value for the dataset. This procedure removed 4 853 sentences from the corpus (from 486 412 to 481 559 sentences). The number of documents from which these sentences were drawn is only 15. The removal of these documents will, therefore, have a minimal impact on the total corpus by reducing the corpus to 30 828 documents from 30 843.

4 DATA ANALYSIS

With the data appropriately cleaned and pre-processed, each article is analysed for two things: sentiment and topics. The following sections outline the procedures followed to do this and conclude by specifying the method to relate the sentiment of the document with the associated topics.

4.1 Sentiment Analysis

In a recent review of lexicon-based sentiment packages, Naldi [26] suggested that sentimentR from Rinker [31] outperforms the other available packages. Although no sentiment analysis method is perfect, and even sentimentR has key flaws, early testing revealed that it outperforms pre-trained model based sentiment analysis procedures including CoreNLP [see 1]. The procedure additionally offers both polarity and strength of sentiment, which is not returned by CoreNLP.

The key benefit of using sentimentR is that it analyses sentiment on a sentence level and not word level as with most dictionary approaches. It incorporates valence shifters to account for linguistic features which might change the sentiment of the text. This section will provide more detail on the actual procedure followed for this study.

[8] We made extensive use of the textclean package by Rinker [32].

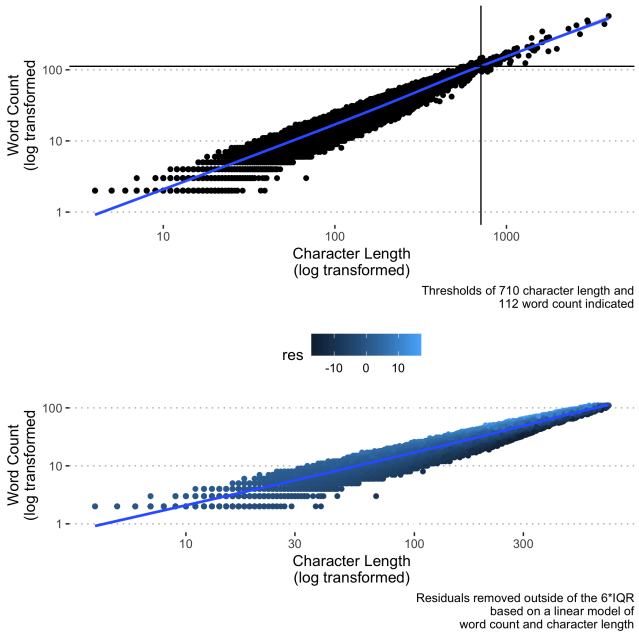

Figure 2: Word count and character length correlation.

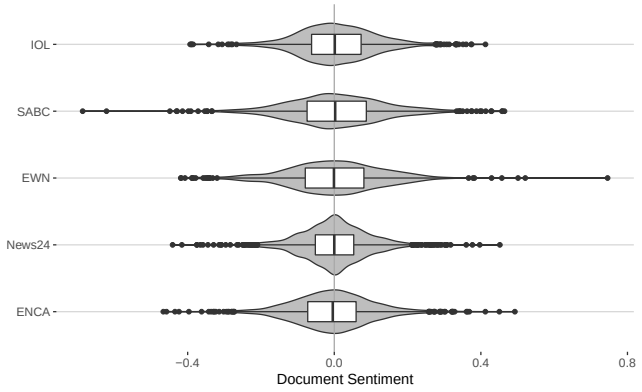

Figure 3: Sentiment density plot.

Table 1: Sentiment per media house.

MH	mean	median	min	max	sd	n
ENCA	-0.001	-0.001	-0.466	0.493	0.107	4334
EWN	0.006	0.004	-0.419	0.746	0.123	3034
IOL	0.011	0.006	-0.393	0.413	0.105	3730
News24	0.006	0.005	-0.441	0.452	0.093	4537
SABC	0.011	0.006	-0.687	0.464	0.126	3132

Firstly, each article was tokenised into sentences by utilising the tokenizers package [25]. The standard implementation did not identify all sentences appropriately since there are still many artefacts left from the scraping process. Sentence breaks were forced in sentences that were longer than 40 words by replacing any commas, semi-colons, and colons into a '. ' sequence, which can be identified by the tokeniser as the end of a sentence.

At this point, the text is sufficiently clean to apply the sentimentR tokenisation procedure. The jockers-rinker lexicon was used, which is an extension of Hu and Liu [16] and Jockers [17]. Furthermore, the valence shifters dictionary from the lexicon package [30] was used to provide the appropriate valence shifters to the sentimentR procedure. An amplifier weight of 2 was assigned since sentiment tends to converge around neutral sentiment when aggregated. The window for valence shifters around polarised words was set to 3 before and 3 after. The question weight was set to zero, which nullifies a question sentence with polarity. This is because the text is produced by journalists, who might make frequent use of evaluative and rhetorical questions. Lastly, the neutral non-verb word "like" was detected. This helps to distinguish phrases such as "I like you" from "I am like you", where the first produces a distinct positive sentiment, whereas the second is more neutral due to the preceding word "am".

Table 1 provides the summary statistics of the results of the sentiment analysis process. Sentence level sentiment was aggregated up to the document level by using the averaging function in the sentimentR package. The function down-weights neutral sentences, while slightly up-weighting negative sentences in the aggregation.

From Table 1 it is clear that sentiment tends towards neutral, with a normal distribution. The most negative article in the corpus is by the SABC, which is a short report of an accident where four were killed and 13 others seriously injured. The most positive article is by EWN about students getting ready for the matriculation examination results in 2018. Consider Figure 3, where the spread for each media house is more evident.

There is a clear spread of either positive or negative documents, but overall they are generally neutral. It is, nevertheless interesting, to investigate what the more negative and positive articles are. It may be that all negative articles are of car accidents, and all positive articles are of wholesome matriculate stories. However, if the theme of all the documents can be identified, it is possible to confirm that negative articles are purely due to negative topics, and positive articles are of inherently positive topics. Another benefit would be to see if the coverage of a topic by a media house fits with the overall sentiment of the topic, or if there is a polarity slant. The next section will outline how topic-modelling was used to extract topics from articles in the corpus.

4.2 Topic Modelling

Three topic modelling approaches were investigated for this study. First, Latent Dirichlet Allocation (LDA) was selected since it is well studied and offers reasonable results in a context where there are many relatively long documents in a corpus. LDA offers results that are easier to interpret and can be related back to specific articles in an easy process. The second consideration was to use a recent application of stochastic block modelling (SBM) to the topic extraction problem by Gerlach et al. [12]. Finally, the third option was to build a global vector (GloVe) model (word embedding model) on which a clustering procedure can be applied to extract topics. The results of all three procedures were reviewed and LDA consistently to gave the most coherent results. The comparison between

the three approaches would nevertheless be a fruitful exercise to systematically explore, but that is outside the scope of this study.

The LDA approach is a method that weighs the relevance of a particular term, usually a word, to a topic (measured by the beta value) and the relevance of a topic to a document (measured by gamma) which is modelled as Dirichlet distributions [5]. LDA is a fuzzy clustering technique which suggests that each data point can belong to more than one cluster. To contrast, a hard clustering technique is one where there are non-overlapping clusters and each instance belongs to one and only one cluster, for example k-means clustering. These fuzzy memberships provide a more varied technique of recommending similar items and finding duplicates. Strict clustering would have restricted the number of topics that could be extracted. LDA provides an output that is easy to interpret and thus makes categorising topics a much simpler job. The central drawback of such an approach is that all documents are related to all topics, which are related to all terms. The researcher must therefore take care on the thresholds and how to handle this limitation of the method.

Topic modelling requires extra pre-processing steps. *First*, the text must be tokenised into single terms. This process involves breaking the text into distinct meaningful parts ("tokens"). To do this, punctuation marks and stop-words were removed from raw text. The list of stop-words were retrieved from the Snowball package and the SMART package which are predefined words that do not contribute towards the topic. Words with less than 3 characters were also removed. Following the findings by Lau et al. [18] and Martin and Johnson [21], tokens were lemmatised and only nouns were retained. The udpipe package performs all these steps [34, 39]. The process first splits the text into single word tokens, then lemmatises the words to a common root word, and lastly tags each word with an appropriate part of speech tag using the CoNLL-U Format of language specific tags (XPOS).[9] Using these XPOS tags, only certain noun and adjective types are retained to improve the performance of the method [see 21]. Specifically, only words tagged with `NN` (noun, singular or mass), `NNS` (noun, plural), `NNP` (proper noun, singular), `NNPS` (proper noun, plural) and `JJ` (adjective) was retained.

One of the limitations of LDA is that it is a soft clustering method, as such it difficult to gauge the topic quality and coherence. There is no objective metric to determine the optimal choice of hyper-parameters. LDA requires a fixed K, which means the analyst is required to define the number of topics beforehand. In order to judge the quality of the results, it necessitates an iterative process to be performed with domain knowledge on the topics.

The length of documents plays a crucial role: poor performance of the LDA is expected when documents are too short, even if there is a very large number of them. Ideally, the documents need to be sufficiently long, but need not be too long. In practice, for very long documents, one can sample a fraction of each document and the LDA still yields comparable topics [24]. When a significantly larger number of topics than needed are used to fit the LDA, the statistical inference may become inescapably inefficient.

In theory, the convergence rate deteriorates quickly to a non-parametric rate, depending on the number of topics used to fit the LDA. This means that the researcher has to be very meticulous in selecting K, the number of topics, and avoid overestimating the number [24].

[9] See https://universaldependencies.org/format.html.

Table 2: LDA topic model examples.

Document ID	Gamma	Topic ID	Beta	Term
14196	0.982	9	0.032	union
14196	0.982	9	0.030	south
14196	0.982	9	0.029	cosatu
7055	0.959	16	0.120	land
7055	0.959	16	0.025	compensation
7055	0.959	16	0.022	expropriation
16939	0.986	35	0.110	eff
16939	0.986	35	0.077	malema
16939	0.986	35	0.027	party

Although there is no definitive measure of determining K, there are attempts at informing a reasonable estimation of K. Three metrics were utilised to inform the optimal number of topics to be extracted: topic density by Cao et al. [6] and divergence measures by Arun R. [2] and Deveaud et al. [8]. For convenience, these measurements are implemented in a package called ldatuning by Nikita [27]. The Deveaud et al. [8] metric indicate a clear peak between 37 and 38 topics. When considering the Cao et al. [7] and Arun R. [2] metrics both minimise at 38 topics. The appropriate number of topics were specified at 38.

The LDA returns a result which specifies a beta value for each word to a topic, as well as a gamma for each topic to a document. The beta values measure how much a particular term contributes to a topic. As such, each topic can consist of multiple words with varying levels of relevance to the topic. Each document could also contain multiple topics, therefore the gamma values indicates the relevance of the topic to the document. Table 2 is an extract from the results. Three topics, 9, 16, and 35 are shown with the top three terms associated with each topic. Topic 9 has the terms *union*, *south*, and *cosatu*, which suggests that the topic is about trade unions. Similarly, topic 16 is associated with *land*, *compensation*, and *expropriation*, which allude to that the topic is on the highly debated land expropriation topic. Lastly, topic 35 is associated with *eff*, *malema*, and *party*, which connotes a topic on the EFF political party. Each of these topics are associated with three documents from the corpus, and the gamma measurement indicates the strength of the association. Table 2 only shows the top topic for each document, however, a document might be on multiple topics. For this reason, the top three topics per document were retained in the dataset.

With the topics extracted for each document, it is possible to relate the sentiment of the document to a topic. The next section explains how this is done.

4.3 Topic Sentiment

Since there is only one measurement of sentiment per document, and there are multiple topics per document, an overall sentiment can be measured towards topics by using the gamma value as a weighting. This would mean that a document that has three listed topics,[10] but the first has a gamma of 0.99, and the other two have

[10] Recall that only the top three topics per document are retained.

Table 3: Weighted sentiment calculation.

Doc.	Topic	Gamma	Doc. Sent.[1]	Gamma Norm.[2]	Wt. Topic Sent.[3]
16939	35	0.9859	-0.0287	0.9992	-0.0287
16939	22	0.0004	-0.0287	0.0004	0.0000
16939	32	0.0004	-0.0287	0.0004	0.0000

[1] Document sentiment.
[2] Normalised gamma: gamma divided by the sum of gamma of top three topics per document.
[3] Weighted Topic Sentiment.

less than 0.01, the document sentiment can be multiplied by the normalised gamma value, which would result in a low impact on the sentiment polarity for the two low gamma topics, while increasing the polarity of the sentiment of the appropriate topic. For instance, consider document 16939 shown in Table 3, the top three topics are 35, 22 and 32. However, only topic 35 has a strong measurement of gamma (0.9859), whereas the other two have a very low gamma (0.0004). Multiplying a normalised gamma value with the document sentiment will result in a weighted sentiment measure per topic. If the topic contributed little to a document, it will inherit none of the document sentiment, inversely, if the document is mostly on one topic, that topic will inherit the sentiment of the document almost directly.

Using the above calculation, it is then possible to measure sentiment towards each topic. Table 4 contains the descriptive statistics of the sentiment towards each topic. Each topic was labelled using knowledge on the domain, these labels are to aid with interpretation, but they are always subjective and not wholly descriptive of the complexities of each topic. The topic labels are nevertheless intuitive to individuals who are aware of South African news and politics.[11]

Some key observations are that the most negative topics are *USA*, *Life Esidimeni* and *EFF parliament.* USA, consists of articles reporting on mostly the President of the United States, Donald Trump. Life Esidimeni is a topic on the scandal surrounding the death of 143 people. The EFF party topic is reporting on the EFFs activities in government, which is usually controversial walkouts and disruptions. The most positive three topics are *Mandela*, *Elections* and *Economics*. Mandela is a topic on the first democratic era President of South Africa. Elections is a topic on election activities in South Africa.

At this point, it is possible to analyse the data to uncover possible signs of visibility and tonality bias. The next section presents this analysis.

5 RESULTS AND DISCUSSION

This section analyses the results of the study and assesses whether they reveal possible latent biases in the coverage of salient topics in South African politics by the chosen South African media houses. We discuss the particular differences in the frequency of reporting for certain topics as well as the polarity of sentiments towards particular topics.

[11] Certain topics were removed from the analysis, because they were not of interest (such as soccer), were not engaged with enough by media houses, or are overwhelmingly positive or negative (such as crime)

Table 4: Descriptive statistics of sentiment for topics.

Topic	mean	median	min	max	n
Mandela	0.019	0.005	-0.252	0.372	661
Elections	0.016	0.003	-0.188	0.294	1037
Economics	0.014	0.006	-0.208	0.289	1145
ANC NEC	0.009	0.002	-0.389	0.231	1425
Higher Edu	0.005	0.000	-0.264	0.274	744
Nkandla	0.005	0.000	-0.225	0.210	468
Captured	0.005	0.000	-0.127	0.309	488
ANC Leadership	0.005	0.002	-0.177	0.276	1716
Land	0.003	0.000	-0.136	0.246	462
Parliament	0.003	0.000	-0.204	0.258	942
Local Election	0.002	0.000	-0.159	0.219	640
Cope	0.001	0.000	-0.204	0.223	454
De Lille Scandal	0.000	0.000	-0.369	0.302	863
Tshwane	-0.001	0.000	-0.193	0.195	496
EFF	-0.001	-0.001	-0.302	0.228	967
VBS	-0.002	-0.001	-0.210	0.188	725
ANC NW	-0.002	0.000	-0.227	0.246	852
Fransman	-0.004	-0.001	-0.284	0.249	523
Unions	-0.004	-0.001	-0.223	0.177	572
MOC	-0.004	0.000	-0.374	0.173	545
State Capture Inquiry	-0.004	-0.001	-0.262	0.223	518
Zuma Gordhan	-0.006	-0.002	-0.186	0.147	664
Public Services	-0.007	-0.001	-0.351	0.226	434
EFF Parly	-0.007	-0.002	-0.255	0.159	632
Life Esidimeni	-0.017	-0.005	-0.345	0.254	446
USA	-0.017	-0.006	-0.423	0.204	372

Note: Based on gamma weighted sentiment.

To review, from Table 1, it is evident that all media houses have a relatively neutral sentiment across their articles. However, it was highlighted using Figure 3 that there is a considerable deviation from negative and positive articles. To investigate which topics might drive the negative or positive sentiment, three topics per document was extracted. This section will investigate the sentiment on these topics by each media house as well as the level of engagement with each topic by each media house. The first will inform on possible tonality bias, whereas the latter will highlight any possible visibility bias caused by over or under-reporting on certain topics. To aid this process, Figure 4 combines all necessary information for the analysis. The plot groups all topics per media house, and places them vertically to observe patterns across media houses. For each topic, two measurements are recorded.

First, Figure 4 reports the weighted topic sentiment on the x-axis. For each topic, the black line indicates the median value for the particular topic, as well as the upper and lower quartiles for the topic sentiment. Notice that the line is the same per media house grouping. The point indicates the median sentiment of the particular media house on the particular topic. The point is only plotted if the media house's sentiment is lower or higher than the upper and lower quartiles for the topic. This indicates that the media house is either over positive or over negative relative to the normal sentiment towards the topic. Consider for instance the economics topic, we can see that ENCA has lower than normal sentiment, and EWN has a higher than normal sentiment towards economics. Both, however, still are showing positive sentiment towards economics.

Second, to the right of the sentiment for each media house is the engagement frequency measures. The bar chart measures the percentage of articles on the topic that is engaged by the MH. This is calculated by taking the number of articles on the topic by the

Table 5: Top five under and over-engagements.

MH[1]	Topic	AOT/AOT|MH[2]	AOT[3]	AOT|MH[4]
Under Engagement				
News24	USA	5.630	373	21
IOL	USA	6.434	373	24
EWN	EFF	7.825	984	77
SABC	Nkandla	8.686	472	41
ENCA	Cope	8.972	457	41
Over Engagement				
IOL	Cope	56.018	457	256
ENCA	Nkandla	40.466	472	191
SABC	USA	39.410	373	147
News24	ANC NEC	36.798	1443	531
EWN	State Capture Inquiry	33.081	529	175

[1] Media House.
[2] Engagement measurement as used in 4.
[3] Total number of articles on topic = AOT.
[4] Total number of articles on topic by media house = AOT|MH.

media house, divided by the total number of articles on the topic. This measurement is meant to capture whether a media house over or under engage with a topic relative to the other topics they cover. The colours indicate whether the percentage is over or under the upper or lower quartiles of the media house's average number of engagements with all topics. For instance, highlighting Nkandla, we can see that ENCA over-engages with the topic and SABC seems to be under-reporting on the topic.

5.1 Engagement

The various media houses exhibited interesting patterns in reporting frequency as it pertains to particular topics. In analysing the potential visibility bias of the media houses we observe the following characteristics of the data. Table 5 provides a summary of the top 4 under and over engagements. The topics with the most over engagement are Cope (by IOL), Nkandla (by ENCA), USA (by SABC), ANC NEC by News24 and State Capture Inquiry (by EWN). The most under-reported topics are USA (by News24 and IOL), EFF (by EWN), Nkandla (by SABC) and Cope (by ENCA).

The observations worth elaborating on are that the SABC under-reports on the Nkandla scandal. Surveying the other topics on which SABC under-reports, highlights other notable topics: Captured, Cope, and EFF. Captured is a topic on ministers and individuals who are drawn into the state capture narrative, and are therefore known as 'captured' individuals. Both EFF and Cope are breakaway parties from the ANC government. Recall that the SABC is a state-owned media organisation. It could, therefore, be interpreted as under-reporting which displays a visibility bias. Contrasting these topics with the engagement from other media houses shows that cope is also under-reported on by ENCA and News24, while being highly over-reported by IOL. EFF is also under-reported by EWN but again over-reported by IOL. Lastly, Nkandla is also under-reported by EWN, and highly over-reported by ENCA. It is therefore difficult to outline an explanatory narrative of a systematic visibility bias, particularly in terms of under-reporting. The individual outliers are nevertheless interesting to observe and suggest that further research should be done to elaborate on this.

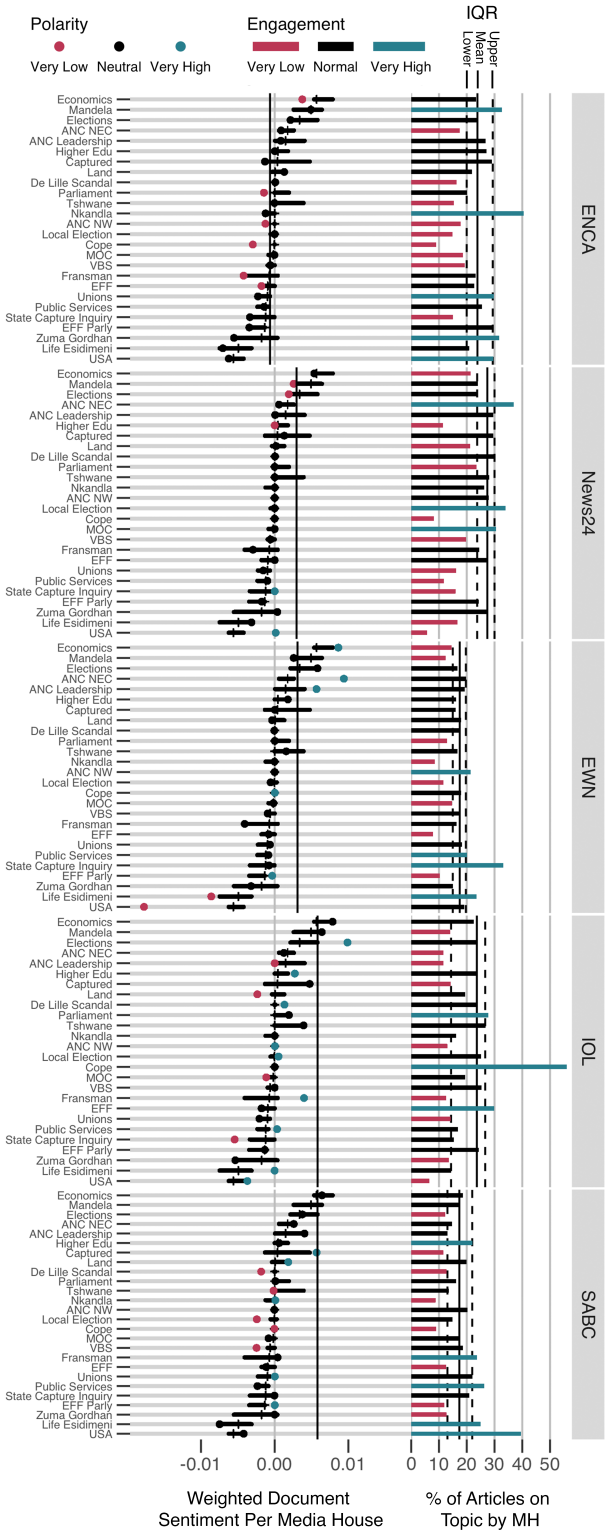

Figure 4: Relative sentiment and engagement plot.

Table 6: Top five polarised sentiment.

MH	Topic	Wt. Topic Sent.[1]	Q1[2]	Q3[3]
Negative Slant				
EWN	USA	-1.773	-0.623	-0.421
ENCA	Cope	-0.296	0.000	0.000
IOL	State Capture Inquiry	-0.543	-0.336	-0.003
IOL	Land	-0.235	-0.035	0.130
SABC	Local Election	-0.242	-0.055	0.026
Positive Slant				
SABC	Mandela	2.015	0.258	0.640
EWN	ANC NEC	0.939	0.059	0.259
IOL	Fransman	0.398	-0.407	0.040
IOL	Elections	0.988	0.215	0.580
IOL	Life Esidimeni	-0.001	-0.747	-0.315

Note: All values are multiplied by 100 for ease of interpretation.
[1] Weighted Topic Sentiment.
[2] Lower Quartile.
[3] Upper Quartile.

5.2 Polarity

Changing focus to the sentiment on topics by media houses in Figure 4, the instances where sentiment by a media house has either been overly positive or overly negative relative to the average sentiment of the topic can be highlighted. Table 6 offers a summary of the most notable observations. The topics with the most negative slant are USA (by EWN), Cope (by ENCA), State Capture Inquiry (by IOL), Land (by IOL), Local Election (by SABC). The most positive slants are Mandela (by SABC), ANC NEC (by EWN), Fransman (by IOL), Elections (by IOL), and Life Esidimeni (by IOL).

It is worth expanding on some more observations, particularly focusing on emergent indications of tonality bias by some media houses. Considering that the most negative media house is ENCA (only slightly lower than neutral), we can see that all slant is always negative, specifically for Economics, Parliament, ANC NW, Cope, Fransman and EFF. In contrast, the two most positive media houses, IOL and SABC have both positive and negative slants as indicated in Figure 4. Acknowledging that the overall sentiment of SABC is positively skewed, it is interesting to highlight the topics with negative slant, such as the de Lille scandal, Tshwane, Local Election, and VBS. These topics are interesting to observe because the de Lille scandal is a prominent scandal around the official opposition, the Democratic Alliance (DA), which caused much damage to the party's reputation. Tshwane surrounds the municipal news for the Tshwane municipality, which was lost from the ANC government to a minority rule government, lead by the DA. Lastly, VBS is a scandal surrounding the collapse of the VBS Mutual bank, which was mired in corruption allegations, mostly implicating prominent EFF members.

Again, a coherent explanatory narrative and definitive claims of tonality bias is difficult to justify. There are clear suggestions that there is sentiment slant and coverage bias, but more research is needed to more definitively forward such claims.

6 CONCLUSION

This paper was contextualised by a reminder of the importance of reputational devices such as news media in a modern hyper-connected society. The general public trusts traditional news media to interpret and communicate information fairly. Therefore, uncovering a potential bias in the content that is published can be of significant value to South African citizens as these media houses control and drive public discourse.

Bias was defined as a predisposition to or disproportionate favour for, or against, an object. Where an object can take the form of things such as people, groups, events, languages, or ideas. To narrow our conception of bias to within the context of media bias, borrowing from Eberl et al. [9], who highlight two key forms of media bias, visibility, and tonality bias. The assumption that news media in South Africa is unbiased and that editors uphold objectivity and a balanced view with no favour to the external political or corporate interests was tested. By computationally analysing the news articles of five media houses, the content could be scrutinised based on the frequency of engagement and sentiment of latent topics.

The analysis highlighted significant over and under-reporting on key topics. For example, the political party, COPE, was consistently under-reported on, even though it was the third largest party after the 2009 national elections, and dropped to the 8th largest in the 2014 local elections. Interestingly, the most positive topics are Economics, Mandela and Elections, with the most negative as USA, Life Esidimeni and Zuma-Gordhan. Investigating the most prominent examples of deviation by media houses highlights that the ANC NEC was the most well-covered topic in the corpus, but News24 still managed to over-report on the topic, and comparatively, ENCA under-reports on the topic. Both IOL and SABC have the highest overall positive sentiment, while still recording strong negative slants on multiple topics such as VBS, Local Election, de Lille scandal.

Overall, South African media houses show neutral sentiment when reporting on a wide array of topics. There are no alarming deviations found, except the few instances highlighted. South African media is known to be relatively bipartisan, and it should be no surprise that it is difficult to extract clear and systematic bias from the data. This suggests that South African editors, governing bodies and news consumers are highly attentive to biased reporting, and the examples of the now-defunct New Age and ANN7 outlets highlight the intolerance towards partisanship. To force a substantive conclusion on possible bias to lead to further investigation, it would be to highlight the negative sentiment on particular topics by IOL and SABC. It is worth highlighting the over-reporting of News24, EWN and ENCA on topics with anti-governing party topics.

6.1 Limitations and Suggestions

There are some key limitations to this project, which outline key lessons for future research. *First*, the data collection relied on a sub-optimal crawling and scraping procedure. The process will be much improved if a standardised library is supplied or created to capture news articles for South African media houses. Using search terms of *ANC*, *DA*, *COPE*, *FF*, and *COPE* as a means to capture the largest political parties as per the previous national election in 2009, was convenient, but not optimal. This might introduce a sampling bias in the data by not covering any other parties that are controversial or have grown since the 2009 elections. Moreover, by using these limited search terms, the spectrum of news topics

are drastically reduced, but that comes with the gain of ensuring political relevance of the topics.

Second, the analysis of sentiment could be much improved by utilising a supervised machine learning procedure trained on South African news articles. This mostly due to a limitation of scope in this project, and is therefore a key suggestion for future research in this domain.

Third, topic modelling can be improved, since it relies on a method, that is well explored, but also has fundamental issues due to the soft clustering approach. Future attempts might find better results by employing newer attempts at improving on LDA. An approach with word embedding models could lead to fruitful insights, but this was again beyond the scope of this project, and is a key suggestion for future projects.

Fourth, determining tonality and visibility bias should ideally rely on robust statistical tests to provide confidence to the inferences. This was not done, since it will add a layer of complexity on the interpretation that is beyond the scope of this article. Future studies will be much improved by designing an appropriate statistical test to define bias within this context.

The analysis of the media houses presented a challenge, because of the fact that no previous iterations or models built specifically for South African media. Thus it was important to gather and collect usable data from each media house and use the appropriate pre-processing methods to prepare the data in a usable format.

REFERENCES

[1] Taylor Arnold. 2017. A Tidy Data Model for Natural Language Processing using cleanNLP. , 20 pages. https://journal.r-project.org/archive/2017/RJ-2017-035/index.html

[2] Veni Madhavan C.E. Narasimha Murthy M.N Arun R., Suresh V. 2009. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. *Mycological Research* 113, 2 (2009), 207–221.

[3] Anthony Aue and Michael Gamon. 2005. Customizing Sentiment Classifiers to New Domains: A Case Study. https://www.researchgate.net/publication/215470666

[4] Bhagyashree Vyankatrao Barde and Anant Madhavrao Bainwad. 2017. An Overview of Topic Modeling Methods and Tools. (2017), 745–750.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

[6] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing* 72, 7-9 (2009), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

[7] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing* 72, 7-9 (2009), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

[8] Romain Deveaud, Eric SanJuan, and Patrice Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17, 1 (2014), 61–84. https://doi.org/10.3166/dn.17.1.61-84

[9] Jakob Moritz Eberl, Hajo G Boomgaarden, and Markus Wagner. 2015. One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences. *Communication Research* 44, 8 (2015), 1125–1148. https://doi.org/10.1177/0093650215614364

[10] Edelman. 2019. *19th Annual Edelman Trust Barometer: Global Report.* Technical Report. Edelman, Edelman2019. 66 pages. https://www.edelman.com/sites/g/files/aatuss191/files/2019-03/2019_Edelman_Trust_Barometer_Global_Report.pdf?utm_source=website&utm_medium=global_report&utm_campaign=downloads

[11] Robert M. Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. *Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election.* Technical Report. Berkman Klein Center for Internet & Society. 142 pages. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33759251

[12] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. A network approach to topic models. *Science Advances* 4, 7 (2018). https://doi.org/10.1126/sciadv.aaq1360

[13] Yigal Godler, Zvi Reich, and Boaz Miller. 2018. Social epistemology as a new paradigm for journalism and media studies. *New Media and Society* (2018), 1–18.

[14] Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI '99)*. Morgan Kaufmann Publishers Inc., Stockholm, Sweden, 289–296.

[15] C Richard Hofstetter. 1976. *Bias in The News: Network Television News Coverage of the 1972 Election Campaign.* Ohio State University Press.

[16] Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*. ACM, New York, NY, USA, 168–177. https://doi.org/10.1145/1014052.1014073

[17] Matthew L. Jockers. 2015. Extracts Sentiment and Sentiment-Derived Plot Arcs from Text. , 12 pages. https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf

[18] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves : Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 530–539.

[19] H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2, 2 (1958), 159–165. https://doi.org/10.1147/rd.22.0159

[20] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2015. The Stanford CoreNLP Natural Language Processing Toolkit. https://doi.org/10.3115/v1/p14-5010

[21] Fiona Martin and Mark Johnson. 2015. More Efficient Topic Modelling Through a Noun Only Approach. In *Proceedings of Australasian Language Technology Association Workshop*. 111–115.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781* (2013). http://arxiv.org/abs/1301.3781

[23] MonkeyLearn. 2018. Sentiment Analysis: Nearly Everything you Need to Know.

[24] Andrea Morandi, Marceau Limousin, Jack Sayers, Sunil R. Golwala, Nicole G. Czakon, Elena Pierpaoli, Eric Jullo, Johan Richard, and Silvia Ameglio. 2014. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. *JMLR: W&CP* (2014). https://doi.org/10.1111/j.1365-2966.2012.21196.x

[25] Lincoln A Mullen, Kenneth Benoit, Os Keyes, Dmitry Selivanov, and Jeffrey Arnold. 2018. Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software* 3, 23 (2018), 655. https://doi.org/10.21105/joss.00655

[26] Maurizio Naldi. 2019. A review of sentiment computation methods with R packages. http://arxiv.org/abs/1901.08319

[27] Murzintcev Nikita. 2016. ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters. https://cran.r-project.org/package=ldatuning

[28] Gloria Origgi. 2018. Say Goodbye To The Information Age: It's All About Reputation Now. https://www.fastcompany.com/40565050/say-goodbye-to-the-information-age-its-all-about-reputation-now

[29] Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent Semantic Indexing: A Probabilistic Analysis. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS '98)*. ACM, Seattle, Washington, USA, 159–168. https://doi.org/10.1145/275487.275505

[30] Tyler W Rinker. 2018. {lexicon}: Lexicon Data. http://github.com/trinker/lexicon

[31] Tyler W Rinker. 2018. {sentimentr}: Calculate Text Polarity Sentiment. http://github.com/trinker/sentimentr

[32] Tyler W Rinker. 2018. {textclean}: Text Cleaning Tools. https://github.com/trinker/textclean

[33] Margaret Roberts, Brandon M. Stewart, Dustin Tingley, and Edoardo Airoldi. 2013. The structural topic model and applied social science. In *NIPS 2013 Workshop on Topic Models (NIPS '13)*.

[34] Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. https://doi.org/10.18653/v1/k17-3009

[35] The Press Ombudsman. [n. d.]. Code of Ethics and Conduct for South African Print and Online Media. https://presscouncil.org.za/ContentPage?code=PRESSCODE

[36] Christian Wartena and Rogier Brussee. 2008. Topic detection by clustering keywords. *Belgian/Netherlands Artificial Intelligence Conference* May 2014 (2008), 379–380. https://doi.org/10.1109/DEXA.2008.120

[37] Herman Wasserman, Wallace Chuma, and Tanja Bosch. 2018. Print media coverage of service delivery protests in South Africa : A content analysis. *African Studies* 1 (2018), 145–156. https://doi.org/10.1080/00020184.2018.1426312

[38] Hadley Wickham. 2019. rvest: Easily Harvest (Scrape) Web Pages. https://cran.r-project.org/package=rvest

[39] Jan Wijffels. 2019. udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. https://cran.r-project.org/package=udpipe

[40] Yi Zhang, Jie Lu, Feng Liu, Qian Liu, Alan Porter, Hongshu Chen, and Guangquan Zhang. 2018. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics* 12, 4 (2018), 1099–1117. https://doi.org/10.1016/j.joi.2018.09.004

[41] George Kingsley Zipf. 2006. *The psycho-biology of language: An introduction to dynamic philology.* Routledge.