Lexicon based analysis of topic related word usage

Anke Jorun Unger
Psychology in IT / 2020
ankeju@stud.ntnu.no

Abstract

This paper introduces tools to analyse the word usage of topic related articles in comparison to non-topic related publications. First a content overview of the articles subset dealing with the chosen topic is visualized. Secondly, this work presents a period based analysis of possibly biased or offensive language to examine the development of the used expressions. Additionally a sentiment score for every timeframe is created by using a semi-supervised learning algorithm to create a domain specific sentiment lexicon. The tools were tested on data related to the corona crisis and the LGBT-community. The authors were able to give an overview over the topic of articles and identified hints about the proportion and polarity of controversial content for different timesteps.

1 Introduction

The main goal of this work is to create tools for a contentbased examination of newsarticels. This allows on one hand an easy inspection of the development over time and gives on the other hand an overview over the sentiment and opinion in a large cluster of publications. As the approach is supposed to work on varying sets of newsarticles, the tools are kept adaptable and are being evaluated with different sorts of data. The focus lays on analysing the usage of vocabulary, which could be classified as offensive or biased and to give a document based overview on the sentiment of articles. There have already been a larger amount of bias or aggressiveness detection conducted. Even more so for sentiment analysis tools. But the majority of those approaches consist of supervised learning techniques (Rosenthal et al., 2019; Chowdhury et al., 2019; Nina-Alcocer et al., 2019). Supervised learning is dependent on the existence of labelled data to train the classifiers but most datasets are only for selected use cases, for example the classification of tweets, political speeches or movie reviews. If the training data is not representative for the real world data the algorithm will create poor results. This is especially the case for highly domain specific texts. For incident a financial article will use a different vocabulary than a boulevard magazine or a politician during a speech. A lexicon is easier and faster to create and maintained than a labelled dataset and the quality of it can be easily controlled. By that the approach is far more flexible for different use cases, especially when the lexicon is automatically created or adapted in a semi-supervised approach. New words can be learned “on the fly” for every different domain.
The described goals lead to the creation of the following hypothesis:

Hypothesis 1 Bias, aggression and sentiment regarding topics and subgroups is changing over time and this is traceable in newsarticles.

Hypothesis 2 Controversial debates for example after special events trigger an enhanced usage of offensive, biased or sentimental words.

Hypothesis 3 The approach can be used for varying newsarticels and also for domains without labelled data.

2 Background and Related Work

Aggressive or offensive language detection has the goal to identify language that is targeted at a specific subgroup and often harmful for the victims. Problematic is, that what language is perceived as offensive, is not only dependent on the readers background, but also on the background of the source or author. Most approaches to identify this kind of language work with supervised learning techniques like from Nina-Alcocer et al. (2019). Additionally most of those approaches are trained on and developed for forms of short texts posted by individuals online in a non-formal language. A vast majority is specifically trained on “tweets”, for instance in the work from Ventirozos et al. (2017). This means that they work with very short texts and that background informations about the authors are provided. Some works archived good results by including those background informations. One example is the work of Pitsilis et al. (2018) where hateful language was detected by using informations about the writers as input features. But this is hardly possible for newsarticles, which are sometimes published without a specific reporter as a source given. Additionally the articles mostly represent the political position of the newspaper and not of one person. Biased language, like the offensiveness detection, handles mostly data from online users. A fundamental work was done by Recasens et al. (2013) were types of biased language were identified and learned trough corrected “wikipedia”-articles. Wiegand et al. (2019) pointed out problems of machine learning in this context, due to most datasets not being representative of real world data. Another challenge which is often solved with supervised learning techniques is sentiment classification (Rosenthal et al., 2019; Chowdhury et al., 2019). Sentiment classification means to classify words, sentences or documents regarding their polarity (Jurek et al., 2015). This is mostly conducted via a binary classification with a positive or negative tag as result. Sometimes also continued ratings are calculated, like in the work of Jurek et al. (2015). Lexicon based sentiment classification means classifying without the necessity of tagged training sets but just by using a lexicon or embedding of pre tagged vocabulary. Here the lexicon is used as an orientation to determine the sentiment of words in the text. The lexicon can be created manually, but it can also be automatically learned like in the approach of Colbaugh and Glass (2011) where the Lexicon was adaptable for different types of domains to classify the sentiment of whole documents.

3 Architecture

As can be seen in figure 1 the application is separated in the preprocessing step and two main analysis steps. First the overall language is scanned. Afterwards an examination of language development over time is carried out. The results are visualized in plots.

3.1 Preprocessing

During preprocessing the crawled articles are read in and tokenized. Every white space and every special character is removed. The data can be provided in .txt, .csv or .json format. Those articles are handled separately, depending on the topic of the articles and therefore transformed into two lists with the containing words as elements. The output consists firstly of all tokenized words as lists and secondly of all tokenized words collected into buckets depending on the articles publication date. The bucket size is optionally per year, month or day.

3.2 Language Analyser

The language analyser only uses the preprocessed texts, the publication date of those articles is per default not taken into account for this analysis. But the timeframe of crawled articles can of course be manually adapted to highlight for example differences in the context of words over time.

Context Analysis To analyse the context, in which vocabulary connected to LGBT or the corona-crisis appeared, a seed list is used. With every appearance of a word from the seed list in an article the three words prior to the word and the three following words are analysed. Because words like ’and’ appeared with a high frequency but are not informative, all words listed by NLTK as stop words are removed. The results are directly visualized in form of wordclouds.

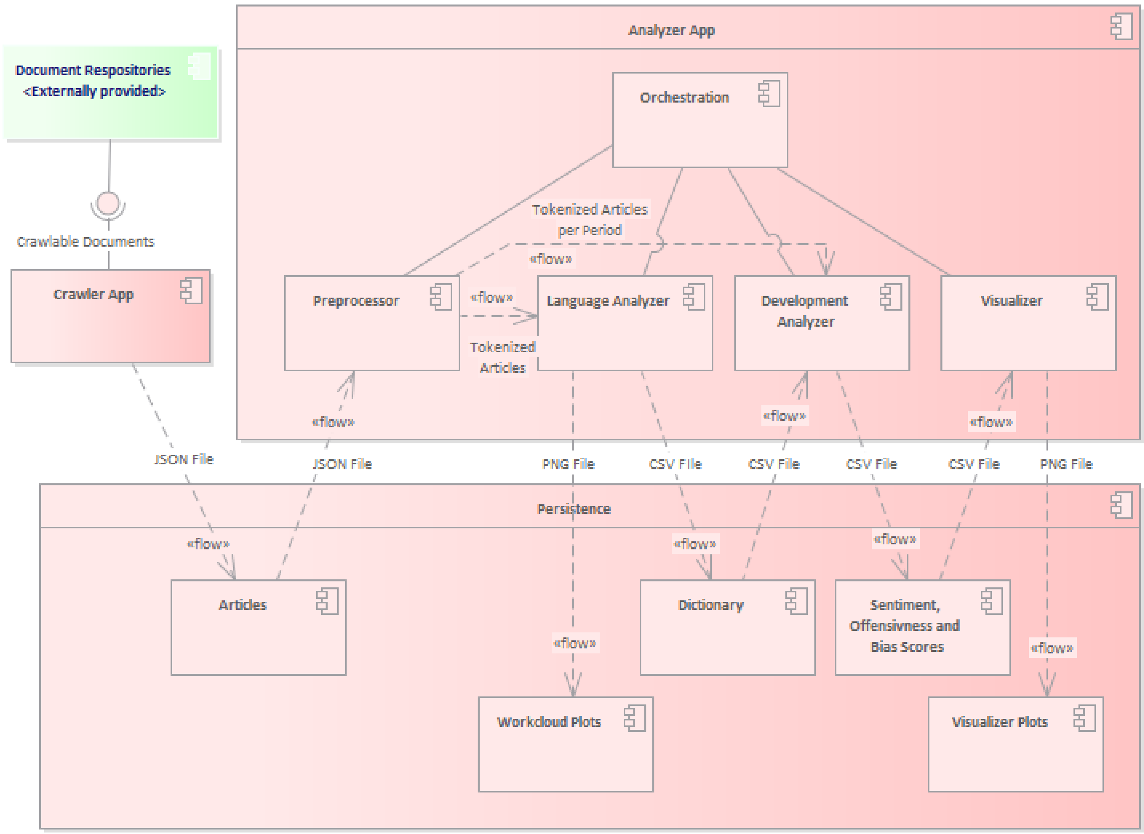


Figure 1: Architecture of the Lexicon Based News Article Analysis

Dictionary learning The Language Analyser uses a seed list of predefined sentiment words, to adapt the sentiment lexicon specifically to different types of articles. First the data gets labelled with the wordtypes. For that NLTKs tokenizer by Bird et al. (2009) is used. Then every adjective conjuncted to words from the sentiment list is recorded. Words with the conjunction 'and' are labelled with the same sentiment, while words conjuncted with 'or' receive the opposite label. Every entry in the dictionary receives a weighting regarding its inverse document frequency. The total number of documents is therefore counted and divided by the number of documents which contain the vocabulary. The result is saved as a csv-file.

3.3 Development Analyzer

The Development Analyzer is called by the controlling unit for every bucket of the dataset to create an overview of changes over time.

Offensive or Biased Language Analysis This analysis filters the reports for occurrences of words from the offensive language or biased language lexicon. Those occurrences are collected and counted. To make the result comparable, it is normalized by the number of words in the articles.

Sentiment Analysis The sentiment analysis checks if a domain specific sentiment lexicon was created in the previous step, else a default lexicon can be used. Every occurring word gets counted and weighted by its inverse document frequency. For every timeperiod a sum of those weighted appearances is created and weighted by the total number of words to create a sentiment score for each bucket of data.

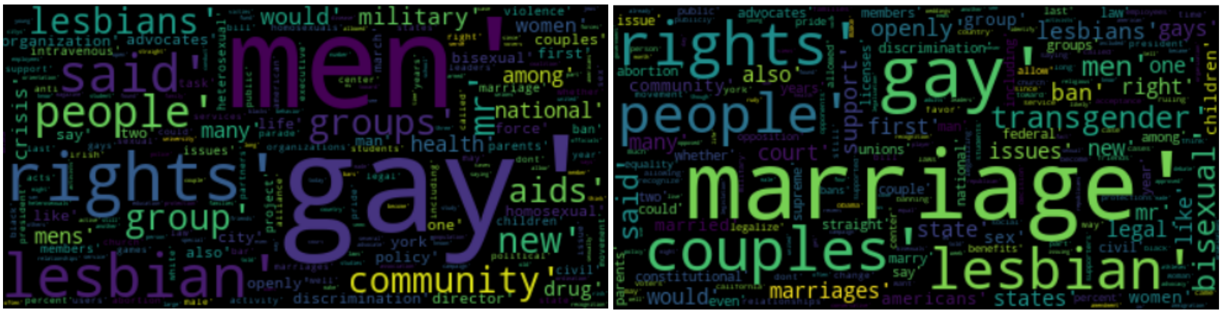
3.4 Visualizer

Line graphs are created to visualize the results of the development analyser. The x-axis is adapted, depending on the chosen step size (day, month or year). As some of the results can have a high variation, an additional time series analysis can be conducted. For that the series was split into its components (baselevel, seasonality, error and trend). Especially the trend gives suggestions about the development of the data.

3.5 Data

Two main datasets were used to test the analysis tools. The data to analyse the language connected to the CoVid-pandemic was crawled from the official “Fox News” website (Fox News Channel, 2020) via the “newsplease”-crawler by Hamborg et al. (2017). Tokenizing the articles lead to 14465312 words from foxarticles. Here 1407057 came from reports about corona and 13058255 words were part of publications without any specific topic. The “Fox News Channel” was chosen as it is part of an ongoing controversy (Martin and Yurukoglu, 2017; McNair, 2017) and especially criticized due to a drift towards offensive and biased reporting (Mills, 2017). To make the context analysis comparable, additional english articles from CNN were retrieved, with 219442 usable tokens with a different crawler (Lucas, 2016). Both crawlers had a filter set, to only crawl for articles in between 23.01.2020 and 11.05.2020. On 23.January was the beginning of Wuhans lockdown and start of the global spreading of the virus (Hua and Shaw, 2020). To analyse articles regarding LGBT, a longer timeperiod was chosen. The data consisted of news articles from 1986 to 2015 with content about the lgbtq community and an additional set of news articles, not specifically about the lgbtq community as ground truth. This lead to 2190579 tokens in articles about LGBT and 1475607 in articles about no specific topic. To identify articles about LGBT and to analyse the context a seed list with the following words was used: (agender, androgyny, androgynyous, bisexual, bigender, butch, cisgender, gay, gays, gender-fluid, genderqueer, lesbian, lesbians, LGBT, LGBTQ, queer, transgender, same-sex, homosexual, homosexuals). To identify articles and sentences about corona, the following seed list was used: (covid, cov, corona, sars, outbreak, pandemic, virus). A sentiment lexicon was used, which was already labelled by Liu et al. (2005) with positive or negative sentiment-tags. The used bias lexicon was created by Recasens et al. (2013) and the lexicon of possible offensive words by von Ahn (2020).

4 Experiment and Results



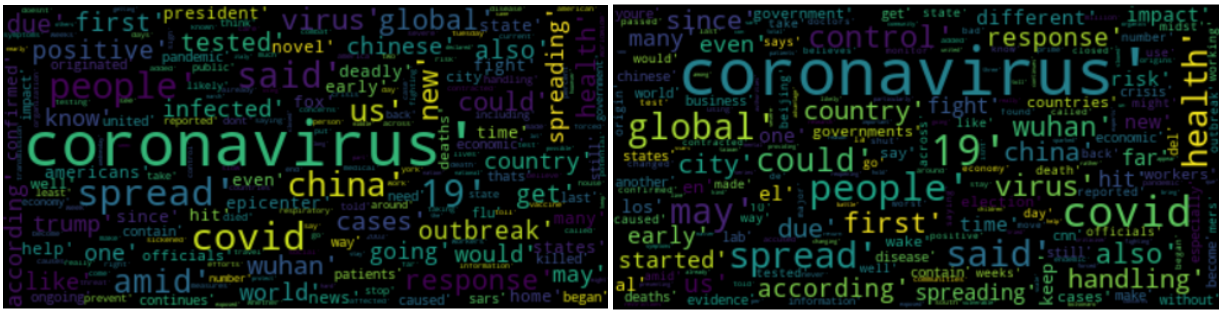
(a) Contextwords in LGBT-related articles 1986-1999

(b) Contextwords in LGBT-related articles 2000-2015

Figure 2: Wordclouds to show the context of the LGBT related seed words

In the experiment articles about LGBT-topics and the corona-pandemic were examined. This was done by filtering for topic related word occurrences.

For both topics, the context of seedwords was visualized with wordclouds. Afterwards a lexicon based frequency analysis of offensive and biased language and the sentiment was applied. As the data from the LGBT-corpus contained articles from a very large time period, results from every year were combined and plotted. The Fox-Dataset contained a shorter timeperiod but far more articles per day. Therefore the stepsize for plotting was day-wise. An additional trend analysis on some of the results was carried out, to visualize developments.



(a) Contextwords in FOX articles about corona virus

(b) Contextwords in CNN articles about corona virus

Figure 3: Wordclouds to show the context of the corona related seed words

4.1 Context

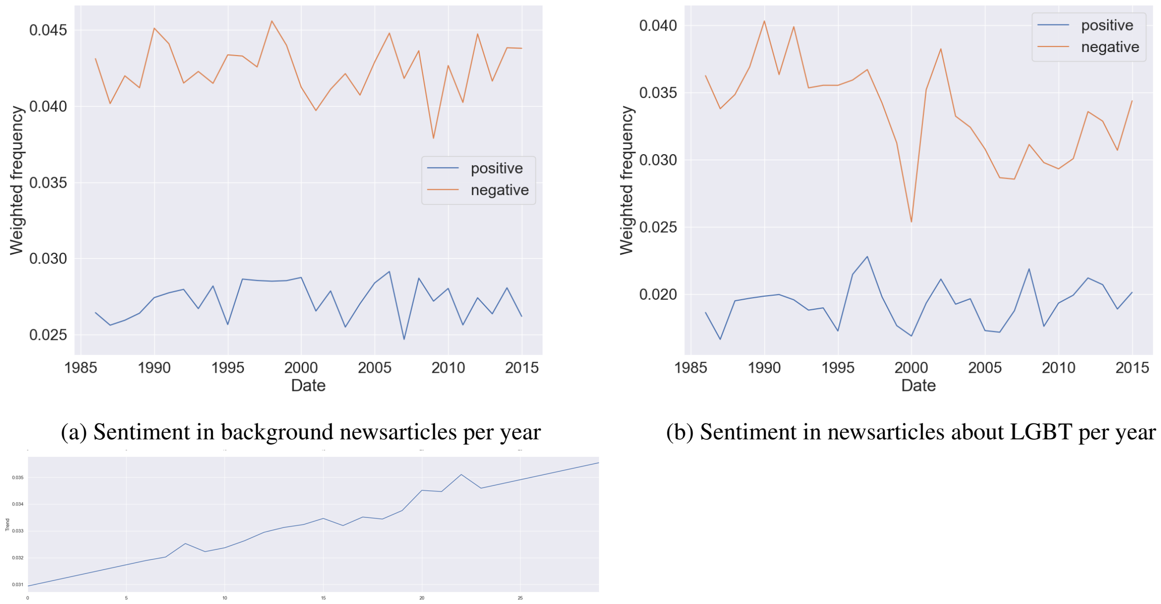
The most common contextword for LGBT related words was 'gay'. The contextwords were visualised in word clouds (see Figure 2a and 2b). After splitting the newsarticles in two subsets before and after the year 2000 it was shown, that the second most common word 'marriage' seems to only have grown popular after the year 2000. After a visualization of the context words from 1986-1999 the three most common words were 'gay', 'men' and 'rights'. The three most common context words from 2000-2015 on the other hand were 'marriage', 'gay', and 'rights'. For the word 'marriage' almost 95% of the total occurrence of the word was after 1999.

The context analysis of the Corona connected articles showed that the focus was mostly on 'coronavirus', 'spread', 'covid-19', 'people', 'said', 'china', 'amid', 'us', 'virus' and 'new'(see figure 3a). As there are no old articles regarding corona, a comparative analysis of the seed words context is not possible. Instead the context words can be compared to the vocabulary used in articles retrieved from CNN-News. CNN is another US-based news and television channel which is often described as opposing to FOX news (Chan-Olmsted and Cha, 2007; Farhi, 2003). Here the most common context words were 'coronavirus', 'covid-19', 'people', 'said', 'spread', 'may', 'global', 'health', 'could' and 'control' (see figure 3b).

4.2 Sentiment

Unique but frequent used sentiment words in the LGBT articles were 'openly', 'supreme', 'advocates' and 'benefits'. Most positive sentiment words appeared in both news corpora but were ranked slightly different. The negative sentiment differed a lot from the words used in the baseline articles. Here eight of the ten most used words were different. Common in the LGBT context but not in the base articles were the words 'discrimination', 'conservative', 'crisis', 'virus', 'opposition', 'risk' and 'infected' but the most commonly used word in both corpora was 'issue'. To evaluate the amount of positive or negative language in the texts a score was created. For that the number of occurrence was counted for each sentiment word, weighted by its inverse document frequency and divided by the total number of words in the article. The sum of scores for every article showed that the proportional amount of negative sentiment was higher in the LGBTQ articles, while the overall amount of positive or negative sentiment words was comparable to the background texts. The score was plotted for every year, as can be seen in figure 4a and figure 4b. A trendanalysis of the sentiment in LGBT articles showed an overall growth in the usage of sentiment words (see figure 4c) but no difference in the development between positive or negative wording.

The 10 most common positive sentiment words were similar for reports about the CoVid-crisis and reports from the same period. The only words that differed were 'positive' and 'great'. In contrast, the six negative words that appeared most often in corona-related articles were not used with a high frequency in unrelated articles. Commonly used were 'virus', 'outbreak', 'crisis', 'emergency', 'infected', 'risk'. All words were directly related to a pandemic or disease. The weighted score for positive words as well as negative was higher than in the base articles. The plots were created with a smaller stepsize on the x-axis and a sentiment score for every day (see figure 5a and 5b). Here the trendanalysis also showed an increase in the overall usage of sentiment words regardless of the polarity of those words.

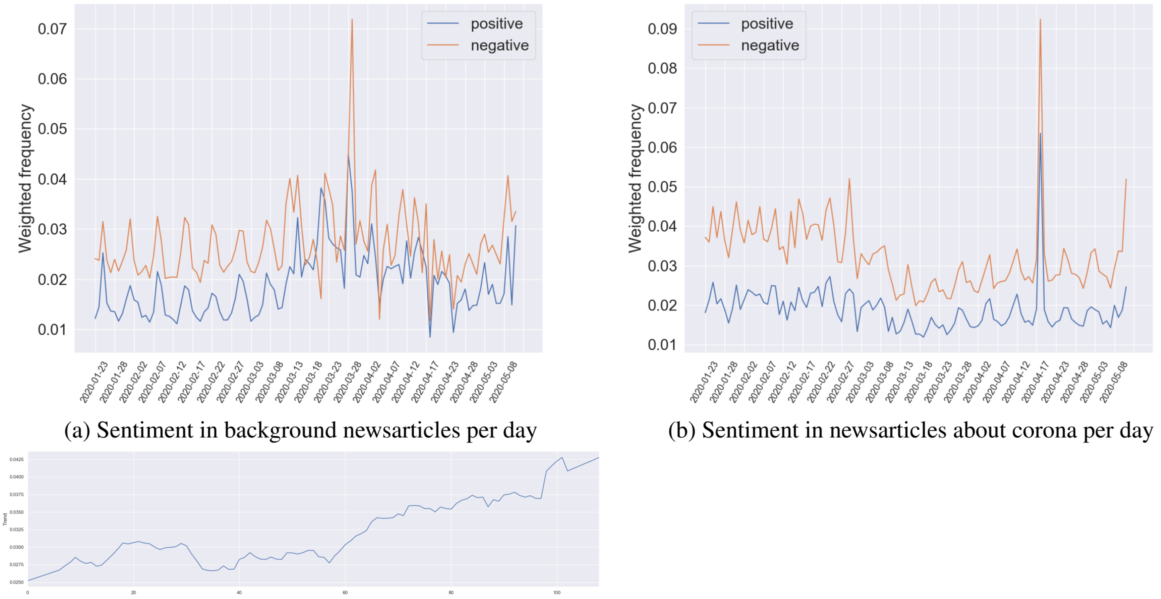


(a) Sentiment in background newsarticles per year

(b) Sentiment in newsarticles about LGBT per year

(c) Trendanalysis of sentiment in articles about LGBT

Figure 4: Sentiment analysis over time for LGBT related articles



(a) Sentiment in background newsarticles per day

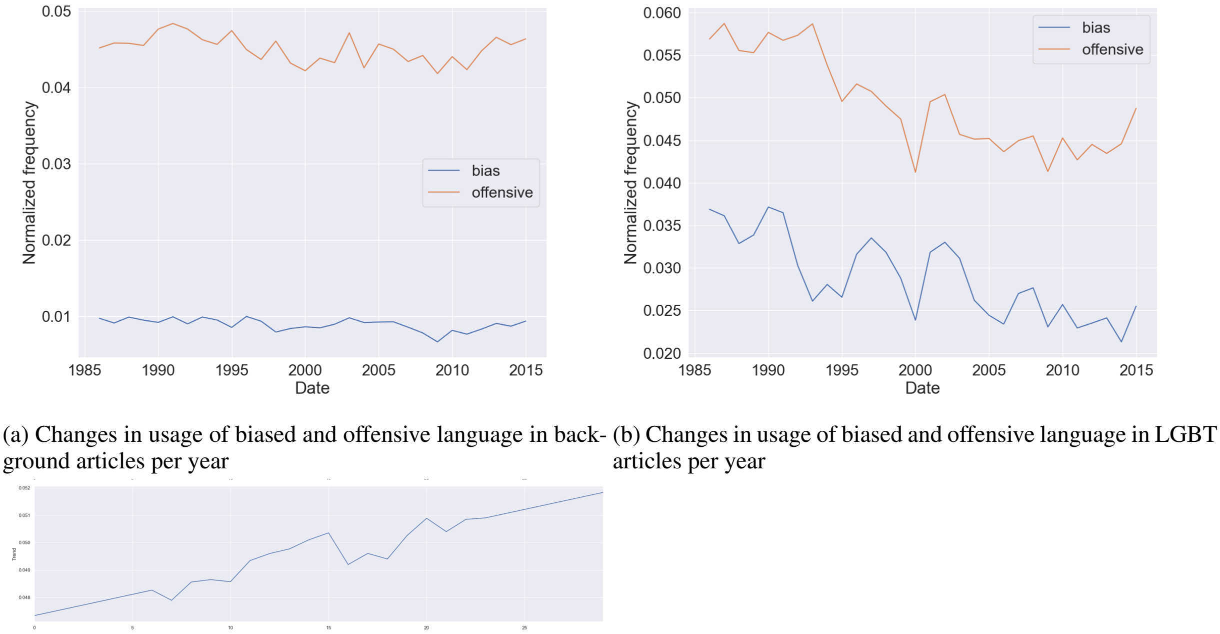
(b) Sentiment in newsarticles about corona per day

(c) Trendanalysis of sentiment in articles about Corona

Figure 5: Sentiment analysis over time for corona-crisis related articles

4.3 Bias and Offensiveness

Most high frequent biased words differ in the two corpora. The only vocabulary that showed to be used frequently in both corpora was ’government’ and ’case’. There is only a small difference in the usage of biased vocabulary in the LGBT-articles over time. So the three most used biased words in 1986-1999 were ’lesbian’, ’military’ and ’issue’ and in 2000-2015 ’lesbian’, ’issue’ and ’issues’ (’military’

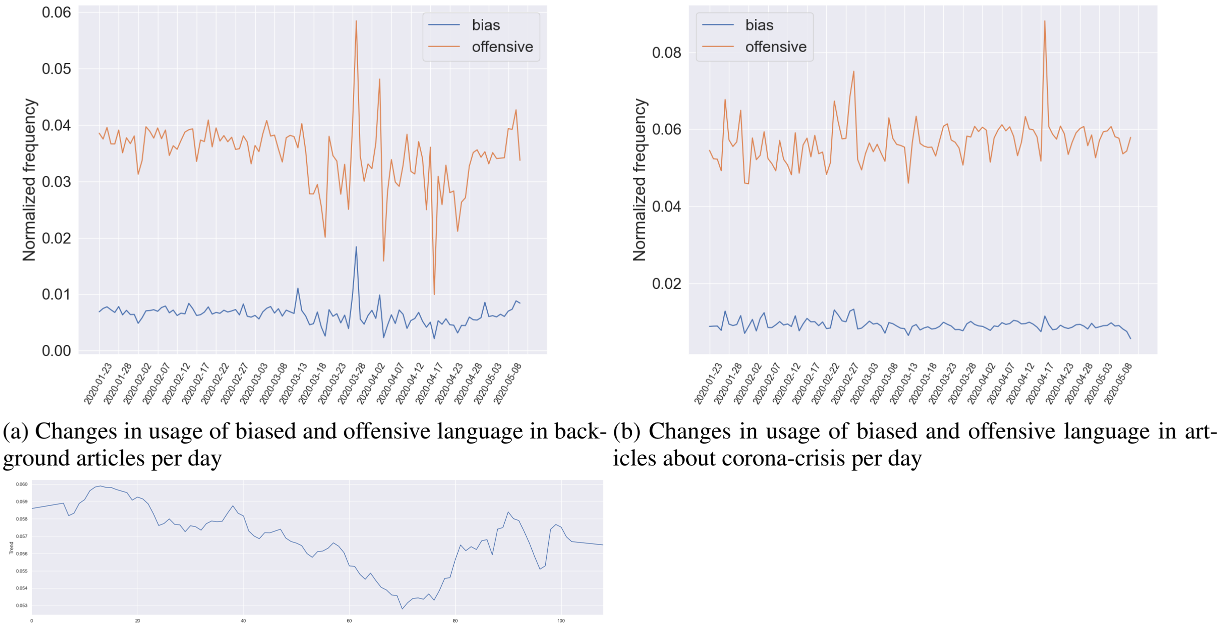


(a) Changes in usage of biased and offensive language in background articles per year

(b) Changes in usage of biased and offensive language in LGBT articles per year

(c) Trendanalysis of offensive language in articles about LGBT

Figure 6: Analysis of biased and offensive language for LGBT-related articles



(a) Changes in usage of biased and offensive language in background articles per day

(b) Changes in usage of biased and offensive language in articles about corona-crisis per day

(c) Trendanalysis of offensive language in articles about Corona

Figure 7: Analysis of biased and offensive language in articles connected to the corona-crisis

also appeared in the top ten but with a lower ranking). Overall the LGBTQ-articles contained with 107449 words (4.91% of all words) more hints of biased language than the baseline-articles with 66465 words (4.50% of all words). The top 5 most used words that could be regarded as offensive was 'gay', 'lesbian', 'homosexual', 'sexual' and 'church' in LGBT articles, while it was 'american', 'republican', 'black', 'death' and 'european' in the background texts. With 2.85% of all words being offensive, the overall usage of those debatable offensive words was far higher than in the groundtruth corpus (0.89%). The proportion of offensive and biased language was plotted in figure 6a and figure 6b. As can be seen in figure 6a the usage of biased words is persistent, with only little variation over time, while a high fluctuation in figure 6b can be seen for LGBT articles. Both developments show a real trend. The amount of offensive words is varying over time for both corpi. A trendanalysis for the usage of offensive words showed an increase (see figure 6c).

The most common biased words in articles about the corona-crisis were 'virus', 'pandemic', 'outbreak', 'government' and 'disease'. None of those words appeared in the background texts with a higher frequency. Overall the number of biased words in the articles was 22.0%, while it was only 1.05% in reports not connected to the pandemic. The usage of words that could be regarded as offensive was very similar in both corpi. The words differing were 'chinese' and 'sick'. Overall the number of possibly offensive words was with 3.48% of all words higher than in the background reports with only 0.19%. As can be seen in figure 7a and figure 7b the proportional amount of biased words is quite consistent in both corpi, while the proportion of offensive words shows a high fluctuation. A trendanalysis for the proportion of offensive language in reports related to the Covid-pandemic showed a decrease.

5 Evaluation and Discussion

Crawling the news article leads to problems with the way the crawler parses the website, for example leading to extremely long run times until the application reached recent articles. At the same time it is important to have a “well mixed” dataset to ensure representativeness. This was especially a problem as many news websites were build differently with different kinds of crawler working for one website while not working for another one. The application was able to cope with that by allowing input articles in very different file formats, as was shown in the experiment. The lexicon was sucesfully extended by new domain specific sentiment words. An example of extracted positive sentiment words trough conjunctions is 'economic', 'responsible', 'stable', 'angerthe' and 'heat-sensitive'. As can be seen by the extracted word 'angerthe' misspelled words were also included into the lexicon, but also correct and highly specific words like 'heat-sensitive'.

It is notable that the context words in LGBT articles changed rapidly over time. Here especially the word 'marriage' appeared only after 1999. The elevated usage could have been triggered by the legislation of same-sex marriage in different countrys. For example the first legislation in the USA in Massachusetts in 2004 (Webb, 2012). The sentiment analysis showed for both datasets a relevant difference in the usage of negative words, while the positive words used were similar to the background articles. This could be a hint that opinions which are emotionally loaded and are therefore expressed with sentimental words, are closer related to the examined topic and that most of those opinions are connected to rather negative emotions. The results show that the usage of sentiment words is continually increasing but no significant difference between the usage of positive or negative vocabulary can be seen. The usage of biased language seems to be slightly increasing with a peak in 1990-1999 and more biased language in LGBT-texts in every year. The difference in the amount of biased language hints between LGBT-articles and the baseline seems to be decreasing over time. It must be noted that biased language must not be used solely in negative contexts but it could be seen as a hint, how controversial a topic is and how emotional or subjective the discussion about that topic is. The reports about the corona crisis had with over 22% a very large proportion of words which could be regarded as biased. Problematic here is that from around 300000 identified words, 10000 of those occurrences were just from the words 'virus' and 'pandemic'. The usage of those words in the context of a disease can not be interpreted as hint of subjective language. It is therefore questionable if those results can be interpreted. The usage of offensive language seems to be increasing for LGBT articles over time, while it is decreasing for articles about corona. This is especially interesting as the rhetoric of some politicians in the USA were criticised to incite racism (Blofield et al., 2020; Sollmann, 2020) and there are signs that the corona crisis is impairing the developments of racial problems in the US (Haokip, 2020; Stechemesser et al., 2020). At least this analysis could not uncover supporting language in Fox News articles for that.

6 Conclusion and Future Work

Most analysis can be used in a ethical and helpful context or can be misused or have unwanted side effects. For example analysing the context of words can be very helpful to analyse text very fast and effective but the information in words, not included in the context get lost. By only noting the context, authors or magazines could be hastily branded or it could result in filter bubbles. The focus on seed words could also lead to a focus oriented or biased search as seed lists can introduce stereotypes or pre assumptions to topics. The same problem occurs in the sentiment analysis. The decision to include or exclude words in the sentiment lexicon is always in some form subjective and therefore biased. Due to the interpretability, fast alteration and the amount of words in languages even a adapting lexicon can never be complete. Further more the lexicon to recognize biased language was created by identifying corrected sentences from wikipedia articles (Recasens et al., 2013). This is a very effective way to generate a dataset but also hard to protect against manipulation, as the wikipedia community is not exclusive. Nevertheless using dictionarys allows for an analysis of domains, where the necessary data for supervised learning algorithms can not be provided. This work showed, that visualizing and interpreting the results of the lexicon based application allowed to get an overview regarding the development of opinions and the overall content of articles from specific time frames or domains.

References

Joanna Almeida, Renee M Johnson, Heather L Corliss, Beth E Molnar, and Deborah Azrael. Emotional distress among lgbt youth: The influence of perceived discrimination based on sexual orientation. *Journal of youth and adolescence*, 38(7):1001–1014, 2009.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

Merike Blofield, Bert Hoffmann, and Mariana Llanos. Die politischen und sozialen folgen der corona-krise in lateinamerika. *GIGA Focus Lateinamerika*, (03), 2020.

Sylvia M Chan-Olmsted and Jiyoung Cha. Branding television news in a multichannel environment: An exploratory study of network news brand personality. *The international journal on media management*, 9(4):135–150, 2007.

Rumman Rashid Chowdhury, Mohammad Shahadat Hossain, Sazzad Hossain, and Karl Andersson. Analyzing sentiment of movie reviews in bangla by applying machine learning techniques. In *International Conference on Bangla Speech and Language Processing*, 2019.

Richard Colbaugh and Kristin Glass. Agile sentiment analysis of social media content for security informatics applications. In *2011 European Intelligence and Security Informatics Conference*, pages 327–331. IEEE, 2011.

Paul Farhi. Everybody wins: Fox news channel and cnn are often depicted as desperate rivals locked in a death match. in fact, the cable networks aren't even playing the same game. there's no reason they both can't flourish. *American Journalism Review*, 25(3):32–38, 2003.

Fox News Channel. Fox news, 2020. URL `https://www.foxnews.com`.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. news-please: A generic news crawler and extractor. 03 2017.

Thongkholal Haokip. From 'chinky'to 'coronavirus': racism against northeast indians during the covid-19 pandemic. *Asian Ethnicity*, pages 1–21, 2020.

Jinling Hua and Rajib Shaw. Corona virus (covid-19)"infodemic" and emerging issues through a data lens: The case of china. *International journal of environmental research and public health*, 17(7): 2309, 2020.

Anna Jurek, Maurice D Mulvenna, and Yaxin Bi. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4(1):1–13, 2015.

Jian Liu, JianXin Yao, and Gengfeng Wu. Super parsing: sentiment classification with review extraction. In *The Fifth International Conference on Computer and Information Technology (CIT'05)*, pages 216–222. IEEE, 2005.

OY Lucas. Newspaper3k article scraping library, 2016.

Gregory J Martin and Ali Yurukoglu. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–99, 2017.

Brian McNair. Fake news–a user's guide. *The Conversation*, 6, 2017.

Colleen E Mills. Framing ferguson: Fox news and the construction of us racism. *Race & Class*, 58(4): 39–56, 2017.

Victor Nina-Alcocer, José-Ángel González, Lluıs-F Hurtado, and Ferran Pla. Aggressiveness detection through deep learning approaches. In *In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings*, 2019.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742, 2018.

Jennifer C Pizer, Brad Sears, Christy Mallory, and Nan D Hunter. Evidence of persistent and pervasive workplace discrimination against lgbt people: The need for federal legislation prohibiting discrimination and providing for equal employment benefits. *Loy. LAL Rev.*, 45:715, 2011.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, 2013.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*, 2019.

Brad Sears and Christy Mallory. Documented evidence of employment discrimination & its effects on lgbt people. 2011.

Ulrich Sollmann. The ambiguity of the psychological limitations of globalization or an uncanny cocktail of viruses. *International Journal of Body, Mind and Culture*, 7(1), 2020.

Annika Stechemesser, Leonie Wenz, and Anders Levermann. Corona crisis fuels racially profiled hate in social media networks. *EClinicalMedicine*, 2020.

Filippos Karolos Ventirozos, Iraklis Varlamis, and George Tsatsaronis. Detecting aggressive behavior in discussion threads using text mining. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 420–431. Springer, 2017.

Luis von Ahn. Offensive/profane word list, 2020. URL `https://www.cs.cmu.edu/~biglou/resources/`.

Rita A Webb. Overview of same sex marriage in the us: The struggle for civil rights and equality. *Retrieved on*, 2(1):7, 2012.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, 2019.