THE UNIVERSITY OF ADELAIDE

PROJECT THESIS

Masters Research Project

Author:
Bhavya PANDYA

Supervisor:
Prof. Nickolas FALKNER

A thesis submitted in fulfillment of the requirements
for the degree of Master of Data Science

in the

Department of Engineer Computer Mathematics and Sciences

June 15, 2021

Declaration of Authorship

I, Bhavya PANDYA, declare that this thesis titled, "Masters Research Project" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Bhavya Kavit Pandya

Date: 18 - June - 2021

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

THE UNIVERSITY OF ADELAIDE

Abstract

Data Science
Department of Engineer Computer Mathematics and Sciences

Master of Data Science

Masters Research Project

by Bhavya PANDYA

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 implementing the various Machine Learning and Deep Learning techniques to determine the sentiment behind the news articles and headlines, we aim to group them bu category in order to extract meaningful insights of the patterns that correlate the news article, news headline, news category, news publishing date with the overall sentiment that the author wanted to convey while exploring the different learning techniques used for sentiment analysis applications.

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Overview

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 account 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.

We aim to perform Text based sentiment analysis of NEWS data, by comparing different models to review the sentiment scores and determine if some aspects of the NEWS are more polarised than others.

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 news category, the sentiment scores and the news articles.

There will hopefully be many figures in your thesis (that should be placed in the *Figures* folder). The way to insert figures into your thesis is to use a code template like this:

1.2 Research Question

1. To use sentiment analysis to determine the sentiment score for News Articles and News Headlines.
2. To find patterns in NEWS data collected over a period of time, to determine the weight of the features in order to predict the correlation between the different features of the News Data to answer the following questions :
   (a) Are headlines more polarised than the article?
   (b) Are NEWS articles in certain categories more polarised?
   (c) Does the polarity of NEWS data depend on events?
3. To compare the sentiment scores generated using different Machine Learning and Deep learning models.

1.3 Motivation

News articles are concerned with real world events which compromise a layer of emotion - good, bad or neutral. Now with advancements in the digital technologies and onus being put on digital creators and publications, the amount of data to be collected and processed is ever increasing and we need sophisticated machine learning and deep learning techniques in order to classify these articles based on the emotions.Sentiment analysis is the use of machine learning techniques for classification of emotions that are present in any form of text data.

With the increase in existence of sheer number of sources from where people opt to get their news and coupled with the ever increasing busy lifestyle people are choosing for themselves, there is much more of a value put on the opinions and reactions based on a certain news article rather than the actual information gathered or the factual analysis of the event that a certain news article may be focusing upon.

When we couple the increased focus on opinions and reactions by different sections of the news cycle, i.e the news publishing websites, social media platforms with the fact that most people don't really take out the time to read full articles, fact-check the sources and the statistics. This means the news consumers are ever so increasingly targeted and baited in exchange for increased engagement on the digital news publishing and sharing platforms.

With the effects of this seen in the increased polarization of society, there is an increasingly growing trend amongst people who lack the ability to debate and relate to opinions which are not congruent to theirs. This is leading to a society where disengagement is the only peaceful measure but in-turn just divides the society in blocks where people inside the box are unable to see events from a different viewpoint, a viewpoint which may challenge their beliefs and position on a certain topic.

This phenomenon has been defined as the Echo Chamber, and the ever increasing proportion of existence of echo chambers among the society is contributed by different factors, firstly the social media recommendation systems play a huge role, wherein these algorithms across platforms recommend and suggest users related content based on the content consumption history and patterns.Secondly, with the general News consumption cycle all geared towards increasing engagement, there

is added value to the views and opinions rather than the actual value that is present in understand the subject and the facts related to the subject.

The existence of these Echo Chambers and the related effects can only be neutralised with the understanding of the type of news that one may be consuming on a daily basis and the added focus put on reading news articles on the same topic from different points of view.

Layered upon the ignorance of humans as a society and the existence of these Echo Chambers, there are the vices of the digital consumption pattern where the people are able to get news hands on, sometimes even as the event is taking place, and thus begins the inevitable race for the publishers to engage people as early as possible and for the consumers to form an opinion and give a reaction as early as possible and this means there are added scenarios wherein whole mobs of people can be left aggravated sometimes even just by reading a headline.

1.4 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. 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 patterns, perform sentiment analysis and then use this sentiment analysis output score to find patterns between the features of the NEWS data : Article, Headline, Date of Publish and Category

- Step 1: Web data scraping and collecting data
- Step 2: Data Pre-processing
- Step 3: Text data Pre-processing
- Step 4: Ground Truth Sentiment Labelling
- Step 5: Model Implementation
- Step 6: Model output evaluation and visualisation
- Step 7: Conclusion and future directions

1.4.1 Figures

There will hopefully be many figures in your thesis (that should be placed in the *Figures* folder). The way to insert figures into your thesis is to use a code template like this:

```
\begin{figure}
\centering
\includegraphics{Figures/Electron}
\decoRule
\caption[An Electron]{An electron (artist's impression).}
\label{fig:Electron}
\end{figure}
```

Also look in the source file. Putting this code into the source file produces the picture of the electron that you can see in the figure below.

FIGURE 1.1: An electron (artist's impression).

Sometimes figures don't always appear where you write them in the source. The placement depends on how much space there is on the page for the figure. Sometimes there is not enough room to fit a figure directly where it should go (in relation to the text) and so LaTeX puts it at the top of the next page. Positioning figures is the job of LaTeX and so you should only worry about making them look good!

Figures usually should have captions just in case you need to refer to them (such as in Figure 1.1). The `\caption` command contains two parts, the first part, inside the square brackets is the title that will appear in the *List of Figures*, and so should be short. The second part in the curly brackets should contain the longer and more descriptive caption text.

The `\decoRule` command is optional and simply puts an aesthetic horizontal line below the image. If you do this for one image, do it for all of them.

LaTeX is capable of using images in pdf, jpg and png format.

Chapter 2

Literature Review

2.1 Sentiment Analysis

Sentiment analysis is the use of NLP techniques for determining the emotions of the text data. This finds applications in fields where the language used needs to be labelled for certain categories of sentiment - positive, negative, neutral and this information can further help in determining the overall sentiment the writer or author or commenter wanted to convey. This helps businesses to get feedback about their product or service and understand the needs of the customer. This adds to the businesses abilities to review the large amounts of data they gather through reviews, comments, emails, messages and then structure the data for a deeper understanding of the overall sentiment pattern of the consumer. Where there are many ways to implement sentiment analysis, most of the process methods are constructed based on the output requirement such as requirement of fine grained analysis, subjectivity analysis, tone analysis and more.

Sentiment analysis problems on a machine level are perceived more like any classification problem, wherein features are extracted from the text data and fed into the classifier system, which simultaneously is fed pre-labelled class tags. The feature extraction is the process here of conversion of the text data into a numerical form of data. The techniques used can range from a simple bag of words technique where each word occurrence frequency is counted for and made up into a matrix to the n-grams technique wherein, the word pair combinations are taken into account before counting their occurrence frequency, these word pairs can be n words long. Other feature extraction methods include word embeddings, which not only determine the occurrence frequency of words or word pairs, but also establish a relationship between the words. These features are used as input to the classification model pipeline for statistical modelling.

2.2 Natural Language Processing

2.2.1 Overview

Using the manual techniques for determining the sentiments of each article is not scalable and hence we need to apply Natural Language Processing Techniques for sentiment analysis purposes.

NLP is a collection of techniques which allow the machine to process natural languages such as audio, text. With these natural languages being a vital model of communication for conveying of thoughts, ideas and emotions, there lies a vast amount of data for machines to work on and extract meaningful patterns and reach results

which via manual computation can be a very time consuming task.

Use of natural language for machine learning applications has proven a difficult task due to the lack of coherent rules present among different languages that people choose to communicate in. Thus NLP is based on the fundamentals of linguistics i.e rules of language but with advancement in new technologies, this has been worked upon to expand for various applications. Some of the more common applications of NLP include : Sentiment Analysis, Text Classification, Auto-Correction, Machine Translation, Text Summarization and many more.

Hence we can conclude that today's NLP applications for machine learning have come a long way from the initial attempts of linguistics analysis, computational linguistics and statistical natural language processing. Modern Natural language processing can be considered more as an agglomeration of linguistics and statistical language processing, with it retaining the vast linguistic knowledge, adding the computability of large amounts of data and the ability to perform complex statistical operations on the resultant data to compute necessary inferences.

Applied NLP allows use of many third-party library packages which have been carefully curated for use on text data eg. nltk. Moving ahead NLP has incorporated the deep learning techniques in methods for analysis of text data in order to obtain meaningful inferences as well as maintain an end-to-end system.

2.2.2 Architecture

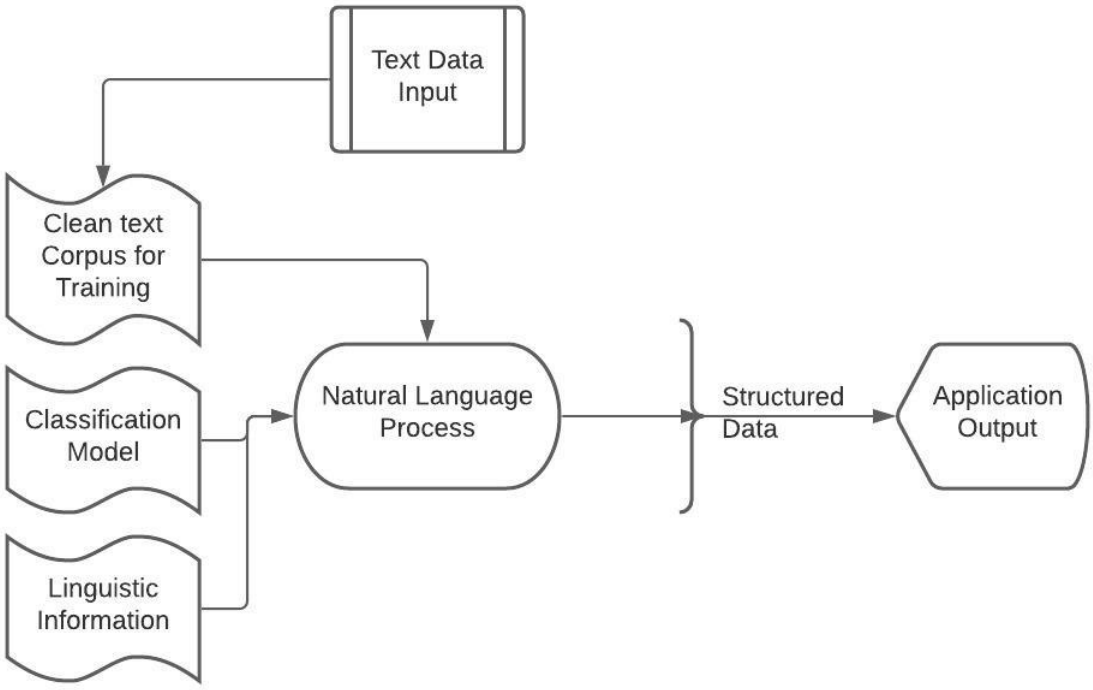

FIGURE 2.1: NLP Architechture

2.3 Natural Language Processing Techniques

2.3.1 K - Means Clustering

An unsupervised learning technique for classifying data containing n observations into “k” numbers of groups where k<n. Here the value of k is used to determine the number of clusters the data needs to be classified in.

The first step while processing raw data is to randomly select “k” data points which act as our initial clusters. This step is followed by the measurement of the distance of the first data point from each of the “k” initial clusters. This data point is assigned to the cluster that is nearest, i.e. the cluster which has the lowest distance value.

For data points existing on a single line, i.e 1-D data, distance is calculated, but for data points existing in 2-D, the euclidean distance between points is calculated and this distance is equal to the pythagorean distance between the two points and the centre of the cluster.

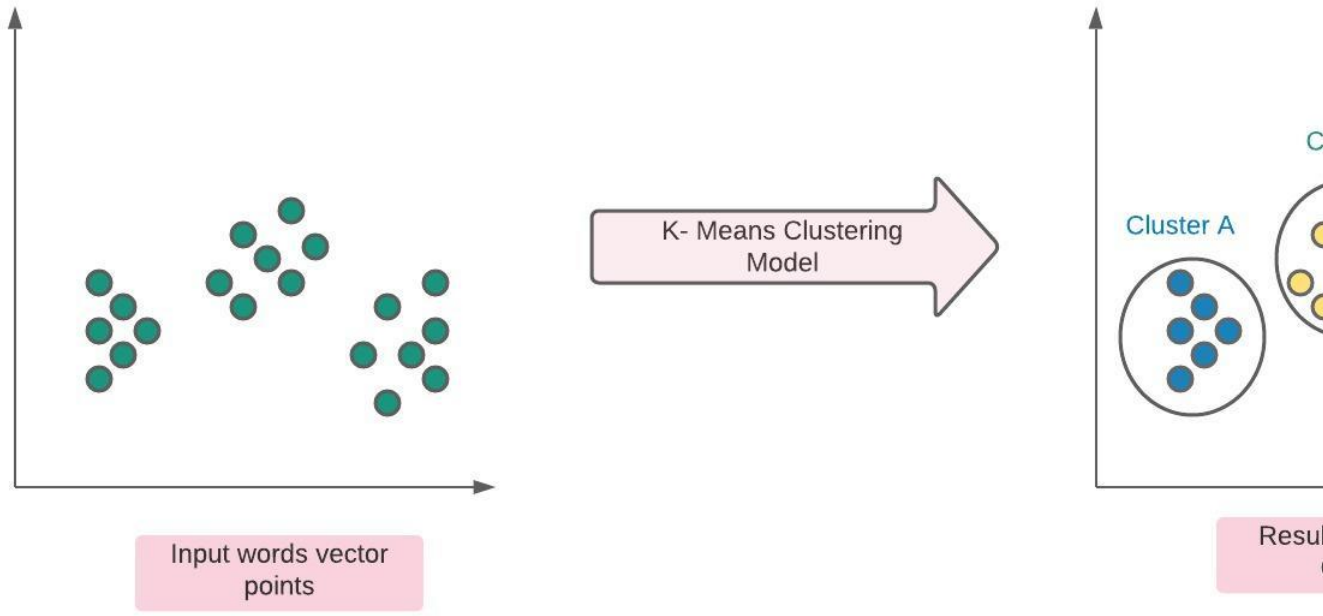

FIGURE 2.2: K - Means Clustering

Based on the different points of reference for distance calculation, the linkages can be classified into 4 types : Single Linkage, Complete Linkage, Average Linkage and Average group Linkage.

Similarly, the process is to be repeated for all the data points and is to be assigned to the relative closest initial cluster. The next step is the calculation of the mean distance of all the data points in each cluster. Now to verify the clustering map output of the given iteration, calculation of the variation inside of each cluster is essential. This step is required in order to remove any skew in the random selection of the

initial clusters. Further, more such iterations of clustering are observed for different data points as initial clusters and each is measured for the variation inside the initial clusters.

The model checks for the most balanced variation values for each clustering iteration for "k" number of iterations and then selects the best clusters. This value of k can be checked for until the variation in each clustering is equal to zero, meaning that there are as many clusters as there are data points, but for more efficient results, we can observe from the clustering vs reduction in variation from the Elbow Plot, wherein we can estimate the point after which the reduction of variation is minimal. This value of k will give the most efficient output for the k-means clusters.

2.3.2 Random Forest Classification

To serve the purpose of classification of data into a certain number of groups, this is a supervised learning technique, which requires pre-labelled data in order to further classify the data into different categories. Random forest classifiers belong to the top of the classifier hierarchy, and are a combination of the different decision tree classifier output. Thus, decision trees function as the building block of the random forest classifier and these individual decision trees combine as an ensemble to provide the required output.

The concept of combining the output of these relatively uncorrelated decision tree outputs is to get an aggregate bagging output and increase the overall accuracy and stability of the resultant prediction. Another advantage that is observed with random forest classification application is the ability of the algorithm to introduce randomness in the feature selection, wherein it doesn't just select the best features but instead searches for and chooses the best features from the randomly chosen subset of features. These random subsets are used for node splitting and can be further improved by implementing random thresholding.

Random forests also mostly allow the problem of overfitting, especially when a large number of features are to be taken into account. As random forests allow easy methods of feature selection, based on their importance and then the subset of these features is used as input to the multiple uncorrelated decision trees, thus, the model is able to eliminate overfitting to a larger extent.

2.3.3 LSTM [Long Short Term Memory]

Long - Short Term Memory is a type of supervised deep learning technique that uses Recurrent Neural Networks (RNNs) to establish long term token word dependencies. For applications wherein the input data is sequential or time-series data, then these models can provide an added advantage wherein the previous state is stored and used for text classification purpose.

These eliminate the problem in RNNs models, which can only hold short term memory and can thus, possess problems to store and use information from previous states to the new state. This is due to the vanishing gradient problem, wherein the gradient i.e values which update the weights in the neural networks are diminished as the sequences move ahead with time, and when the gradient value is extremely low, the model doesn't contribute anything new to the learning process.

LSTMs were introduced in order to cipher through the data, in order to filter the data which is necessary to store and neglect the rest of the data. These LSTMs, act as gates, where, the flow is the same as that of the RNNs but additionally allow ease of application and tuning, as they have three basic layers : Input layer, Hidden Layer and Output Layer and the hidden layers is the layer which consists of the different emory cells and the gate units which allow these recurrent networks to keep or neglect data as per application requirement.

To prevent overfitting with the LSTM models, there needs to be an accompanying dropout layer, which helps reduce the sensitivity of the model learning process to individual learning gates of the hidden layer in the LSTM network.

Also, an activation layer needs to be added after the output is generated from the LSTM network, in order to help in the interpretation of the output values as generated in numerical form. The activation function depends on the type of classification output needed, i.e in case of the binary classification, it is efficient to use the softmax activation function. The best method to choose the activation function is to test the various configurations and compare the resultant output.

2.3.4 BERT [Bi-directional Encoder Representations from Transformers]

These are state of the art ML algorithms, which allow us to perform NLP tasks.As the term suggests these collections of deep learning models, which are bi-directional in nature i.e. it can learn textual information from left to right and right to left.

This is a pre-trained on words chosen from the Bookscorpus and English Wikipedia corpus. Pretrained is based on contextualized word embeddings from these corpuses. These transformers are better at handling long term dependencies of the words as compared to the word embeddings. There are two different pre-trained models that can be accessed using Transformers - BERT base model and BERT large model. The base model allows use of 12 bi-directional heads and the large model allows the use of 24 bi-directional heads.

Transformers are used to detect, split and read entire sequences of tokens from a sentence all at once. This is unlike the LSTM model, where the tokens are read individually and unidirectionally, i.e. either from left to right or right to left. This is known as attention modelling, and can allow the model to get an innate understanding of the relationship between words in a sentence, e.g detect the nouns and the subsequent connecting pronouns. Transformers are transformer encoder stacks that can be pretrained.

To use the BERT model for sentiment analysis applications, we can either use the helpers that are provided by the Transformers library or we can just use the basic BERT model and then build a classifier on top of it.

BERT finds application in many text data based real world applications, such as question - answering, sentence classification, next sentence prediction etc. Some features to be kept in mind while handling BERT models is that it is better to pad input sequences in the right rather than on to the left of the input vector. Secondly,

BERT models can be trained as Masked Language Models(MLM) ,which are better at understanding the Natural language and the word dependencies, whereas the CLS(Sentence Level Classification) models are more efficient when used for sentence prediction.

2.4 Literature Overview

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

$$idf(t, D) = log|D|1 + |dD : td| \tag{2.1}$$

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 TF=Term Frequency

$$TF = termfrequencyindocument/totalwordsindocument \tag{2.2}$$

Let IDF=Inverse Document Frequency

$$IDF(t) = log2(totaldocumentsincorpus/documentswithterm) \tag{2.3}$$

Let t=Term

$$tfidf(t,d,D) = tf(t,d)idf(t,D) \tag{2.4}$$

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 the same vectors if they occur in the same way in the english text.That is, more than the relation between the words is the semantic 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 learn- ing model for a smaller dataset is not very significant when com- pared 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.

Chapter 3

Methodology Analysis

3.1 Dataset Overview

For this paper, we wanted to mine data from News Publication websites, in large amounts, in order to obtain a better perspective on the distribution of the sentiments across the features.

Data mining for websites is done through the web scraping and data crawling techniques. Our objective was to collect and tabulate the data from the different websites in order to further run text processing and text classification techniques and then label each instance with their subsequent sentiments.

It is essential to establish the exact features that we need to identify and extract data about from the websites. So keeping the aim of the project in mind, we have established that the data collected from the websites needs to have the news article, headline, publishing date, and the category the news was published under.

Now due to limited understanding of the 24 hour news cycle of the publishing websites of the different countries and the limited understanding of the society outlook on the news in general, we have preferred to target Indian News publishing websites which predominantly publish their articles in the English language as we have better tools in hand if and when we encounter peculiarities and obstacles.

3.1.1 Indian Digital News Scenario

The numbers that hit digital platforms across the board overwhelmingly show the eagerness and the shift in preferred mode of content consumption of Indian. Along with this, with the increasing popularity of the english language among the young users, and with 500 million new internet users since 2015, there has been an uptake in the people who prefer to get news from sources online - news publishing websites, social media sharing platforms and news aggregator platforms.

Now 56 per cent of youth below the age of 25 years in India, preferring to obtain news from online sources and 64 per cent of people who access the internet being moderately or highly interested in news content. There are just a significant number of people who are in ways affected by the news that is published on the news sites, which is then subsequently shared on the various social media sharing platforms and news aggregator platforms.

Now looking at the distribution of the readership and the popularity of the various top news publishing websites, we can see that these have engagement in numbers

that are larger than populations of any moderate sized countries.

With digital news publishers like, NDTV, The Times of India, BBC News, Hindustan Times and Republic News being in the top 10 preferred digital news websites, with upwords of 100M readership count statistics half-yearly.

3.1.2 Web Scraping

We can select websites from different spectrums if the political bias scales and thus have a better point of reference for comparison regarding the aims of our project. Having identified our geographical domain and the features we need to grab from the different news websites, we further need to go through the individual websites and see the structure of the websites to set up web scraping functions.

Each website has a different code structure and for each the scraping functions are customized to obtain the relevant content. We have to check for the code structure for where each feature information is present. What we find is that most of the websites have a limit on the amount of requests we can send for data retrieval beyond which they restrict the amount of data that can be scraped per day or even block the ip addresses from any further web scraping. This is circumvented by addressing the issue of data stored in cookies on the website and thus we can obtain the access to retrieve data in large scale to have a significant dataset, upon which we can perform the different machine learning and natural language processing techniques.

We have opted to use the spiders like possibility of the scrapers that are available under the *Scrapy framework*. We have opted for Scrapy over other tools such as Beautiful Soup and Flair as Scrapy provided significant advantages with respect to reusability rather than the XML and HTML parser tools offered by *BeautifulSoup* and makes it significantly efficient to scrape data from different News Categories and save it in a file. Additionally, we can take advantage of the Twister functionality provided by Scrapy, through which we can circumvent the blocking issues encountered on the websites and thus, we can use it to smoothly scrape data on a large scale, asynchronously for concurrency.

For data cleaning purposes, we have an added advantage using the scrapy tools as, we can easily choose to eliminate the

Once we have inspected the websites we can lock down our websites, which in our case was two Digital News Websites - NDTV News and Republic News. Navigating all the different urls in each website and further implementing spiders for extracting the relevant data into the feature columns, we are able to obtain data directly into CSV format file.

Once we have the dataset ready, we can read files into our notebooks, using the pandas library in python and perform further exploratory data analysis, which yields the below shown results.

We see that our spider was able to extract data into the following feature columns : Date Published, Article Text, Article Headline, Article Description, News Category and News Sub - Category. We choose to add two independent columns for Category and Subcategory after observations in the data collected to differentiate between the

News pertaining the different sub-categories which are further clubbed into world news and India news. From here on, for matters of categories, we will be using the data contained in the sub-category feature column as category information.

3.2 Model Pipeline

Firstly the research aims to establish the polarity of the news articles that appear on 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.

3.2.1 Text Data Pre-Processing

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.

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.

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.

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, still there will be a large number of repetitive tokens and will underfitting 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.

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 tokenization we can use the 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 tokens 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 purposes with the different text classification models.

3.2.2 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].

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.

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. 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.

3.2.3 Ground Truth Sentiment Labelling

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 -1 to +1 scale with -1 being negative, +1 being 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 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.

3.3 Model Comparison

Chapter 4

Key Findings and Outputs

4.1 Dataset Overview

4.2 Model Pipeline

4.3 Model Comparison

Chapter 5

Conclusion

5.1 Benefits

5.2 Conclusion

5.3 Model Comparison