VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering
VIT Chennai
Vandalur - Kelambakkam Road, Chennai - 600 127

Final Review Report

Programme: B.Tech – Computer Science and Engineering
Course: CSE4022
Slot: E1
Faculty: Dr. Premlatha M.
Component: J

Title: News summarization and Sentiment Analysis

Team Member(s):
Megha Nath (20BCE1581)
Shashank Sharma (20BCE1768)

Abstract

Around the globe, there are more than 6500 ‘daily’ newspapers, selling close to 400 million copies every day. Additionally, there are blogs, micro blogs, periodicals, magazines, fanzines etc. How can we make sense of all this information? How can we classify it and aggregate it so that we can perform quantitative analysis? This project explores one possible answer to these questions by classification of news articles by sentiment and topic and by summarizing it so that one can make sense out of the massive influx of news in these times.

Our vision is to create the capability to track how sentiment on a topic has evolved over time, how different news outlets cover the same topic and, in the limit, to be able to predict future behaviour through sentiment trends and summarise the news articles.

So, our primary objective in this project is to classify the news according to the sentiments it represents (positive, negative, neutral) using various other machine learning algorithms and classify the news according to the category it falls into.

Finally, for user convenience we have also summarized the given news article using Word2Vec and Seq2seq model.

Keywords

News, Text Summarization, Word2Vec, Seq2Seq

Introduction

People nowadays lead fast paced lives. The advent uses of news online social media such as articles, blogs, message boards, and news channels, and in general Web content has dramatically changed the way people consume information. The rate at which information is consumed has never been imagined before.

Events happening around the world are also recorded and distributed, in the form of news, at similar historically unprecedented speeds. For this very reason, humans need information in a manner that can be faster to comprehend and analyse. This begets that news be conveyed to people in a concise and lossless manner.

Today, many newspapers are published online. Some of them publish dedicated online editions, while others publish the pages of their print edition in PDF. In addition to newspapers, there are a wide range of opinionated articles posted online in blogs and other social media. The sheer amount of content that is put out via text, opens up the possibility of detecting positive or negative mentions of different types of organizations in the articles published online, thereby dramatically reducing the effort required to collect this type of information.

Before the era of Internet, the only way for an organization to track its reputation in the media was to hire someone for the specific task of reading newspapers and manually compiling lists of positive, negative and neutral references to the organization, it could undertake expensive surveys of uncertain validity and sentiment analysis would help with that.

Also due to huge availability of text in numerous forms, a lot of unstructured data has been recorded by research experts and have found numerous ways in literature to convert this scattered text into defined structured volume, commonly known as text classification.

Therefore, we have identified two needs in terms of web documents online:

- Text classification: Identifying the notion of a piece of text, whether it has positive or negative connotations, or they are neutral

- Summarization: Presenting news articles in a summarized way which is concise, lossless and scannable.

To resolve these issues we employ multiple ML models to classify the news articles among two degrees:

a) Positive

b) Negative

Next, for summarization we employ Word2Vec and Seq2Seq models to obtain a summary for our news articles.

Acceptable Summaries will have the following characteristics:

Concise: Information should be short. Information provided should be as short as possible so that it takes a human less time to read. An average human reads at a pace of 2 minutes per page. This requires content of the summary to be short so as information can be gained faster.

Lossless: No loss in terms of information or knowledge. Information should be able to convey the totality of the event that occurred. Any information missed out would mean unsuccessful summarization. A proper balanced trade off needs to occur between the shortness and losslessness of the summary.

Scannable: Information should be presented with more keywords. No keywords should be missed out. Keywords help humans read a body of text faster. It happens because upon reading the keywords the brain can fill out the context on its own, in turn reducing the time taken to comprehend written information. Thus summary should have a balanced trade-off between scannability and shortness.

Literature Review

I) Automated Classification of Societal Sentiments on Twitter With Machine Learning

In this research paper, the authors have developed an automated framework to understand positive, negative and neutral sentiment extracted from tweets on the social media platform, Twitter. The results obtained from the framework's analysis helps us understand societal sentiments during profound events like the COVID-19 pandemic. The framework is hybrid as in it uses both deep learning and machine learning methods for tweet classification. The authors' main contribution include the following:

a) A novel hybrid model which combines state-of-the-art deep learning techniques for tweet classification. The model makes use of lexicon based automated tweet sentiment analysis in combination with deep learning methods.

b) A methodology to evaluate the ML techniques using various evaluation measures.

The authors aim in developing the model was to understand how people online feel about imposed government policies, regulation and guidelines in high tense global emergencies like COVID-19.

According to the authors, VADER is the most used tool for sentiment analysis on social media sites. VADER is a rule-based model which was developed especially for sentiment analysis using social media messages because of its impeccable ability to recognize and understand the sensitivity of messages. VADER works by providing a polarity score for positive, negative and

neutral classes and also gives a compound score. This compound score can be used as a threshold value to distribute the messages between all the three classes. VADER was also chosen as it relies on a dictionary of lexicons and their associated sentiments. This enables VADER to extract sentiments based on the contextual meaning of a phrase. The authors suggest scaling up

VADER to process the sentiments of enormous volumes of tweets from the site would not be sufficient, hence they were motivated to amalgamate VADER with ML techniques. The model employs unsupervised learning, particularly employing a lexicon based approach that associates the three classes (positive, negative and neutral) with words. The machine learning techniques used were Gaussian Naive Bayes, MLN Naive Bayes, Decision Trees, Random Forest, Logistic Regression and Long Short Term Memory (LSTM). The models were evaluated on the following basis: Accuracy, Recall, Precision and F1 Scores.).0.5 was the threshold for the compound score wherein the following scores were the following classes:

a) Compound Scoresample >0.05 => Positive

b) Compound Scoresample < = -0.05 => Negative

c) - 0.05 < Compound Scoresample <0.05 =>Neutral

The authors also employed a five fold cross validation technique to fine tune the hyperparameters. The results showed that the COVID-19 tweets were lengthier than usual reaching the historic 140 character limit. The high character count was attributed to the gravity and complexity of the COVID-19 situation which led them to express their sentiments in a rich manner. This also begets a need for ML for tweel sentiment analysis. Most tweets were either positive or neutral suggesting a positive attitude towards the situation. Words related to positive sentiment were "help", "mask", "need", "people", and "thank". Most negative tweets were related to death and words relating to the former U.S. President Donald J. Trump was also used negatively suggesting a political dimension to the sentiments of COVID-19. The word "amp" (ain't my problem) was flexibly utilized to express all three sentiments.

II) Reinforced Abstractive Text Summarization With Semantic Added Reward

In this paper the authors' aim was to improve the quality of summary statements using a reward function usually used in text summarization based on Reinforcement Learning. The dataset used was the Gigaword dataset. The authors have recognized that the recent summarization models are based on sequence-to-sequence neural networks which are made up of encoders that understand the input sequence and decoders that generate the output sequence. The authors identified the following problems with using sequence-to-sequence neural networks:
1) Pov (Out of Vocabulary) problem
2) Generation of a particular word repeatedly
3) Exposure bias at test time
4) Non-optimized learning for Evaluation metrics
The authors have also identified that ROUGE-L cannot be used as a reward function as it is not completely free from the bias problem because ROUGE uses the same vocabulary as the reference summary and the generated summary in the same order.
The authors contributions are as follows:
1) ROUGE-SIM: It is a metric that is a modification of the ROUGE metric based on n-gram match. When the longest common subsequence is calculated, the similarity between the word embeddings should be used to allow matching of similar words.
2) ROUGE-WMD: This reward function mixes Word-Mover's Distance and ROUGE-L.
This model allows a selection of a diverse vocabulary, such as abbreviations or similar words. Repetition was reduced leading to high robustness of the model. The grammatical errors were evaluated manually using GRAMMARLY (online tool). Most of the grammatical errors were caused by repetitive problems, which improved significantly without post processing.
The model proposed by the authors is based on a simple LSTM (long short term memory) sequence-to-sequence model with Attention, a pointer mechanism that handles Out of Vocabulary (OOV) words, and an intra decoder attention that handles repeated words. The sequence-to-sequence attention module uses a single layer bidirectional LSTM as the encoder and a single layer LSTM as the decoder. The intra-decoder attention module was built to prevent the decoder from generating a repeated phrase. This module proceeds with creating a context vector containing information about the sequence decoded in the previous timestamp during the word generation process. By using this context vector, the decoder when generating a word can use the information about previously generated words in the decoding process minimizing repeated phrases. The pointer generation module solves the OOV problem by copying unseen or rare words which cannot be generated from the input sequence. The learning employed is policy learning. Policy gradients using Reinforcement learning can solve metric optimized learning and bias problems.
Reward Functions:
ROUGE-L: ROUGE is an indicator for evaluating natural language generative models such as text summarization and machine translation, measuring performance through comparisons with answer sets. ROUGE-L is computed using Longest Common Sequence of the ground truth summary and generated summary. In the text summarization task, ROUGE-L's precision is the shortness of the generated summary and its ability to recall how much information the generated summary contains from the reference summary.
ROUGE-SIM: ROUGE-SIM does not reject the length of Longest Common Subsequence for words with a similarity score of 0.8 or more to words appearing in the text. By allowing semantic matching, the penalty for using similar words and overcoming the lower abstractiveness problem and the bias problem that could occur during the learning process is reduced.

ROUGE-WMD: Word Mover Distance measures the semantic distance between word documents using word embedding. Excluding stopwords, a document can be seen as a distribution of words.
The semantic distance is the cost of moving all of one distribution to another, i.e. , the semantic distance between two documents is the minimum of the cumulative sum of the distances between words from each document.
The outcomes were measured quantitatively using F1 scores for ROUGE1, ROUGE2 and ROUGE-L. The abstraction of the text summary was measured using novel n-grams. Novel n-grams are indicators of the percentage of n-grams in the generated summary that are not included in the input sequence. ROUGE-WMD has a lower value in 4-gram novelty than ROUGE-L, but shows better expression in 1,2 gram. The frequency of using abbreviations or synonyms not included in the input sentence was high. ROUGE-WMD was higher in 4-gram novelty than ROUGE-SIM.

III) **Fact-Driven Abstractive Summarization by Utilizing Multi-Granular Multi-Relational Knowledge**

In this research paper, the authors have identified that the limiting factor for existing solutions of text summarization techniques is that they are primarily concerned about extracting facts from the source text and tend to overlook important facts like locations, times, consequences, purposes, participants etc. Also they inadequately model complex semantic relations among facts and the corresponding factual information. These are some areas of improvement the authors have recognized and thus developed a model called FFSum to mitigate them. Recent works tackle the problem of generating summaries containing utterly different semantics and meanings by employing a fact-driven strategy. The idea is to extract factual pieces from source text and then encoding them into the summarization framework to improve the generated summary. These works tend to consider only the coarse granular factual pisces and overlook the fine grained facts embedded in such texts. More detailed information in a precise summary should be composed of a multitude of fine granular pieces of information since events/facts typically come with their arguments.
The existing models tend to ignore these multi-granular facts and produce imprecise summaries that may confuse end-users. This paper aims to introduce a new fact-driven model for summarization.
FFSum aims to extract the fine grained facts from text as most recent works do, but then use those facts and facets to construct a factual graph.
FFSum is implemented upon BART, a state-of-the-art sequence-to-sequence summarization framework. FFSums uses the BART's checkpoints to warm-start its own generation framework.
It optimizes BART by using a fact-driven graph attention network that integrates multi granular facts representations at the encoding stage. FFSUM further employs a hybrid pointer (Ptr-Net) in the decoder for abstractive summarization. The hybrid pointer allows the generation framework to retrieve fact and facet knowledge from the factual graph and copy faithful tokens from the source article. Thus, by including multi granular factual pieces of information, FFSum provides richer context that boost factual correctness.
The datasets used are:
A. CNN/Daily Mail
B. BBCXSUM
The paper's contributions are as follows:
a) Exploiting multi granular factual information for precise text summarization
b) A fact-driven graph attention network that integrates factual information into summarization effectively.

c) Shows how the graph encoding and hybrid pointer networks can be combined to collect multi-granular factual information for better text summarization.
StuffIE was utilized to extract facts and facets. StuffIE was used to find the coreference resolution between facts and facets. The coreference resolution provides nested relationships between facts and facets. Then, the factual graph is constructed using FACT LABEL and FACET LABEL. The FACT LABEL edge is used to connect the subject, object and predicate phrases in
a fact triple. Connect three nodes of a fact using edges ‘sbj’ and ‘obj’. The FACET LABEL is used to link facts with its facets. The FFSum encoder consists of two modules, BART encoder which learns text tokens’ representations and FGAT to learn graph nodes’ representations. The decoder contains two modules as well, BART decoder to learn target representations and Hybrid Ptr-NEt to copy source tokens and retrieve graph nodes for generation.
The paper also introduces its own factual correctness evaluator FFCC which can do the following:
a) Explicitly identify multi granular factual consistency in which both facts and facets are present.
b) FFCC focuses on factual consistency on phrase-level multi-granular factual information beyond entity tokens.

Dataset Description

1. For text classification based on sentiment analysis: Scraped data

2. For text classification based on category: https://www.kaggle.com/rmisra/news-categorydataset

3. For text summarization: Scraped data

Pre processing
Removing non-alphabetic characters, replacing URLs, remove multiple spaces, remove any single characters hanging between 2 spaces, ordering regex etc. For word2vec we also used bigram
There are 25654 columns in our Scraped data
Our dataset contains 6 attributes and 2,10,294 records

headlines	text	date	author
Massive fire at Apple supplier Foxlink's Andhra facility, production halted	Apple supplier Foxlink has halted production at its assembly f	27 Feb 2023,Monday	Pragya Swastik
BJP+ to retain Tripura and Nagaland, hung House in Meghalaya, predict exit polls	BJP along with its allies is set to retain power in Tripura and N	27 Feb 2023,Monday	Apaar Sharma
Buildings collapse after fresh 5.6-magnitude earthquake in Turkey	A fresh 5.6-magnitude earthquake shook southeast Turkey on	27 Feb 2023,Monday	Pragya Swastik
4 killed as pickup van hits motorcyclists, pedestrian in Gujarat	Four people, including a woman, died after they were hit by a	27 Feb 2023,Monday	Shreyasi Banerjee
Father-in-law of Asaduddin Owaisi's daughter dies by suicide in Hyderabad	Dr Mazheruddin Ali Khan, the medical superintendent at Owa	27 Feb 2023,Monday	Apaar Sharma
Farmer gets â‚¹0 profit after selling 825 kg of onions in Maharashtra, pic of receipt viral	A farmer named Bandu Bhange got â‚¹0 profit after selling 825	27 Feb 2023,Monday	Anmol Sharma
Raccoon, tegu lizards found in passenger's bag at Chennai airport	Customs officials at Chennai airport have seized a raccoon an	27 Feb 2023,Monday	Shreyasi Banerjee
UP govt to fine upto â‚¹2,000 for not segregating waste from Mar 4	Uttar Pradesh government will fine people who do not segreg	27 Feb 2023,Monday	Ashley Paul
â‚¹10.42 lakh cr wealth of investors wiped off in 7 trading sessions	Seven consecutive sessions of decline in the equity market en	27 Feb 2023,Monday	Pragya Swastik
I've mostly been offered cop roles since I cut my hair: Mandira	Mandira Bedi revealed that ever since she cut her hair, she ha	27 Feb 2023,Monday	Udit Gupta
It's a no from my side: Mrunal as fan asks her to marry him	Mrunal Thakur replied to a fan who proposed marriage to her	27 Feb 2023,Monday	Udit Gupta
Anupam Kher's dad had a very unique personality: Vivek Agnihotri	Filmmaker Vivek Agnihotri presented the Viewers' Choice Best	27 Feb 2023,Monday	Amartya Sharma
Was payback for poor fellow: Arbaaz on SRK's 'Pathaan' success	Actor Arbaaz Khan has said that the success of 'Pathaan' was	27 Feb 2023,Monday	Amartya Sharma
CBSE alerts students against rumours about question paper leak	The CBSE has alerted students and parents against rumours a	27 Feb 2023,Monday	Deepika Bhatt
Chip shortage has eased, not fully over: Hitachi Energy India CEO	Hitachi Energy India CEO N Venu has said that the semicondu	27 Feb 2023,Monday	Aishwarya Awasthi

Table 1. Scrapped Data

Implementation

Our work presents various Machine Learning (ML) approaches and algorithm comparisons for of texts and for doing sentiment analysis efficiently.
The text is classified based on three classes' positive, negative and neutral classes. This work suggests that it is efficient to use naïve bayes classifier for the purpose of sentiment analysis. As it gives better accuracy as compared to other classifiers used for sentiment analysis.
Our project also provides a method to summarize the news articles using **Word2Vec** and **Seq2seq** model.

Cleaning data

Firstly, lemmatization then since the data was scaped it required a lot of cleaning i.e. removing all the rows for which the data was in the wrong column, rows for which the URL could not be reached and rows with NA values.

Pre processing

Removing non-alphabetic characters, replacing URLs, remove multiple spaces, remove any single characters hanging between 2 spaces, ordering regex etc. For word2vec we also used bigram

Text classification based on sentiment analysis

A) Stemming and Lemmatization of Title and Content column

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.
Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

B) Categorize the Title column into positive, negative, and neutral headlines.

C) Vectorization

D) Sentence Segmentation

i) Tokenization

ii) Lemmatization

iii) Removing Stop Words

iv) Vectorization

Libraries used:

- Nltk
- Pandas
- numpy
- Seaborn

- Matplotlib

Text summarization
The dataset scraped using beautiful soup consists of around 11 thousand examples and contains Author name, Headlines, URLs of Article, Short text, Complete Article. We gathered the summarized news from Inshorts and only scraped the news articles from the various major news sites from which the data was taken. Time period ranges from January to February 2023.
Summarization using Word2Vec and Seq2seq model

Working Models
News Classification using Sentiment Analysis

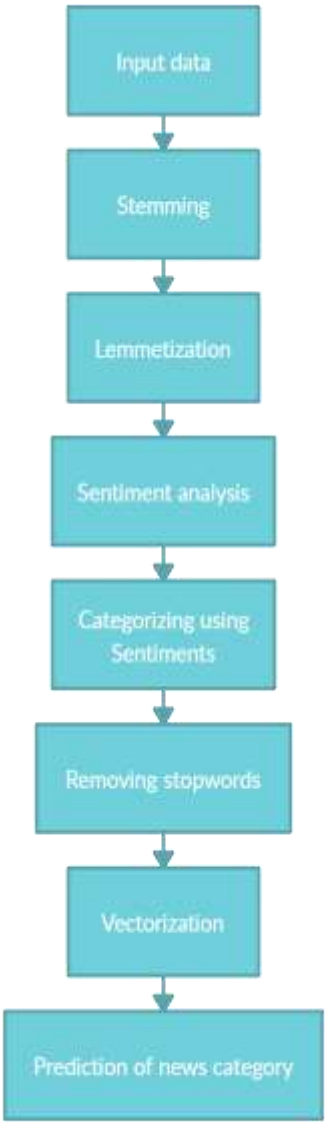

Figure 1. Sentiment Analysis Workflow

News Summarization

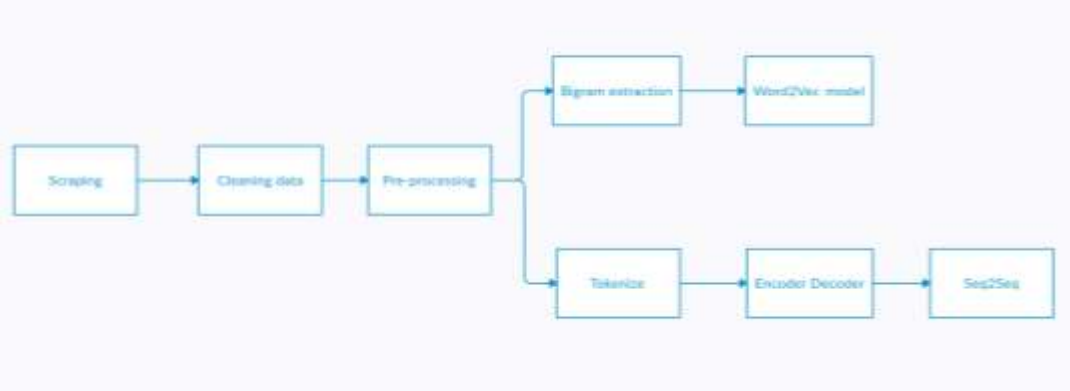

Figure 2. News Summarization Workflow

Models used:

- Naïve bayes algorithm

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

- XGBoost

XGBoost is short for Extreme Gradient Boost. Unlike Gradient Boost, XGBoost makes use of regularization parameters that helps against overfitting. XGBoost only accepts numerical inputs. Therefore, it will be up to us ensure the array type structure you pass to the model is numerical and in the best cleansed state possible.

- MLP

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN).
The term MLP is used ambiguously, sometimes loosely to any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons.

- Word2Vec model

Word2Vec is a method to construct word embedding. It can be obtained using two methods (both involving Neural Networks): Skip Gram and Common Bag Of Words (CBOW)
CBOW Model: This method takes the context of each word as the input and tries to predict the word corresponding to the context.
Skip Gram Model: We input the target word into the network. The model outputs C probability distributions.
Skip Gram works well with small amount of data and is found to represent rare words well. On the other hand, CBOW is faster and has better representations for more frequent words.

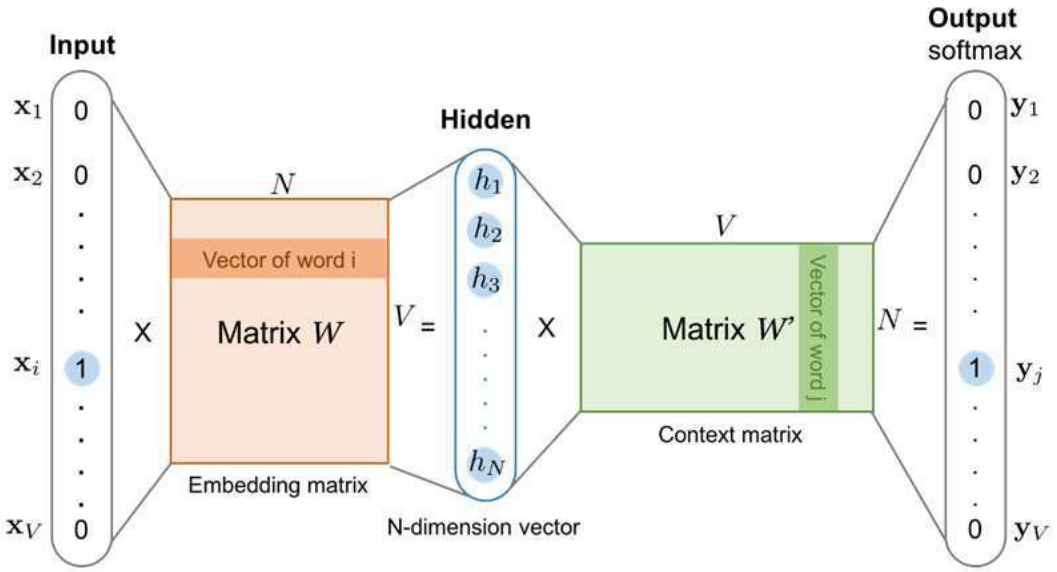

Figure 3. Word2Vec Model

- Seq2seq model

Seq2Seq is a method of encoder-decoder based machine translation that maps an input of sequence to an output of sequence with a tag and attention value. The idea is to use 2 RNN that will work together with a special token and trying to predict the next state sequence from the previous sequence.

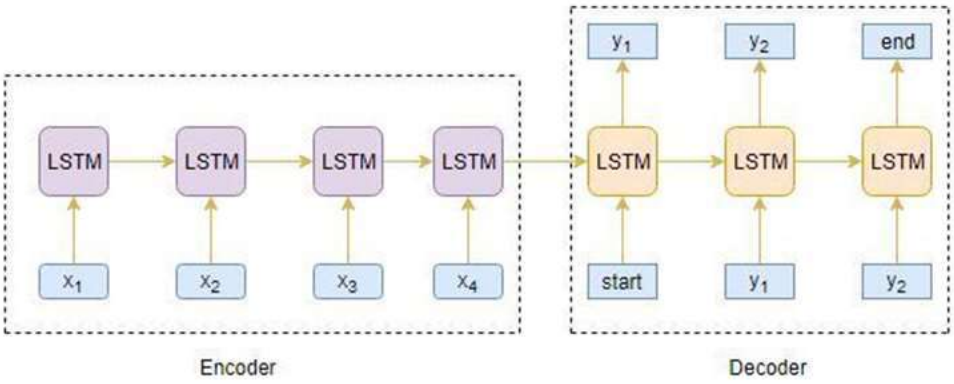

Figure 4. Seq2Seq Model

Results

The scrapped data used showed that the ‘Daily Mail’ released the most number of articles during our considered time period. It is also revealed that the ‘Entertainment’ category is the most popular due to the sheer number of articles under this category. Also, the word that appears most of the times in our considered timeframe of news articles is the word ‘google’.

During sentiment analysis, we found that news titles are usually neutral as opposed to content which is classified as positive mostly. Negative titles are the least common. Similarly neutral content are also the least common.

For news classification, XGBoost gives the best accuracy which was finally used to classify the news titles and articles.

For text summarization, 100 worded paragraphs on average could be summarized in 15 worded sentences.

Pre Processing

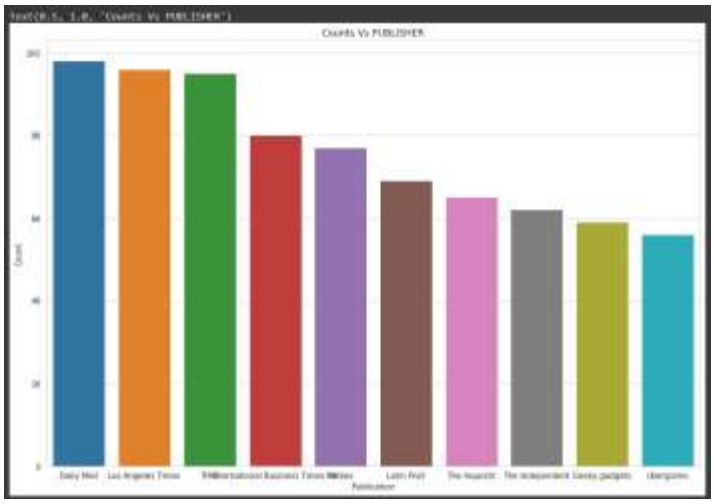

Figure 5. Publication Counts

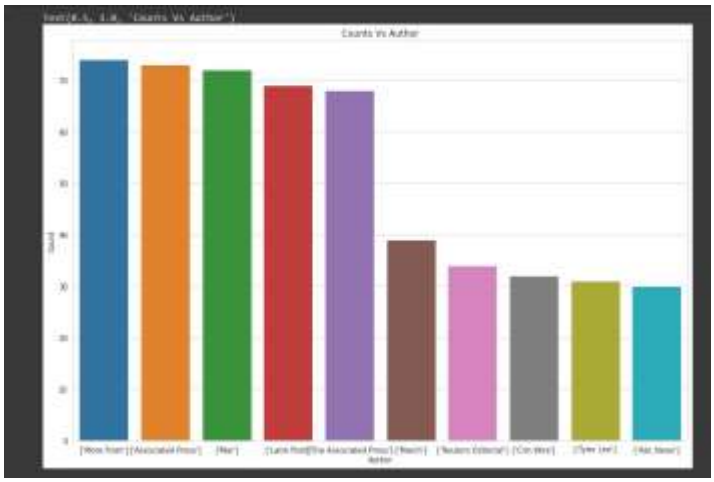

Figure 6. Counts of publications of each author

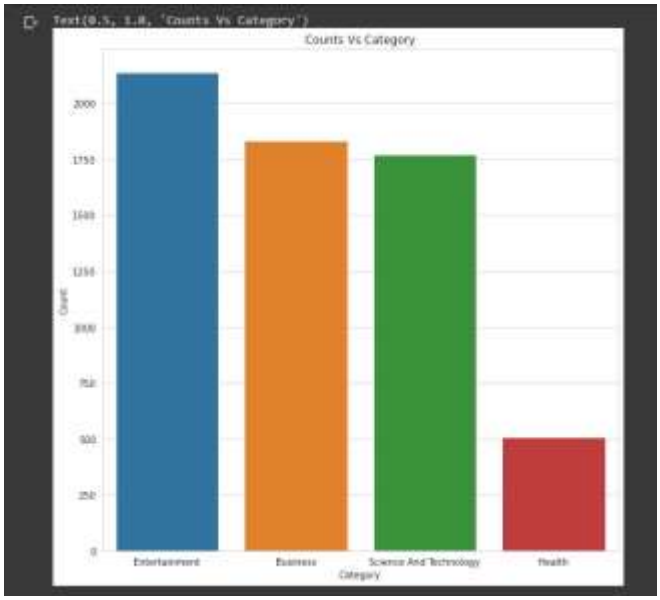

Figure 7. Counts of publications under each category

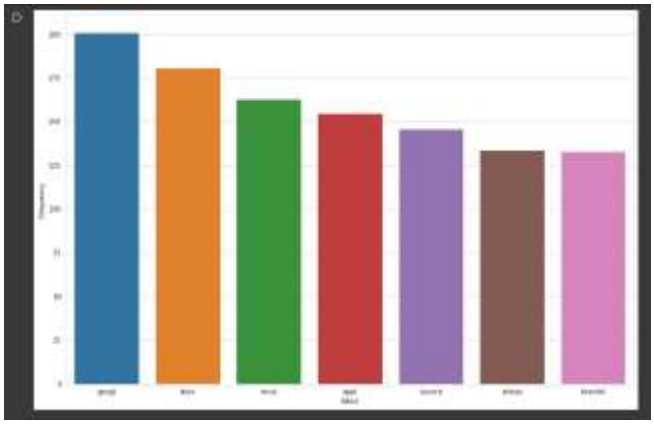

Figure 8. Frequencies of occurrence of each word

Sentiment Analysis

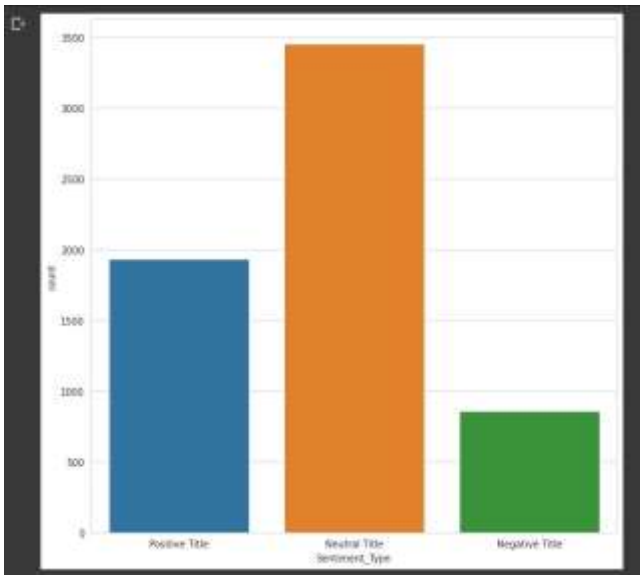

Figure 9. Frequencies of titles under each Sentiment Type

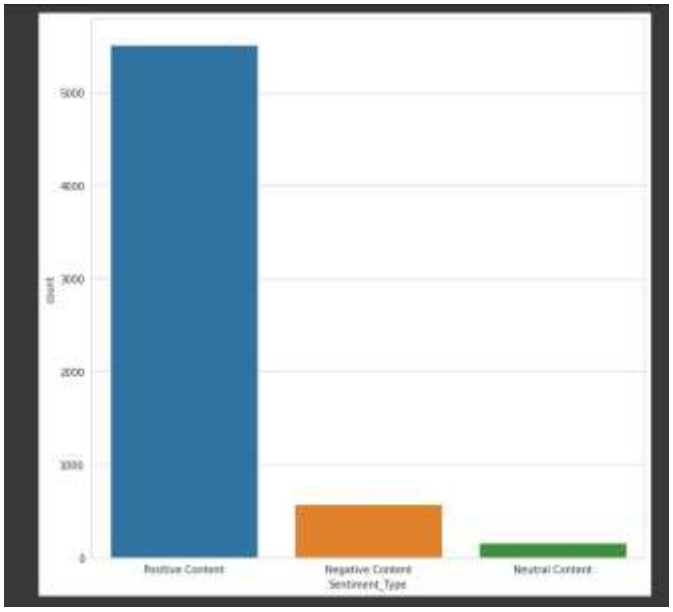

Figure 10. Frequencies of content under each Sentiment Type

News Classification

	Naïve Bayes	**MLP**	**XGB**
Train	0.97316	1.0000	0.86847
Test	0.85101	0.85976	0.82984

Table 2.Comaprison of accuracies of each model

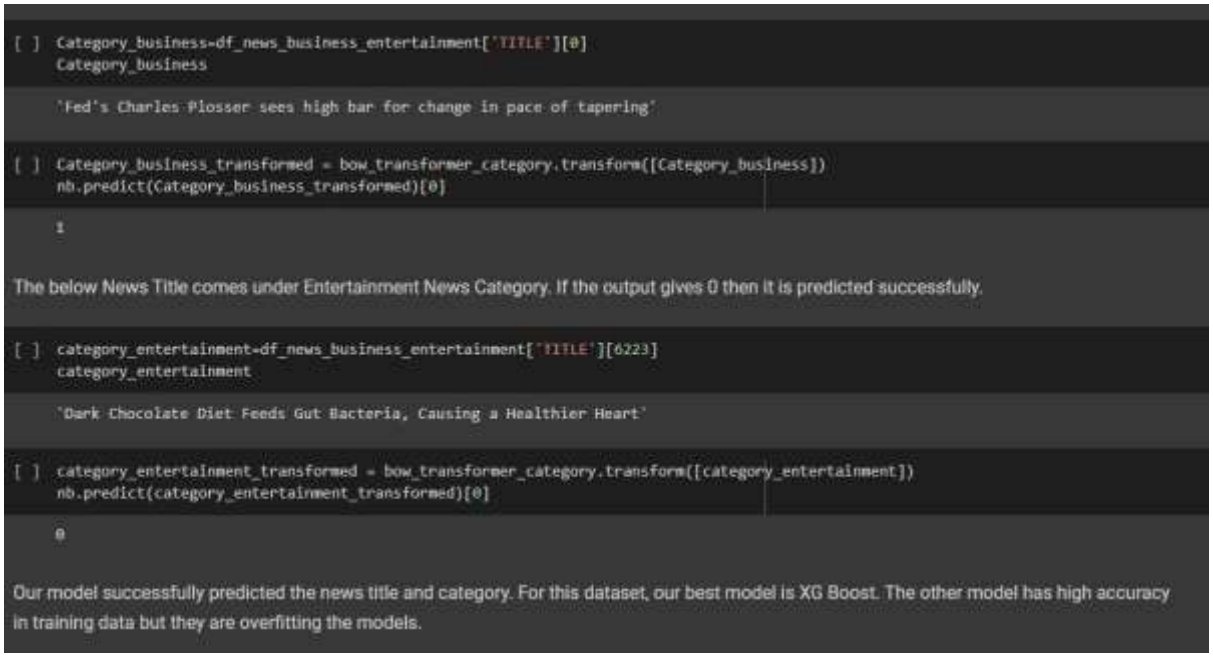

Figure 11. Classification of test set

Text Summarization

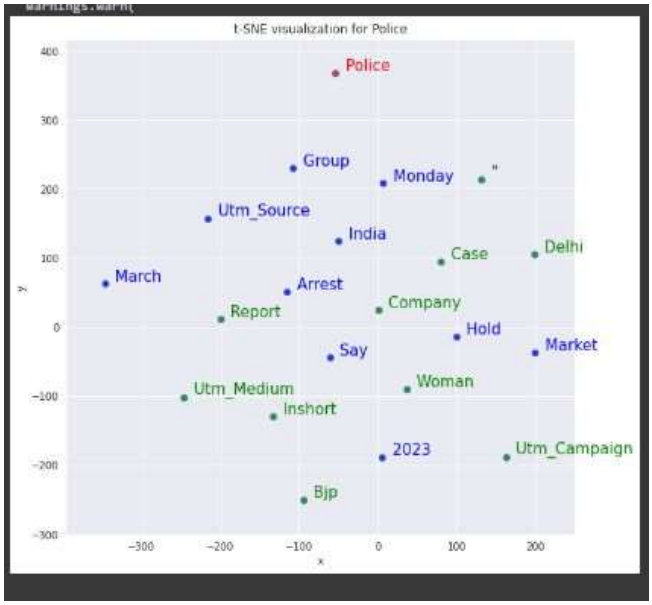

Figure 12. t-SNE visualization for Police

Figure 13. Word Cloud

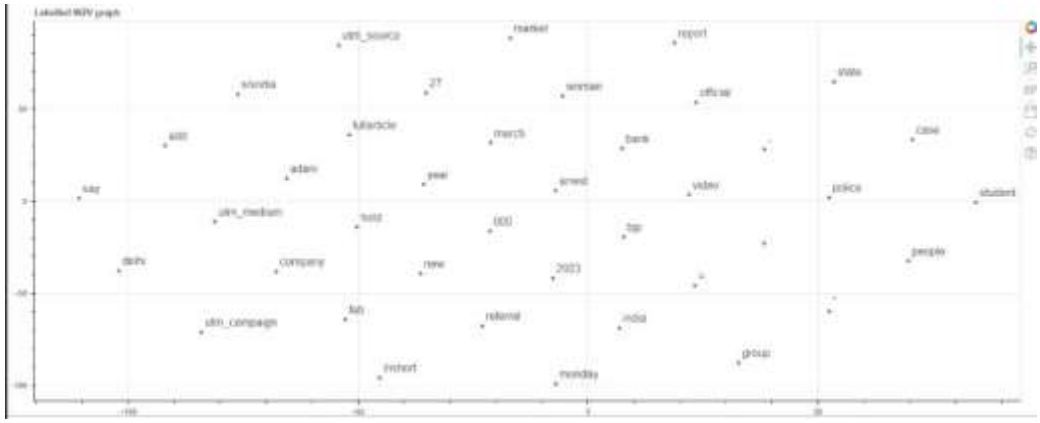

Figure 14. Bokeh Plot

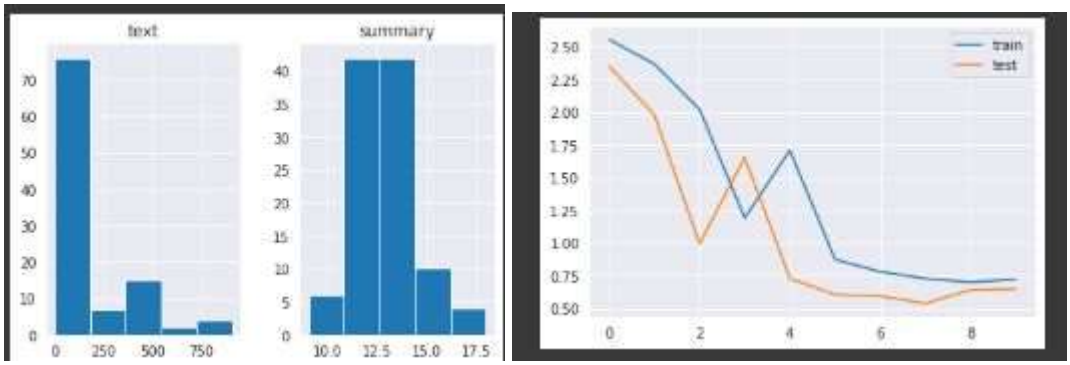

Figure 15. Model output and model accuracies

Figure 16. Summary outputs

Conclusion

The field of sentiment analysis is an exciting new research direction due to large number of real-world applications where discovering people's opinion is important in better decision-making. The development of techniques for the document-level sentiment analysis is one of the significant components of this area. Recently, people have started expressing their opinions on the Web that increased the need of analysing the opinionated online content for various real-world applications. A lot of research is present in literature for detecting sentiment from the text. Still, there is a huge scope of improvement of these existing sentiment analysis models. Existing sentiment analysis models can be improved further with more semantic and common-sense knowledge so using those models would definitely help In this project we explored various ways in which we can perform sentiment analysis, classification and summarization generating one-sentence summarization that mimics the style of news titles given some paragraphs.

We built multiple models for sentiment analysis and classification. Naive Bayes and MLP models were overfitted. Our best model for news category prediction was XG Boost. Since, there was no much difference between training and testing data accuracy of XGB model. But in the other cases, the training accuracy is very high and testing is very low, therefore these models will not give best results on the unseen data.

For summarization we managed to build word2vec model to produce word embeddings and seq2seq model to provide a one line summary for the given piece of news which performed fairly well on most of the test cases but still wasn't as good as a human summary.

Future Work

The field of sentiment analysis is an exciting new research direction due to large number of real-world applications where discovering people's opinion is important in better decision-making. The development of techniques for the document-level sentiment analysis is one of the significant components of this area. Recently, people have started expressing their opinions on the Web that increased the need of analysing the opinionated online content for various real-world applications. A lot of research is present in literature for detecting sentiment from the text. Still, there is a huge scope of improvement of these existing sentiment analysis models. Existing sentiment analysis models can be improved further with more semantic and common-sense knowledge so using those models would definitely help

For summarisation we can use other language models and explore other new and upcoming summarization techniques like Attention-based Deep Recurrent Neural Networks, BERT etc.

Apart from this we can also explore an implementation of this project in real time which would take news articles in real time and apply all the different techniques in this project on them and give user all the output in real time along with a recommender system which recommends similar articles.

References

1. *Vyas, P., Reisslein, M., Rimal, B. P., Vyas, G., Basyal, G. P., & Muzumdar, P. (2021).*
2. *Automated classification of societal sentiments on Twitter with machine learning. IEEE*
3. *Transactions on Technology and Society, 3(2), 100-110.*
4. *Jang, H., & Kim, W. (2021). Reinforced Abstractive Text Summarization With Semantic*
5. *Added Reward. IEEE Access, 9, 103804-103810.*
6. *Mao, Q., Li, J., Peng, H., He, S., Wang, L., Philip, S. Y., & Wang, Z. (2022). Fact-Driven*
7. *Abstractive Summarization by Utilizing Multi-Granular Multi-Relational Knowledge.*
8. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1665-1678.*
9. *Ramesh Chandra Belwal1 · Sawan Rai2 · Atul Gupta Extractive text summarization using clustering-based topic modelling Accepted: 22 September 2022 © The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022*
10. *Prashant Raina School of Computer Engineering Nanyang Technological University Singapore prashant007@e.ntu.edu.sg Sentiment Analysis in News Articles Using Sentic Computing 2013 IEEE 13th International Conference on Data Mining Workshops*
11. *Wataru Souma1 · Irena Vodenska2,3 · Hideaki Aoyama4,5 Enhanced news sentiment analysis using deep learning methods Journal of Computational Social Science (2019) 2:33–46*
12. *D. Rajagopal, E. Cambria, K. Kwok and D. Olsher, "A graph-based approach to common-sense concept extraction and semantic similarity detection," in Proceedings of the 22nd International Conference on World Wide Web Companion, Rio de Janeiro, 2013*