BIA 660 - Web Mining
Sentiment Analysis on News Websites

Project Report

Chetan Popli | Abhishek Desai | Yash Amitabh

TABLE OF CONTENTS

1. Motivation and Research...2
2. Problem Statement……………………………………………………………….3
3. Background and Related Work………………...4
4. Methodology……………………..……………………………………………...6
5. Analysis of Experimental Result...8
6. Conclusion and Future Scope...11
7. Bibliography...12

Motivation and Research Question

Media bias in India existed across the largest newspapers throughout the country, and political forces shape this bias. For example, funds from the government are critical to many newspapers' operations and budgets, and the current Bhartiya Janata Party (BJP) government has previously refused to advertise with newspapers that do not support its initiatives. This pressure leads the media to endorse government policies, creating unbalanced reporting where media bias can affect political behavior in favor of the incumbent. Many media outlets enjoy a symbiotic relationship with the government, in turn receiving attention, funding, and prominence. These trends damage India's democracy and also put journalists critical of the government in danger, threatening their right to physical safety.

Funds from the government are critical to many newspapers' operations and budgets, and the current Bhartiya Janata Party (BJP) government has previously refused to advertise with newspapers that do not support its initiatives.

This ability of media bias to influence political support in India can contribute significantly to democratic backsliding by harming journalists, preventing freedom of expression and government accountability, and influencing voters. Media bias in itself causes democratic backsliding because the media neither holding the government accountable nor informing the public about policies that strengthen the incumbent's power can increase authoritarian practices.

In addition, government efforts to constrain the media harms journalists, undemocratically violating citizens' rights and physical safety.

With the help of this project, we made an attempt at trying to understand whether we can understand these political biases without knowing where a media house receives their funding from, and solely on the basis of the article text and description. If this kind of an analysis can be performed, the general population can make an informed decision as to which media house they make use of to get their daily news, with full knowledge of the underlying political biases the media house might possess.

Sentiment Analysis: the process of computationally identifying and categorizing
opinions expressed in a piece of text, especially in order to determine whether the
writer's attitude towards a particular topic, product, etc. is positive, negative, or
neutral.

Problem Statement

Aim: To extract news headlines from each source, and check for outliers in case of reporters'/writers' sentiments

We tried to extract the texts about the news of the places we wanted to and did a sentiment analysis on the words used to describe that news and grade it accordingly. We want to provide a platform for serving good news and create a positive environment and give an easy interface to the user to fetch the news articles based on their genre and opinion from multiple sources as well as give maximum information to the user using minimum time and resources.

Three news sites were used for this project namely:

- News express
- Indian express
- NDTV

Background and Related Work

Electronic and press media have increasingly usurped newspapers as the primary source of news for the general public. Additionally, they are occasionally referred to as a society's mirror. Summarizing various reports on several subject matter occurrences helps news readers to quickly browse through news topics. Since readers tend to follow events, keywords, and subjects, presentation and identification are significantly implied techniques for all news. By synthesizing large bodies of newspaper text information into summaries, a proliferation of studies has been seen that have employed both long-form and unstructured newspaper data with social media, such as Twitter and Tumblr, to retrieve the opinion of the general public.

In [1], the paper examines the various patterns of words that appear in a well-renowned English newspaper published in Bangladesh called 'The Daily Star', as well as tries to deduce the relationship between 2018-19s possible social and political context. The study is carried out using contemporary text mining techniques such as Word Clouds, Sentiment Analysis and Cluster Analysis. The output is indicative of the country's passion for cricket, political turmoil and certain Rohingya-related issues.

In [2], the paper performs sentiment analysis on news articles to predict whether it will cause any change in the stock prices of a particular company. An attempt is made to take the non-quantifiable information such as financial figures and analyze its sentiments to predict the rise or fall of the stock price. This is performed using methods such as Naive Bayes and Random Forest.

In [3], a dataset consisting of 1329 articles collected from various Telugu newspapers is labeled with their bias towards a political party. They proposed the use of a headline attention model, which mimics how a person reads a news article, and assigns more bias to it inherently.

In [4] the proposed method entails the definition of metrics to measure political bias in GPT-2, after which a reinforcement learning framework to mitigate such political biases based on text is generated, using rewards from a classifier or word embeddings. The proposed method helped in reducing bias according to both the established metrics and human evaluation, while keeping the readability and semantic coherence intact.

In [5], historical data from the Thompson Reuters News Archive Data from 2003 to 2012 based on financial market performance is used to forecast financial news sentiments. A combination of Deep Learning methodologies such as Recurrent Neural Networks (RNN) along with Long Short-Term Memory (LSTM) is used in this method, and a noticeable increase in the forecasting accuracy is seen when negative articles are used to create the training dataset.

In [6], a simple methodology is proposed to quantify the positive, negative and neutral sentiments observed in news articles which are a part of the BBC News Dataset, which was able to express the applicability and validation of the adopted approach, which is essentially a lexicon-based approach for sentiment analysis of the said news articles.

Lastly, in [7], the results for the analysis of sentiment in political news articles are classified. In the proposed methodology, a two-step classification model is used, which distinguishes between the subjective sentiment text first and then it takes into account the positive and negative sentiments of the same. A shallow machine learning approach is used where a minimal number of features are required to train the classifier, which include the sentiment-bearing Co-Occuring Terms (COTs), along with negation words. Contrary to other research carried out, the use of negations as features does not have a positive impact on the evaluation results, and also suggests that the lack of sentiment lexicons and parsers needs to be integrated in articles to improve the databases of news articles created.

Methodology

a. Data Collection:

The data used in the project is all collected from media houses, namely - NDTV, Indian Express and NewsExp. These media houses have their own political leanings and biases, hence we classify them as Left Wing, Right Wing and Neutral.

The headline title and the description for the respective articles for the above-mentioned websites are collected with the use of Selenium, which is a library in Python used to scrape data from dynamic websites.

Selenium refers to a number of different open-source projects used for browser automation. It supports bindings for all major programming languages, including Python. The Selenium API uses the WebDriver protocol to control web browsers like Chrome, Firefox, or Safari. Selenium can control both, a locally installed browser instance, as well as one running on a remote machine over the network.

This data is then stored as text files in the respective folders for the various media houses. Unlike BeautifulSoup4, another library in Python widely used for web scraping, we chose to use Selenium since the former cannot perform certain functions such as traveling pages and clicking buttons, which was required particularly in our project to scrape a sufficient number of articles.

b. Data Pre-processing:

For this step, we exported the headline text and descriptions for all scraped articles and merged them into a single dataframe. Along with this, we append the source and bias (Left, Right, Neutral) for each article. We preprocessed the data by processes such as converting the article headlines, removing punctuations, stop words, tweet text, ads, etc. and lemmatizing the data.

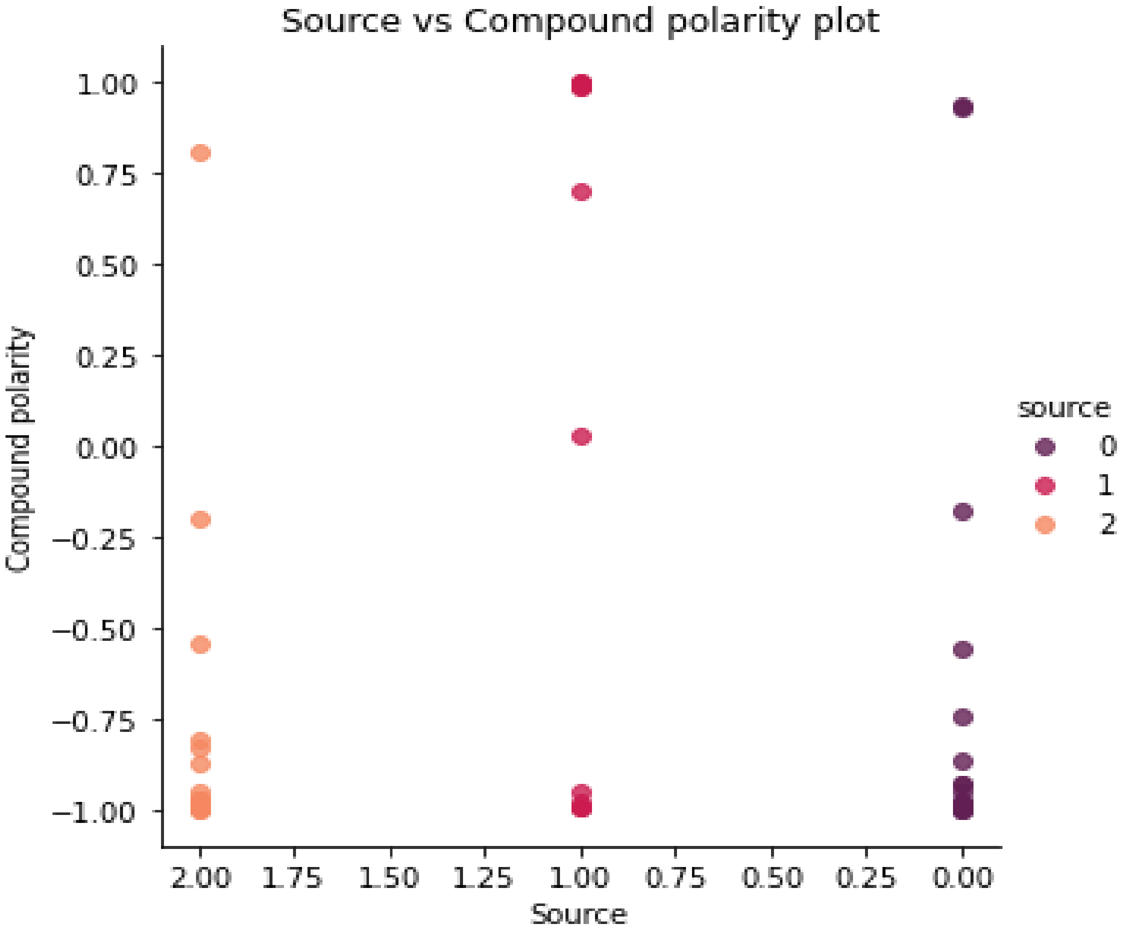
c. Model Generation:

We have predicted anomalies using isolation forest.
Isolation Forest: It detects anomalies using isolation (how far a data point is from the rest of the data). This method deviates from the mainstream philosophy that underpinned most existing anomaly detectors at the time. Instead of profiling all normal instances before anomalies are identified, Isolation Forest detects anomalies using binary trees. The algorithm has a linear time complexity with a low constant and a low memory requirement, which works well with high volume data.
The output we get from VADER is in the form of positive, negative and neutral scores between 0-1 and a compound polarity score between -1 to 1. If the compound molarity is lesser than 0 the result is negative, greater than 0 is positive and equal to 0 is neutral.

VADER:Valence Aware Dictionary for sEntiment Reasoning is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.

There are also alternate methods to VADER for sentiment analysis such as TextBlob but VADER sentiment analysis is social media focused and is appropriate for our project.

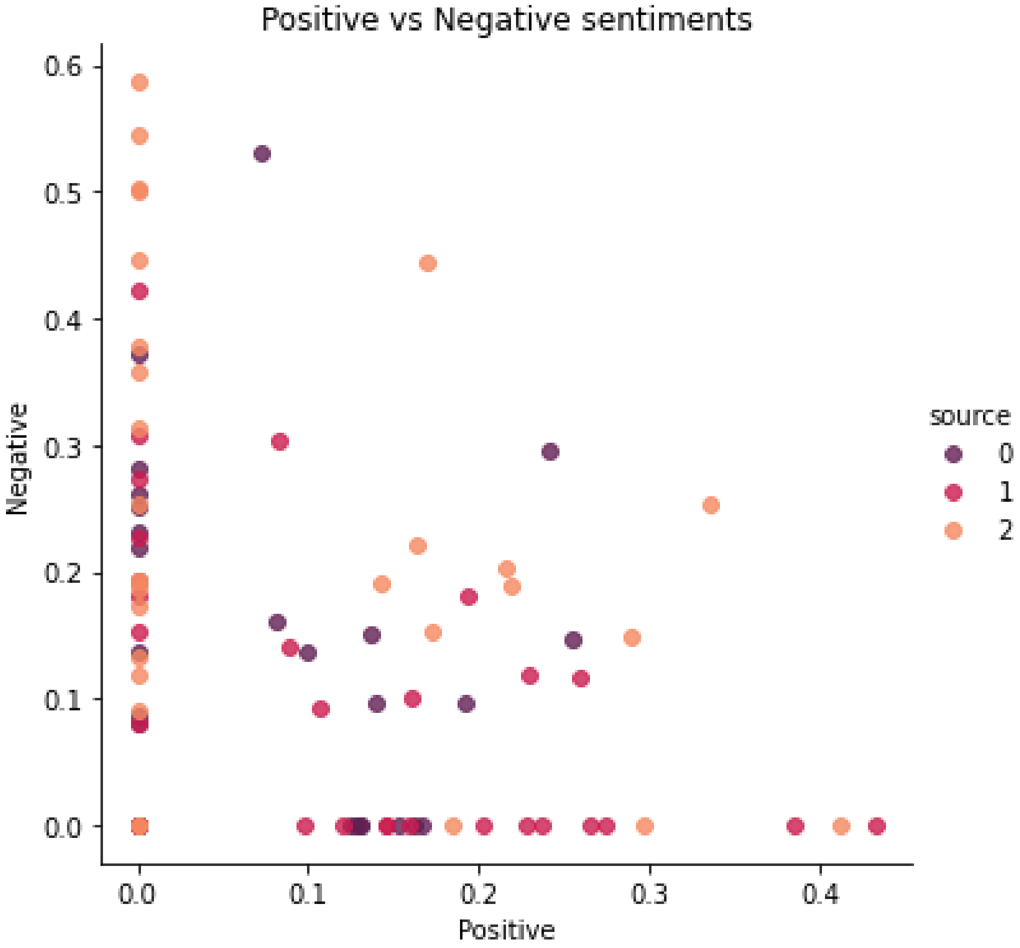
Analysis of Experiment Results



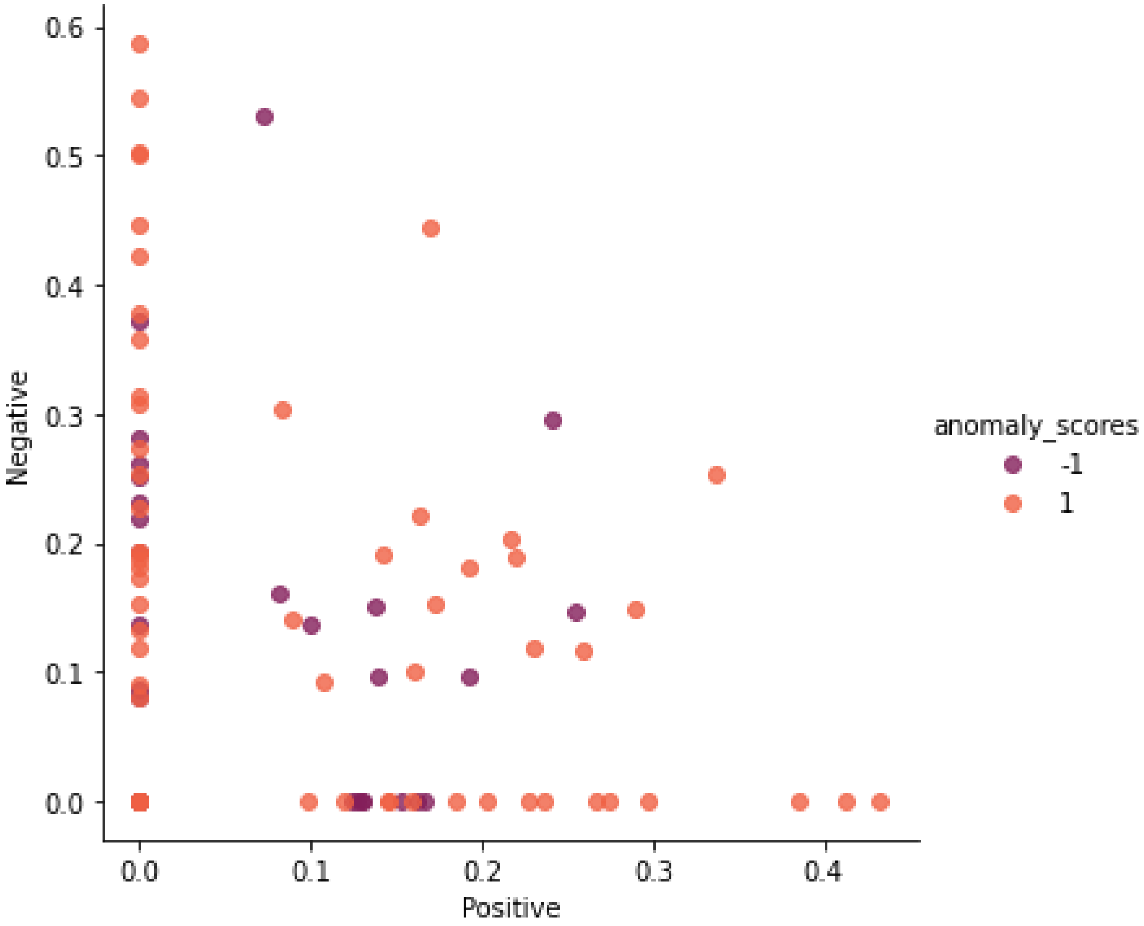
Plot 1- The output we get from VADER is in the form of positive, negative and neutral scores between 0-1 and a compound polarity score between -1 to 1. Graph 1 shows us the range of scores for the compound polarity scores.

Compound polarity:

The compound score is the sum of positive, negative & neutral scores which is then normalized between -1(most extreme negative) and +1 (most extreme positive). The more Compound score closer to +1, the higher the positivity of the text. Above text is 49.2% Positive, 0% Negative, 50.8% Neutral



Plot 2- Plotting a scatterplot for the positive vs. negative sentiment scores obtained using vader for all sources indicated in different colors.



Plot 3- Indicates the anomalies found in the dataset using the formula for anomaly score = total length (positive)/total length(neutral + negative)

Conclusion

In this project, we learned how to extract data from websites, then convert the data to useful information using sentiment analysis, and then used Isolation Forest to find out anomalies. From our testing on the keyword "Kashmir" we found 30 anomalies in a total of 90 entries.

There is a higher number of anomalies than expected. The reason is that using only 1 keyword results in too broad a "search area".

We were unable to get a clear idea about the political bias that a media house might have through this project, but were able to analyze the sentiments behind the headline and its description. This can definitely be improved in the future by collecting more data from each media house.

Besides this, the insights obtained from our project can be used in the future to give an idea to the people who follow certain media houses about the underlying political bias the media house might possess, helping them segregate the actual facts from these political and ideological leanings which might turn out to be misleading for them. Furthermore, the same can be carried out by the anomalies that we have detected in this project.

Result and Future Scope

Due to website constraints, we are only able to use one keyword at a go. We aim to increase the number of keywords so that we can more specific events which will help reduce the number of anomalies.

We can use methods other than VADER for sentiment analysis to get more parameters, thus being able to make better anomaly detections.

Bibliography

[1]: Hossain, A.; Karimuzzaman, M.; Hossain, M.M.; Rahman, A. Text Mining and Sentiment Analysis of Newspaper Headlines. Information 2021, 12, 414. https://doi.org/10.3390/info12100414

[2]: Gupta, O. "Sentiment Analysis of News Headlines for Stock Trend Prediction." *International Journal for Research Trends and Innovation* 5.12 (2020): 13-17.

[3]: Gangula, Rama Rohit Reddy, Suma Reddy Duggenpudi, and Radhika Mamidi. "Detecting political bias in news articles using headline attention." *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2019.

[4]: Liu, Ruibo, et al. "Quantifying and alleviating political bias in language models." *Artificial Intelligence* 304 (2022): 103654.

[5]: Souma, Wataru, Irena Vodenska, and Hideaki Aoyama. "Enhanced news sentiment analysis using deep learning methods." *Journal of Computational Social Science* 2.1 (2019): 33-46.

[6]: Samuels, Antony, and John Mcgonical. "News sentiment analysis." *arXiv preprint arXiv:2007.02238* (2020).

[7]: Bakken, Patrik F., et al. "Political news sentiment analysis for under-resourced languages." *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016.

[8]: wikipedia.com