News Tone Analysis

Bhavya Pandya
University of Adelaide
a1785085@student.adelaide.edu.au

Nickolas Falkner
University of Adelaide
nickolas.falkner@adelaide.edu.au

ABSTRACT

Consumption of news in text format is not a new phenomenon, but the medium of consumption has changed and transformed at a lightning speed.

With growing acknowledgement from the social media and digital content consumers about them falling into a loop of an endless 'Digital isolated bubble', there may not be a direct solution of all the recommendation algorithms on these platforms and the targeted advertisements, but the tone analysis of the news and the comparison between the news article and the headline will allow the public to be more conscious about the type of content they are consuming and the whether is the news media accurately representing the actual content of the article.

With the shift towards the digital platforms, there is much more of a scope to store, model and classify the data to achieve the required results.With the data in text format, we need to work on Natural Language Processing techniques in order to make the Computer understand and map the patterns. We can use text classification techniques either by using heuristics, machine learning or Deep Learning methods.

Furthermore, by learning about different features in the news articles in history that correlate with the world events,we will try and predict the effect of news articles on upcoming world events.

1 INTRODUCTION

With the increasing digital presence of large news corporations and even the local news outlets on all the different social media platforms as well as websites, we have seen a shift of public preference from the print and broadcast mediums of news consumption to a digital based news consumption, mainly through three platforms - Social Media applications, News Aggregator applications and websites.

The common factor for all the three platforms is that they are interlinked and that means that readers are only able to view a headline on their respective social media or news aggregator platform. This means there is an added incentive for the news agencies both local and large corporations to have headlines which entice the readers to actually engage the actual full article.

Another factor to take into count is the emergence of digital bubbles, meaning that people who follow one type of news are then further subject to similar recommendations through targeted advertisements and can find themselves in the midst of a ' digital isolated bubble' where they are recommended the same type of content continuously.

A learning system for each news information source to explain and predict the market behaviour.

1. To analyse the articles to determine the accuracy of the actual headline on the front page in comparison with the actual article 2. To analyse over a period of time whether if media house(small, medium and multinationals) is having a certain pattern of sentiments represented in the articles they choose to publish 3. To analyse the articles over a period of time, to find the patterns of how news articles can correlate with the actual world events.

Sentiment Analysis is defined as the process of computationally identifying and categorizing opinions from a text data and determining whether the writer's attitude towards a particular topic is either positive, negative or neutral. In this paper, we want to use sentiment analysis of news articles to determine whether if the tone of the news article actually reflected in the news headline and whether we can determine a pattern based on the data collected from news articles over a time period and use the patterns obtain to correctly predict the correlation between the real world events and the news articles.

2 RESEARCH QUESTIONS

(1) To use sentiment analysis to determine how accurately the tone in the article is reflected in the news headline.
(2) To find patterns in historical data of news article to determine the weight of the features in order to predict the correlation between the news article and world events.

3 PROJECT OUTLINE

This research project is geared towards 2 main outcomes, text classification and sentiment analysis. For text classification, we have performed data mining and data cleaning methods along with NLP techniques for cleaning the news text data and make a corpus of word pairs, known as n-grams.

1.News text is +ve,-ve or neutral and by how much 2.Correlate with the past events happening in the same news category and use this to predict the tone of any news article that may be published under similar circumstances in the future.

To achieve the above aims for the research we need to set the following sets of objectives to get to the desired output.

We will do a web scraping methods and overview to understand how the datasets are curated, how the websites are arranged and how to collect, store, organise and clean the data.

Secondly, we have to select from the different dataset to have data that is normalized and annotated and labelled without it causing an effect on the output accuracy. For the event dataset and news dataset standardization, we can make the news and even category features into categorical features and then compare the both. For text data cleaning we need to perform the different NLP pre - processing basic techniques to remove all the words in the text data that is not determinant of the sentiment of the news. We then make the token pairs and store them as vectors to train the model on this data for text classification.

After text classification is performed, we can use it to find the patters, perform, sentiment analysis and then use this sentiment

analysis output and compare with the event data to find the correlation with the real worl events.

4 BACKGROUND AND RELATED WORK

News tone analysis has been done previously on financial data to predict the impact of the sentiment of the news article on the variations in the stock prices.The whole system is based on the technique of Natural Language Programming called Sentiment Analysis which provides two types of functions : Opinion Analysis and Opinion forecasting.

According to the design principles of [15], a general workflow of a sentiment analysis pipeline has a workflow which starts with Web Scraping and web crawling techniques to get the text data from the webpage. [20]shows that the data extracted from the web pages has to be further processed before it is modelled using machine learning and deep learning techniques in order to initially create a parsing tree which can then be traversed and processed to convert the raw text into tokens.

Tokens as explained by [2]is the division of the text data into smaller portions depending on the task at hand. Calculating the number of tokens after data cleaning has been done using techniques such as lemmatizations, stemming, removal of stop words and Part of Speech tagging as shown by [13], we find that there is a tendency for the number of tokens to keep on increasing exponentially.

This has been worked upon by [18] by using the process of n-grams and by [2] making use of deep learning techniques to decrease the number of features that are extracted from this text data corpus.

Researchers have developed various ways to group tokens whether as unigrams, 2- grams or 3-grams. As shown by ..... that results of performing TF-IDF techniques on the n-grams results in three categories of output essentially : High frequency n-grams, low-frequency n-grams and medium frequency n-grams.

They show how low-frequency n-grams can lead to overfitting of the data and high frequency are essentially anomalies in the grammatical structure, both of which need to be omitted from our model which we are to train in order to get the required accuracy of the result.

TF is the technique to determine the frequency of a term t in a document d.

IDF is the technique to determine the log-normalised value of the inverse document frequency i.e. the total number of document / no. of documents in which the term t appears idf(t,D)=log| D |1+| dD:td |

TF-IDF is the product of the tf(t,d) and idf(t,D) values, whose result if high signifies that the term has a high frequency in one of the documents and low frequency in other documents. This is implemented using sklearn for feature extraction in text using the TFIDFvectorizer.

Let t=Term Let IDF=Inverse Document Frequency Let TF=Term Frequency

TF=term frequency in document/total words in document

IDF(t)=log2(total documents in corpus/documents with term)

tfidf(t,d,D)=tf(t,d)×idf(t,D)

The types of NLP approaches to extract the features and train the model include a dictionary based approach in which the model makes use of the pre-existing libraries such a WordNet in order to form the necessary distinction between the positive and negative emotions behind the text of an article and whether if the article was factual based or opinion based. Secondly, there are approaches which involve various word counting methods as shown by [14] , to then find the resultant patterns.

This is known as the vectorization process and as shown by ‘[22] this can be done using either Word2vec, GloVe or fastText techniques.

For word2vec, two words will have almost same vectors if they occur in the same way in the english text.That is, more than the realtion between the words is the semantical relation between the different words. [26]This is a major part of keyword extraction, such as we can say that both Serie A and Champions League are both related to soccer thus to sports directly.This is how the vectors that are used to train the model for text classification provide an added value to the output and we can categorize then in the same cluster.

These features are then used as input to a classification model which essentially finds the patterns in these features which in textual format can be seen as tags and labels. This as shown by [11] can be done using supervised modelling techniques and as shown by [27] and [23] using unsupervised techniques to find the patterns and group these features.

The types of unsupervised techniques that have been worked upon include Principal Component Analysis, Autoencoder and Boltzmann machine techniques. The advantages as depicted by [18] are the ability to extract non-linear features, capture non-linear relationships between words and find the meaning of the content behind the words.

As determined by [10] the overall accuracy gain for a deep learning model for a smaller dataset is not very significant when compared to performing sentiment classification.

To give the actual sentiment score and the polarity to any given document in the text data, as shown by [6]. there is a statistics based method which is a knowledge-based technique.

5 METHODOLOGY/PIPELINE

Firstly the research aims to establish the subjectivity and the polarity of the news articles that appear on the front page of the news websites over the world. For this research the project aims make a collection of different articles and make a corpus which can be further analysed to find the patterns in polarity of different news publications, authors and news topics.

For news data collection and analysis, we aim to understand the methods of data mining and text extraction. Furthermore for this we aim to understand how datasets and databases are curated and how to select the datasets for collection news data in terms of text of news article, news headline, news summary and related data such as publication website,date and category of news.

Lastly, the data we collect needs to be ready for classification and modelling for next semester and for that for this research we

started with basic NLP techniques implementation on a single article and then moved further towards a more standardized function for multiple articles and features.

5.1 Data Mining

Our goals require us to scrape the data from the different websites and then apply different Natural Language Processing techniques. As mentioned earlier We not only wanted to know what my dataset looks like, but also how it was curated and web scraping softwares are tools to aid in extraction of text data from websites. Web Scraping techniques can be classified into two clusters, first one where every article can be scraped manually using python libraries and requests and secondly where we can use specific text mining software to automatically extract data on a large scale or we can use pre-trained datasets

In our approach we have to scrape the data in text format from the front page of the different news publication website. For web scraping, BeautifulSoup library in Python to extract data and store it in a tabular format. Before we start the process of web scraping, we need to understand the website structure of the different websites because each website is made in different formats like HTML, CSS etc. Here we access the source code of the website in order to understand the block and class structure of the website. From this we use Beautiful Soup functions to store different data fields like Headline, Article short summary, Article link and the News Class(Sports, Business, Technology and Politics) into a data frame using Pandas package library.

In my observation, selenium is most preferable for large scale application but other methods can lead to a "BLOCKED IP" if there are continuous requests for url.

For data set selection, with the selection of datasets, we need to check to find the common features in the datasets and the datasets should be well tagged, annotated and classified. Now we are at a stage of our research where we know how to curate datasets and extract text data.

The Dataset 1 is the news text dataset The Dataset 2 is the list of recurring and significant events

For dataset 1 we have 3 critical fields which have text data of news articles, headline and summary.

As we can observe that we can use keyword extraction on the event datasets in order to predict the "News Category" of the event and then map the value in columns [eventime and newsCategory] in both the datasets. This will allow us to correlate and predict as per our objective

5.2 Text Pre-Processing and Parsing

So after we have explored our datasets for any missing values, any anomalies and basic features, we can start cleaning the dataset and perform necessary operations firstly for DATA CLEANING and then to make the data in a format which can be trained in a classification model.

Data fields except the timestamp for news and articles are in String format. For the category field, we can normalize the values in the field in order to make it categorical.

We have to divide the data into three sections first to then be able to compare the accuracy, polarity and subjectivity of the different texts. The three different sections being : Headline, Short Summary and Article. In the main article page there are further unwanted parts of the text data that we omit like images and URLs. Text for all purposes is viewed as a sequence of characters which further form words, phrases, sentences and paragraphs. Our pre-processing step uses the Natural Language ToolKit [NLTK] library in python to implement Tokenization, Stemming and Lemmatization techniques. In order to get the implicit meaning behind the different articles, we need to first break down the text into " Tokens" using functions in python for removing the white spaces, punctuations and handle contradictory words i.e. words which have a 'Not', 'Will' attached to it. Even though this allows us to convert the text into different smaller parts but still there will be a large number of repetitive tokens and will underfitting of the model. To overcome this problem, we use Stemming and Lemmatization techniques using ProterStemmer and WordNet Lemmatizer respectively. By application of lemmatizing after the stemming process where the base word for tokens is taken and suffixes are omitted, it doesn't convert irregular words like 'feet' into its base form and does give non-word tokens for words, e.g 'Wolves' is converted into ' Wolv' which doesn't make much sense for analysis. Thus by using the WordNet dictionary by the techniques of lemmatization, we convert the tokens into lemmas which give us meaningful tokens. Even after the optimization of tokens, there are many words which may result in ambiguity due to either capital words or acronyms present in the article, headline or summary.

5.3 Feature Engineering

These tokens in the corpus still need one last process to be ready for the feature extraction step in the project workflow and that is to convert the text data into vectors in order to form n-grams [collection of n number of words].

As we can observe the frequency distribution of the 1-gram, 2-gram and 3-grams to evaluate the vectorization process. The process of word to vector is done for the model to be able to extract features which are used to determine the polarity and subjectivity of the text data. Our model will form n-grams that are token pairs to then be run through layers of convolution filters in order to determine the log-normalised value of their occurrence frequency in the whole collection of articles. The 1-D convolution windows are preferred to the linear Bag of Words technique which when applied to all the documents gives a large number of 2-grams.This number of 2-grams is exponentially increasing as we increase the number of articles analysed. In our approach we work with dense representation of the tokens due to its increased comparability of vectors which are similar, i.e it makes a recursive filtering of the 2-grams through the convolution filter so as to understand the larger bracket for similar words.e.g cats and dogs and combined to come under a larger bracket of animals.By establishing more complex features from our model for the input text data, we further implement max pooling over time to reduce the number of features of our model while making sure our model is not over fitting the data. This is done by treating the different documents as a coagulation of the probability distribution of different topics. We use GenSim to find the different patterns in the clusters of the topics with respect 2 to time, by determining the hidden and derived topics

from the text data by applying algorithms such as Latent Dirichlet Allocation[LDA]. This allows us to automatically label these articles. For LDA we want to completely remove these words since it's highly likely the words with invalid characters appear only once and therefore don't have much value for the topics anyway. For word2vec we want to keep the order and the number of words per sentence the same; therefore we will replace these words with a random word 'abc". We will remove frequent words and stopwords since they probably bring little meaning and maybe even create noise when we want to classify later on.

5.4 Keyword Extraction

To find out the most relevant and the key words from the text, defined as per our context. We can leverage the output of this text analysis technique.by using NLP tools to break down the barrier between human language and machine language.This is visualized in terms of a word cloud.

This can allow us to obtain key tags for the event feature in the event dataset and allows to to convert this into a categorical feature which is easier to match with the actual text dataset. For our unstructured data, this is the optimum method to find the key words. By implementing word co-location techniques using n-grams allow the model to count separate words as one. This essentially helps in the process of monitoring of the real world events.

6 PROJECT TIMELINE

Semester 1

- Week 1-2 : Discussion about project topic with supervisor
- Week 3-4 : Project readings and preparation of project proposal
- Week 5-6 : Learn and setting up environment to implement different NLP techniques
- Week 7-8 : Data mining and pre-processing
- Week 9-10 : Implementing different NLP techniques
- Week 11-12 : Project progress report and Project Presentation preparation

Semester 2

- Week 1-2 : Continuation from last semester, implementing different NLP techniques
- Week 3-4 : Visualization of tone of headline and article and correlation
- Week 5-6 : Implementing the model to find pattern and correlation between prediction and real world events
- Week 7-8 : Comparing output and evaluation
- Week 9-10 : Final report and presentation
- Week 11-12 : Final report and presentation

7 EARLY RESULTS

For the research we have started with data mining in order to understand how the dataset is created from the data that is extracted from the web using different machine learning techniques.

7.1 Data Mining

To extract any kind of data from the websites, we need to understand the websites and the different frameworks they are built on.These frameworks, such as HTML, CSS, JavaScript or their combinations determine how the website works, so we need to choose a different approach for every website. Even for news article data extraction, the process is distributed into two sections, individual and multiple news article data collection at the same time. We first see that the CSS language facilitates the websites to make the content looks more creative and more designs, while the JavaScript allows the website to connect smoothly in a user friendly manner with the backend support for the database.The HTML language is where we find most of the text data stored. We used basic HTML knowledge to differentiate the divisions, classes, tags and identify the data available in different websites. For BBC websites we tried to extract the article and related data only for news article.We determined that the website has the divisions : article body, story introduction and other information, when we inspected the article link on a browser. We use this url with BeautifulSoup find function to find the article body and get the content text data.This allows us to access the content in HTML format although as a list of paragraphs.We can further use this for text parsing with different formats, also we can access the text data one paragraph at a time. The original inspected output also has a title tag in the same division and a division id that contains the news tags.Thus we can now extract full article text data, acces the data in HTML format at variable text data lengths. Next we access the front page of the BBC websites, this shows us multiple categories and sections where we can further find multiple articles.We thus need to understand the HTML framework basics to find how the data is structured. We can make the classes to extract , store and perform operations on each of the classes that have stored data about the particular news article.To access these classes and functions we can use attributes. This allows us to make a block of code which can allow us to collect data of multiple text articles from the front page of a particular website. This way the data can be collected everyday but it requires a manual entry of updated url each day.There are also multiple versions of the same publications that are published on the internet such as BBC News has a UK version and an international version.

7.2 Exploratory Analysis

For our dataset we need all this data for multiple news articles and multiple news publications.Having understood how the data looks on the websites and tried the different tools to extract this data, we use datasets which are pre-trained from HuggingFaces dataset package. There are different types of curated datasets that have been pre-trained, cleaned, annotated and are made available for research purposes.The news dataset has the following features : articleLink, Author, Headline, articleText, Publication, Category and timeStamp. This dataset in itself can provide us with the text data in the article as well the data about the article such as link, category etc.

We will use this dataset to extract the different features from the text data to perform the text classification to determine the tone of the news article, headline and summary. We also have a second event dataset that has features such as Events and eventTime. We perform keyword extraction on the text dataset to match with the event categories in this dataset.We add another feature to the dataset

that contains event labels in the dataset that have been matched for keywords in the article and timestamp of article.

7.3 Text Pre-Processing

There is string data in all features and the category data is converted into categorical data by using labels.This makes the dataset more standardised.We clean the dataset for anomalies in features that give information about the article, e.g. publisher, timestamp and author. For the features such as headlines, summaries and the actual body, we will use NLP techniques for pre-processing this data.

Furthermore basic NLP techniques need to apply on the 3 critical features individually to clean the text data for any urls, , stop words, HTML tags and punctuations. Also with the average number of words for articles being close to 400 words, we can use techniques to determine the basic and root form of words in order to restrict the number of tokens from the corpus. In stemming we find the basic form of the word and the stem word may or may not be in the dictionary Whereas, for lemmatization, we convert into the root form of the words which can be found in the dictionary. For tokenizing we can use NLTK python library to choose from white space tokenizer, word punkt tokenizer and treebank word tokenizer.

After this, we get a clean corpus, free from any contractions, any irregular expressions, all converted to their basic root word form and converted into token of different n-gram pairs. These tokens are stored as vectors for text classification purposes.These tokens are used as training data for sentiment analysis purpose with the different text classification models.

8 PROGRESS REPORT

Project status as of now is as follows : We have performed web scraping using the different machine learning techniques using request to get the data on a particular news article url.After further understanding the what type of data is stored and where it is stored on the website, we make use of classes and functions to make a standardized code block for the web scraping of multiple news article from the front page of a news website and also collecting the related data. We then checked and compared the different pre-trained dataset and found that there are news datasets which collect the following information such as news article link, category and time as well as the article text data.This dataset is curated carefully over many years for news from large number of publications around the world for long period of time.

We provide an overview of the Web Scraping Methods.The methods and process of Dataset selection and feature data normalization, Event dataset and news dataset standardization We perform Text Pre - Processing by implementing the NLP basic techniques and compare the output at the different stages.

Our task for this semester was divided into two parts,firstly about getting the required data and secondly about the implementation of text pre-processing and basic NLP techniques.This includes the text data in all three features to be cleaned and reduced to be clean of all words which can contribute to lower output accuracy. The NLP techniques allow us to understand and clean the data and then convert to a text corpus which is the penultimate output stage for our research this semester.After we have established the text corpus for the text data in each of the features individually, we convert into vector and form pairs to check for output for different number of n-gram pairs. Converting text data into vectors allows the data to be used to train the classification models. This is our final output stage for the research this semester.

9 PLAN FOR NEXT SEMESTER

My plan for next semester is to : Compare output for SVM, Randomforest and Deep Learning n/w.Used for text classification.Used for sentimental analysis on 3 critical features.Compare output of SVM and TF-IDF and for keyword extraction.Used to determine the news category of event in event dataset.Used to predict the correlation between the events and news tone.

9.1 Modelling and Pattern Mining

LDA and Word2Vec are used to extract sensible feature data from the documents. By applying a few supervised classifiers to be able to predict what text belongs to which author, publication and news class. We have to first encode our categories in each of these sectors for us to then apply LabelEncoder. We train the model by applying the grid search with cross validation arguments.This helps us determine a dictionary from the grid search output and then train the model on the hyperparameters selected from the best outcomes of the grid search. We evaluate both our models trained using LDA features and Word2Vec features to find any patterns which allows us to select the more accurate model for our topics and tasks. Logistic regression with the w2v features works as follows: Once we have vector embeddings for the words in our vocabulary through word2vector, we sum up all the vectors of a document and divide it by the number of words in that document. We get one vector with the average word vector of that document. This will be our input for the logistic regression model. For LDA this vector is calculated directly because the methodology for LDA is to directly give a vector with topic distributions. Therefore LDA gives a straightforward comparison.

9.2 Sentiment Analysis

TextBlob library is used with python to determine two factors from the extracted features : Polarity and Subjectivity of the text data from the article, summary and the headline. This is applied on all three separately and the results are then compared to give output in terms of varying factors such as News Publication, News Author and News Topic with respect to time. To find polarity we determine the overall emotion of the text data by comparing our features with pre-existing features which are further customised depending on the topic of the news and to find the subjectivity of the text data, i.e. the level of opinion involved in the representation of the news we determine by the application of weighted encoders on each article on a -5 to +5 scale with -5 being most negative, +5 being most positive and 0 being neutral. Aggregate tone of each word in a sentence gives the tone of the sentence and the aggregate tone of the sentence is marked to make the aggregate tone of the paragraph, similarly scaling to give the tone of the article.This is then represented with respect to time and is used to compare the variation in the tone of the news in terms of the polarity and subjectivity over time. We apply news focus techniques in order to give differential weights for recurring topics as a news article

will not have the same effect on the readers with each passing day. Application of news focus techniques is needed to determine the varying impact for each article on public opinion and events.

9.3 Correlation with real word events

This in a sense comes around to the expected output, where as per our objectives, we need to determine the tone of the text in news articles, headlines and summaries, compare them and then map them with events in the world to find any correlation and pattern between the tone of news and events in history.

Thus the event feature which is mapped categorically by implementing keyword extraction on the event data, is then compared with the news text data with respect to the time of occurrence and the news category.

By knowing what are the key events at the time and in that news category, the news was printed, we can determine the correlation between the sentiment of the news and the event that occur ed. This can be then used to find any patterns if any, between the sentiment of news articles around the real world events.

10 DESIRED OUTCOMES

This research aims to find the existing patterns between the variation in polarity and subjectivity of a news article with respect to real world events, when mapped over time over three sectors : News Publication, News Author and News Topic.

This when represented after determining the features from our deep learning model, will give the Polarity - Positive, Negative, Neutral; the Subjectivity - If it is an opinion or a factual text data among the three sections of a news article : Headline, Short Summary and News Article. This allows us to compare whether the summary and headline are an accurate description of the sentiment that is represented in the article.

For events in the real world, the model should be able to categorize the tone of the text and predict the tone of the article - based on the category and time of the event.

The model should find the sentiments for a supposed news article and its headline when matched with the time, event category and the news publication.

11 CONCLUSION

In this study we have proposed a model to evaluate three structures of a news articles : The headline, the summary and the article body in terms of its accuracy and its polarity and subjectivity to determine if any patterns are existing among the news publication, news author or news topic when the output of the news tone(Polarity and Subjectivity) are mapped with respect to time.

This allows us to find a correlation between the varying patterns in news polarity and the actual real world events.We find the patterns existing on a time of the day, day of the week, week of the month, month of the year and the year.

We have selected the pre-trained datasets for the features that gives us the information regarding the text data in headline, body and summary as well as the data related to the article such as the date, publication website and the link url.By performing the basic NLP techniques such as removal of contractions, stem words, converting the words into lemma form, in order to make the data into smaller tokens which are then converted into vectors which are access in n-gram pairs.

We have proposed to find the patterns that exist between the variation in polarity and subjectivity of a news article with respect to real world events, when mapped over time over three sectors : News Publication, News Author and News Topic. This when represented after determining the features from our deep learning model, will give the Polarity - Positive, Negative, Neutral; the Subjectivity - If it is an opinion or a factual text data among the three sections of a news article : Headline, Short Summary and News Article. This allows us to compare whether the summary and headline are an accurate description of the sentiment that is represented in the article.

We focus on implementing the neural networks to find the best technique to analyse the sentiments of articles that are appearing on the front page of the various newspapers across the world.

REFERENCES

[1] adeoyewole. 2018. Simple Guide to Scraping News Articles in Python. (2018). https://medium.com/@adeoyewole/scraping-news-articles-in-python-53c567282e25

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*. 30–38.

[3] Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, and Gurpreet Kaur. 2016. Opinion mining and sentiment analysis. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 452–455.

[4] Axel Bruns and Stefan Stieglitz. 2013. Towards more systematic Twitter analysis: metrics for tweeting activities. *International journal of social research methodology* 16, 2 (2013), 91–108.

[5] datenstrom. 2017. TFIDF. (2017). http://datenstrom.gitlab.io/cs532-s17/notebooks/TFIDF.html#TFIDF

[6] Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. 2007. Large-Scale Sentiment Analysis for News and Blogs. *Icwsm* 7, 21 (2007), 219–222.

[7] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. 2011. Good friends, bad news-affect and virality in twitter. In *Future information technology*. Springer, 34–43.

[8] Vasileios Hatzivassiloglou and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*. 174–181.

[9] Tuan-Anh Hoang and Ee Peng LIM. 2012. Virality and susceptibility in information diffusions. (2012).

[10] Muhammad Usama Islam, Faisal Bin Ashraf, Ali Imam Abir, and MA Mottalib. 2017. Polarity detection of online news articles based on sentence structure and dynamic dictionary. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*. IEEE, 1–5.

[11] Vishal Kharde, Prof Sonawane, et al. 2016. Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971* (2016).

[12] Andrew Kohut, Carroll Doherty, Michael Dimock, and Scott Keeter. 2010. Americans spending more time following the news. *Pew Research Center* (2010).

[13] Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. 2014. Towards building a social emotion detection system for online news. *Future Generation Computer Systems* 37 (2014), 438–448.

[14] Kevin Lerman, Ari Gilder, Mark Dredze, and Fernando Pereira. 2008. Reading the markets: Forecasting public opinion of political candidates by news analysis. (2008).

[15] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[16] Michael H MacRoberts and Barbara R MacRoberts. 2018. The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology* 69, 3 (2018), 474–482.

[17] Farhad Malik. 2019. NLP: Introduction To NLP Sentiment Analysis. (2019). https://medium.com/fintechexplained/sentimental-analysis-an-introduction-7fc21d9b8625

[18] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* 5, 4 (2014), 1093–1113.

[19] Andrew Montalenti. 2019. Machine learning for news. (2019). https://blog.parse.ly/post/7790/machine-learning-nlp-parse-ly-currents/

[20] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining.. In *LREc*, Vol. 10. 1320–1326.
[21] Julio Reis, Fabrıcio Benevenuto, Pedro OS de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. *arXiv preprint arXiv:1503.07921* (2015).
[22] Adam Hale Shapiro, Moritz Sudhof, and Daniel Wilson. 2020. Measuring news sentiment. Federal Reserve Bank of San Francisco.
[23] Wataru Souma, Irena Vodenska, and Hideaki Aoyama. 2019. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science* 2, 1 (2019), 33–46.
[24] Ubale Swati, Chilekar Pranali, and Sonkamble Pragati. 2015. Sentiment analysis of news articles using machine learning approach. In *Proceedings of 20th IRF International Conference, 22nd February*.
[25] Nianxin Wang, Huigang Liang, Yu Jia, Shilun Ge, Yajiong Xue, and Zhining Wang. 2016. Cloud computing research in the IS discipline: A citation/co-citation analysis. *Decision Support Systems* 86 (2016), 35–47.
[26] Kajal Yadav. 2020. Scraping 1000's of News Articles using 10 simple steps. (2020). https://towardsdatascience.com/scraping-1000s-of-news-articles-using-10-simple-steps-d57636a49755
[27] Wenbin Zhang and Steven Skiena. 2009. Improving movie gross prediction through news analysis. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, 301–304.