JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

Multi-Label News Category Text Classification

Shilpa Patil
Department of Studies in Computer Science,
Vijayanagara Sri Krishnadevaraya University,
Ballari, India.
Dr. V Lokesha
Department of Studies in Mathematics,
Vijayanagara Sri Krishnadevaraya University,
Ballari,India.
Dr. Anuradha S G
Department of Computer Science and Engineering,
Rao Bahadur Y. Mahabaleshwarappa Engineering College,
Ballari,India.

Received:2022 March 15**; Revised:**2022 April 20**; Accepted:**2022 May 10

Abstract

News classification in the present day is a challenging research problem due to availability of huge digital data demanding the accurate classification for user easy. The world is witnessing the transformation in digital age giving rise to change in every aspect of human life. Each news category arranges the news story prior to distributing it. So every time guests visit their site, can without much of a time find news of their interests. Presently, the news stories are ordered by hand by the substance chiefs of information sites. The aim of our article is to build a text classifier for news category and further analyze the sentiments of text-based headlines. The Experimental study is carried on 10 News categories available in dataset ('**News_Category_Dataset_v2**' from Kaggle). The results reveal sentiment associated with news categories as polarity index with positive or negative values.
Keywords: News Category, Kaggle, polarity index, dataset, headlines**.**

I. Introduction

During these days, Internet has become one of the most important sources for various news information. The advanced technological development of ICT has made tremendous collection of data in social media like face book, Twitter, etc., for different entity-based. Thus, Automatic text classification using Natural Language Processing (NLP) is gaining popularity among researchers. NLP is a fundamental task for text classification. This text classification is a method for categorizing the text-based entities into some predefined categories or labels. Each label either belong to multiple or exactly one or no category (topic or class) at all.

The major heritage of information is on news articles available on different websites which are broadly classified on various categories. This news article provides in search of a particular news category as it allows timely and efficient information retrieval. Many researchers have done tremendous work on News classification to develop text classifiers. Researchers have contributed their study in the field of Business, Sports, Technology, Politics, Entertainment, Health and Economy, on different Blogs, using Hash tags for news categories and so on.

NLP (Natural Language Processing) studies interactions between system thinking and human languages, in the sense of how computer programs, process and analyses most natural language data. Using NLP, we classify textual data. In Text classification, assign predicted categories for textual data according to its content. In this article, we explore methods to analyze textual data and extract features to build classification model for the subset of 10 categories: CRIME, TRAVEL, WOMEN, BUSINESS, COMEDY, POLITICS, SCIENCE, SPORTS, MEDIA, RELIGION from news category dataset.

II. Related Literature survey

Hussain et al., [2020] [14] In this article "Design and Analysis of News Category Predictor" with standard dataset of British Broadcasting Corporation (BBC news) consists of five categories - Business, Sports, Technology, Politics and Entertainment. The four multiclass news category predictors on same dataset are evaluated and compared i.e., Naïve Bayes, Random Forest,

K-Nearest Neighbours (KNN) and Support Vector Machine. Different categories are evaluated by analysing confusion matrix and also measure test dataset by precision, recall and overall accuracy. As results have shown adequate accuracy. However, SVM model is better than 4 supervised learning models (98.3%accuracy) whereas lowest accuracy i.e.,KNN model. **Pelicon et al.,[2020] [13]** In this article "Zero-shot Learning for Cross-Lingual news sentiment classification" is to classify news for a given dataset with positive, negative and neutral news to slovene news with sentiment category and also news in other language without data training. This system is based on multilingual BERT model and evaluated on news sentiment testset in croatian. This experiment shows improved results in monolingual setting and achieves substantially majority baseline classifier in cross-lingual setting. **Showrov et al.,[2021] [12]**In this article "News classification from microblogging dataset using supervised learning" ,identifies news from Twitter dataset and find best known model for microblogging dataset, that began with data crawling and then apply four supervised learning algorithms, ended by selecting best one and also best template for crawled dataset. Finally, all five datasets have applied four models (Naïve Bayes, Decision Tree, SVM, Random Forest) and found Naïve Bayes is the best algorithm than other models. **Qu et al.,[2006][11]** In this article "Automated Blog classification: Challenges and Pitfalls" investigates the efficiency of using machine learning for categorization of different blogs. To categorize 120 blogs into specified topics are Personal Diary, News, Political and Sports. Then the Baseline feature is unigrams weigh by TF-IDF with 84% accuracy. The result analysed for a given blog for classification under more than one category. And also proved to be effective as a particular blog can fall into multiple categories. **Song et al.,[2015] [10]**In this article "Classifying and Ranking Microblogging Hashtags with News categories" classifies hashtags for news categories. This paper represents 3 modules: training domain classification model, domain classification of hashtags and domain sensitive ranking, to get hot hashtags in each domain. Then microblogs finds a related hashtag to its text and classify the hashtag with a domain. **Fuks, O [2018] [9]** In this article "Classification of News Dataset" represents data source from Kaggle Dataset which contains 125000 news from past 5 years from Huffpost. This dataset contains 31 different topics. In this work, samples description size must be greater than 7 words. After removing left over samples are 113342 and 25 labels. Data pre-processing included removal of stopwords, punctuations and stemming each word. This paper has been experimented on machine learning techniques: Naïve Bayes, multinomial logistic regression, kernel SVM and random forest. As a result, dev set achieves 68.85% accuracy for logistic regression and confusion matrix for logistic regression with TF-IDF has obtained overall accuracy. **Hui et al.,[2017] [8]** In this article "Effects of Word class and Text position in Sentiment-based News classification" identify the contents of news to be exposed for classification. Then analyses and determines the key parts of news contents for sentiment-based category. Included training classifiers are text parts of speech and text position. Evaluated on 250 English news label with sentiments for sentiment voting system. The results of sentiment and polarity-based category has F score of 0.422 and 0.837. This study shows that finer categories are less effective and on polarity orientations, outcomes for text positioned at headlines and then using adjective words. **Fanny et al.,[2018] [7]** In this article "A Comparison of text classification methods K-NN, Naïve Bayes and Support Vector Machine for News Classification" study text classification with methods as K-NN, Naïve Bayes and SVM, to classify the news category in English. This experiment works on six news categories that involves Entertainment, Health, Sports, Economy, Politics, and Technology. With 90 samples for each category from Fox News, New York Times, ABC News and BBC Datasets. The main goal is to find the improved and quality method to categorize a news. **Katari et al.,[2020] [6]** In this article "A survey on News Classification Techniques" contributes the working and evaluation of different algorithms with merits and demerits for multiple news classification based on machine learning technique which have been studied. The survey of observation have been done for betterment choice of algorithms on datasets. Overall, this paper studies various research works and techniques, used to classify the news headlines with accuracy. The underlying techniques are K-NN, Naïve Bayes, SVM, Hierarchical Multi-label and clustering using Semi-Supervised Learning. **Singh et al.,[2018] [5]** In this article "Intra News category classification using N-gram TF-IDF features and Decision Tree classifier" study on multiple feature based news classification on BBC News dataset with inter and intra news classification. Sports category for intra class whereas technology, business, sports, politics and entertainment for inter class classification. Firstly, pre-processing each news articles from different news categories, feature extraction with TF-IDF using unigram, bigram and trigrams. Different classifiers used but decision tree gives effective results than others. Hence, it achieved 96% accuracy in intra class news and about 98% for inter class classification in true identification of observed sample of news dataset. **Saigal et al.,[2020] [4]** In this article "Multi-Category news classification using support vector machine based classifiers" represents by measuring the quantitative analysis on different SVM classifiers in multiple category. Least Squares SVM, Twin SVM and Least-Squares Twin SVM(LS-TWSVM) are classifiers on news dataset. This dataset performs pre-processing activities, feature set as TF-IDF to be performed on each document. The efficiency of each algorithm are evaluated on UCI News datasets ie., Reuters and 20Newsgroups. The report has achieved with an accuracy of 92.96% for LS-TWSVM that outperforms other two classifiers. **Singh et al.,[2019] [3]** In this article "Comparison between Multinomial and Bernoulli Naïve Bayes for Text

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

classification" is to predict the sentiment of news data either to positive or negative using Naïve Bayes classifier ie., Multivariate Bernoulli Naïve Bayes and Multinomial Naïve Bayes. It has concluded that Multinomial Naïve Bayes has little variation which is better than Bernoulli Naïve Bayes but Multinomial Naïve Bayes gets 73% higher accuracy then Bernoulli Naïve Bayes with 69% accuracy that implies the evaluation of algorithms may not differ on given dataset. **Asad et al.,[2020] [2]** In this article "Classification of news articles using supervised machine learning approach" is done on news channel for automatic classification which is based on the title and description of class of the news from 2014 to 2018. This news classification is on dual languages ie., Urdu and English. As model can predict the class based on title of the news and can also predict category of class and sub-category based on automated news description. Here, feature engineering includes counter vectorization, TF-IDF, word2vec, are 3 features which are ready-to-use two machine learning algorithms such that Logistic Regression with 98% accuracy. **Basha et al.,[2021] [15]** In this article "Design and Implementation of NEWS Classification Predictor using Machine Learning**"** used self-study dataset to train classifiers such as KNN and Naïve Bayes and also compared accuracy, average precision, precision and recall may or may not have combination of feature selection techniques. The results conclude that all documents collected from online articles for self-study have five categories – business, sports, technology, politics and entertainment i.e., from CNN, Washingtonpost and newyork times. Then, each category predictors performance is measured by analysing confusion matrix. This research study shows that all category predictors have adequate accuracy grades obtained. **Meng et al.,[2020] [1]** In this article "Text classification using Label names only: A language Model self-Training Approach" investigate problem of weakly supervised text classification for label of each class and also train a classifier on unlabelled data. Authors used pre-trained neural language models as general source of knowledge for category understanding and representation learning model for document classification. The model exploits LOT class [label-name-only text classification] model in three steps:

i. They have constructed vocabulary for each class contains semantics correlated words with label names using a pretrained language model.
ii. LM gathers high quality category indicative words in unlabeled corpus to train context word-level category.
iii. LM model generalizes via self-Training on unlabeled data. LOT class obtain 90%accuracy on four Benchmark datasets on AG News, DBPedia, IMDB and Amazon corpora yields weakly-supervised models and also evaluated with strong semi-supervised and supervised models.

III. Proposed FrameWork

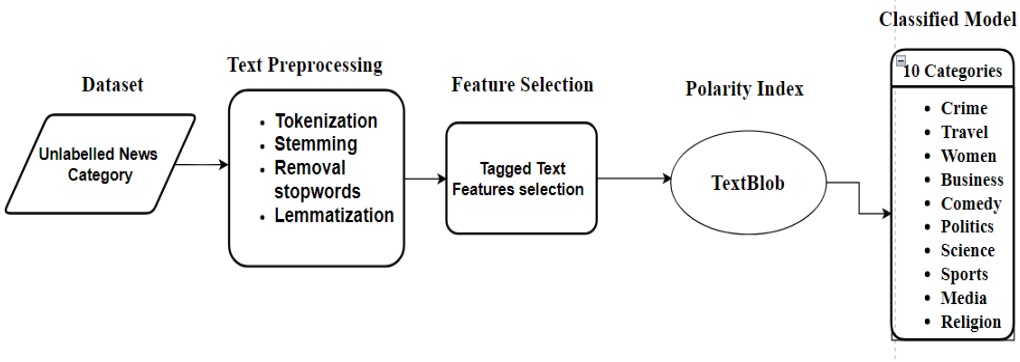


Figure 1: Proposed Data flow of text classification model

IV. News Category Text Analysis

A. Dataset

In this article, we use '**News_Category_Dataset_v2**' from Kaggle that reviews headlines from Huffpost inbetween 2012 to 2018 year and classify it with few categories. This dataset contains 30 categories with records of 2,00,853, that it exceeds to 5MB limit.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

	category	headline	authors	link	short_description	date
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama...	She left her husband. He killed their children...	2018-05-26
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.	2018-05-26
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...	2018-05-26
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre...	The actor gives Dems an ass-kicking for not fi...	2018-05-26
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-...	The "Dietland" actress said using the bags is ...	2018-05-26
...	...	...	...	...	...	...
495	COMEDY	Fake Donald Trump And Sean Hannity 'Pillow Tal...	Lee Moran	https://www.huffingtonpost.com/entry/donald-tr...	"I've been watching you."	2018-05-16
496	ENTERTAINMENT	Rapper T.I. Arrested For Disorderly Conduct, P...	Jenna Amatulli	https://www.huffingtonpost.com/entry/rapper-ti...	He reportedly went back to keep arguing with t...	2018-05-16
497	POLITICS	Conservatives Cook Up Farm Bill Gambit to Get ...	Matt Fuller	https://www.huffingtonpost.com/entry/freedom-c...	This is getting very House of Cards-y.	2018-05-16
498	POLITICS	Top Novartis Lawyer Resigns Over Michael Cohen...	John Miller, Reuters	https://www.huffingtonpost.com/entry/novartis-...	U.S. lawmakers demand details from Novartis an...	2018-05-16
499	TECH	Why Isn't There A Birth Control Emoji?	Nicole Lee, Engadget	https://www.huffingtonpost.com/entry/why-isnt-...	Two women decided it was time to have one.	2018-05-16

500 rows × 6 columns

Figure 2:'News_Category_Dataset_v2' dataset

As shown in figure 3, we use 10 categories: **CRIME, TRAVEL, WOMEN, BUSINESS, COMEDY, POLITICS, SCIENCE, SPORTS, MEDIA, RELIGION**. All these categories contains 73,066 with memory usage of 1.7MB. Therefore, the dataset get imbalanced. Each news has several attributes for Text classification. Here, we concatenate 'Headline' and 'Short Description' attributes into one as 'Text' attribute for data input of our model.

	y	text
61596	RELIGION	Watch Conan O'Brien Get In A Snowball Fight Wi...
199977	TRAVEL	Finding Sodom In Madaba, Jordan's Historic Cit...
141611	TRAVEL	Best Beaches Coast to Coast
72144	CRIME	Drunk Driver Found Hiding In Nativity Scene Af...
96698	WOMEN	This Is What a Feminist Will Look Like
74820	POLITICS	Ted Cruz Ties 'Amnesty' For Undocumented Immig...
110974	WOMEN	Why It Doesn't Matter That The Emma Watson Thr...
61375	COMEDY	#BernieHillaryRomComs Is A Perfect Substitute ...
186766	TRAVEL	Neither Here Nor There: Notes From Abroad
10127	POLITICS	Kristi Noem Says Her Story Shows How The Estat...
26794	POLITICS	What The First Polls Say About Comey's Firing
33126	POLITICS	Betsy DeVos Accused Of 'Whitewashing' That Hist...

Figure 3: Selection of 10 different Categories from 'News_Category_Dataset_v2' dataset

B. Text Preprocessing

The procedure of text categorization precludes categorizing the whole text into predefined entities. For the level of complexity on dataset which includes large number of text to be classified into predefined class. This class categorization can be explored with an advance technique on our dataset, that is text preprocessing and text extraction.

Text pre-processing is one of the step for preparing raw text apt to machine learning model which includes text cleaning, removal of stopwords, tokenization, Porter Stemmer and Wordnet lemmatizer.

Text extraction is to represent unique patterns with the partition of words in text into individual word count statistics in certain order. N-grams is a sequence of N tokens(words) in text with consideration of order and relationship between words. The steps to be followed as:

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

Table 1: Steps of Text Preprocessing

Text Cleaning	An unstructured data is preprocessed with collections of significant and insignificant data cleaned from perverted data. And also meaningless information such as !, :, ;, ', " etc, irrelevant sentences, dates etc. By transforming all characters either lower and upper case into lower case. To eliminate homologous words interms of their case.
Removal stopwords	Filtering stopwords is to delete insignificant words occur frequently or infrequently in sentences or text and also remove words with no specific meanings such as a, an, or the. This reduces execution time and complexity in computation. Here, download a list of universal stopwords for English vocabulary using NLTK(Natural Language Toolkit) libraries.
Tokenization	First step of preprocessing is to convert of text-based entity into a list of tokens. Tokens represents an entity which substitutes with idiosyncratic class of tokens. Hence, all words are delimited in sentences and all punctuations are disposed which cannot represent any entity.
Stemming (Porter Stemmer)	The crucial step of preprocessing is stemming, as it reduces the complexity to some extent by storing single word for various forms present in it.
Lemmatization (Wordnet lemmatizer) –	Lemmatization is a process of grouping different inflected words as a single item. It extracts high quality information from Natural Language. But similar to stemming. Example of Inflected words are read, reads, reading, reader.

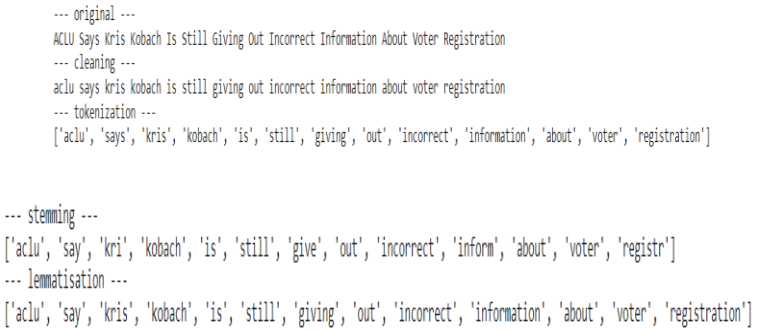


Figure 4 : Sample text Preprocessing

C. Tagged text feature selection

In our model, we use Named Entity Recognition (NER) for the process of tagging named entities in unstructured text with predefined categories such as people names, organizations, locations, time, expressions, quantities etc. When NER model apply on each text of the dataset. This dataset has envisioned each news headlines to all recognized entities in the form of new column (named "tags"), as the text with same entity that appears for number of times.
For ex:{('Texas','GPE'):1,('Last Week','DATE'):1, ('Only 1' , 'CARDINAL'):1}.

text	label	Feature 1	Feature 2	...
News 1	Y1	X1	X1	...
News 2	Y2	X2	X2	...
...	...	...	...	

Figure 5: Feature selection for label text

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

Figure 6: Tag attributes

```
0         [(Texas, GPE), (Last Week, DATE), (Only 1, CAR...
13        [(Trump's Crackdown On Immigrant Parents Puts ...
14        [(Trump's Son Should Be Concerned', WORK_OF_AR...
15        [(Edward Snowden, PERSON), (Vladimir Putin, PE...
16                                        [(Booyah, ORG)]
                                ...
200837                                                []
200849    [(Maria Sharapova, PERSON), (Victoria Azarenka...
200850    [(Giants Over Patriots, ORG), (Most Improbable...
200851    [(Aldon Smith Arrested, PERSON), (49ers, CARDI...
200852                        [(Dwight Howard Rips, PERSON)]
Name: tags, Length: 73066, dtype: object
```

Figure 7: Assigning tags for each text on category -based

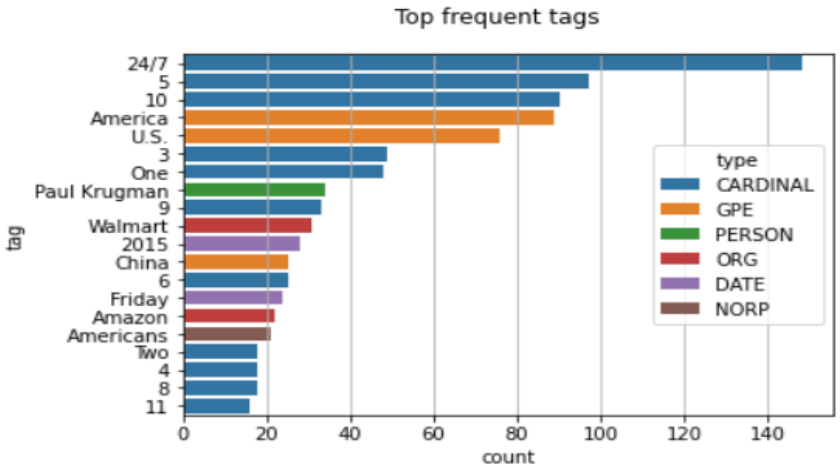


Figure 8: Top frequent tags of Business category

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

- **Word Cloud**

Word cloud are small graphs, tell us about common words occurred in all news articles of the news category dataset. Generating word cloud for a corpus of text is simply concatenating all articles into one group of string. To instantiate the word cloud class and with different parameters such as back ground color, font-size, colour map to display the words in, etc. This word cloud is completely a mixture of top occurrence of words from 10 categories such as **CRIME, TRAVEL, WOMEN, BUSINESS, COMEDY, POLITICS, SCIENCE, SPORTS, MEDIA, RELIGION**.

Figure 9: Word cloud of 10 categories

- **Length analysis**

To have word frequency as a feature in data frame. We need to use the count vectorizer from scikit-learn as the most popular library of machine learning in python. This vectorizer converts a collection of text documents to a matrix of tokens. This tokenization process is to split a string into a list of strings or tokens. Then to analyse the length of the text by the insight of dataset that one of the category is systematically larger than another and whole text counts for the number of words, Characters and Sentences. To find average length of text analysis to build the predicted classified model.

Table 2 : Summary of Length Count

Length Count	Definitions
Word Count	Counting number of tokens in sentence which is separated by a space.
Character Count	Number of characters to be counted for each token in sentence.
Sentence Count	Count number of sentences in text
Average Word Length	Total sum of words length is divided by number of words count(character/word)
Average sentence length	Total sum of sentences length will be divided by the number of sentences(word /sentence count)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

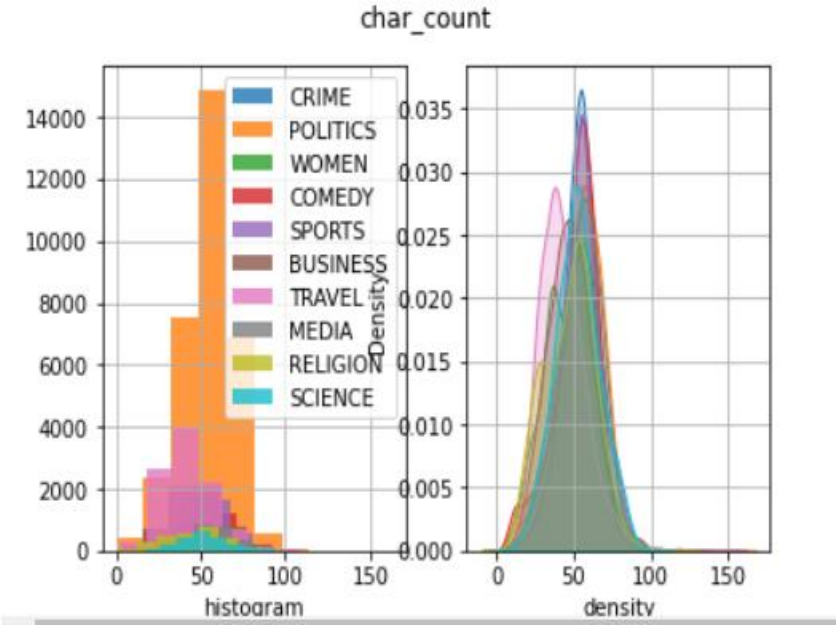


Figure 10: Histogram and Density map of character count

D. Polarity index

In our research work, Polarity index is computed using TextBlob library, a pre-trained model to fit our data according to its context. Further we train our dataset for the evaluation of sentiment scores of the whole textual data with the estimation of an average sentiment scores. The experimental view demonstrates most of the news headlines having neutral sentiment, except politics news headlines skewed on negative tail and Science category spike is on positive tail. The polarity index is represented on density map as shown in the figure 11.

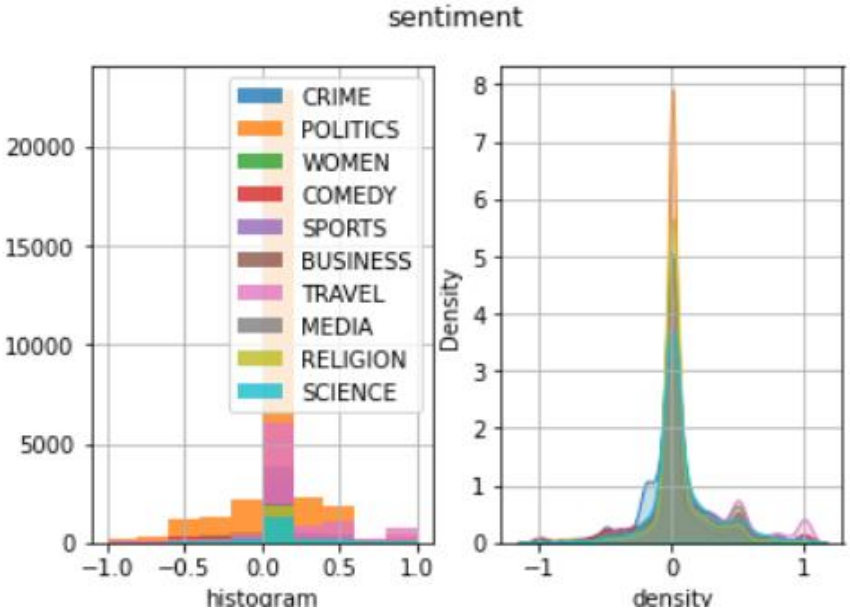


Figure 11: Polarity index

E. News category analysis and Feature Extraction on TF-IDF of Unigram and Bigram

The dataset shows statistical analysis of data – as sample count per category and average count of words per news description. With composition of dataset, we have taken only 10 different categories which shows Univariate and Bivariate

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452
distributions. A probability distribution of just one or two variables with class label frequency bar plot that shows Histograms of the samples. If the distributions are variant, each category is predictive with different patterns.

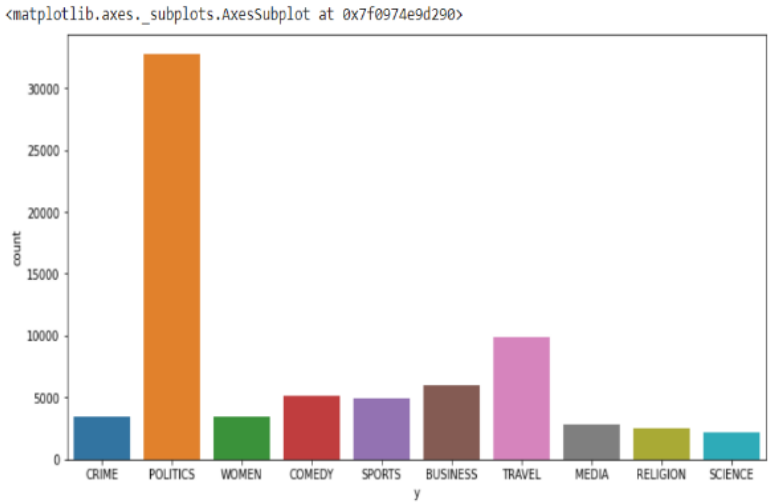


Figure 12: Frequency count of 10 categories

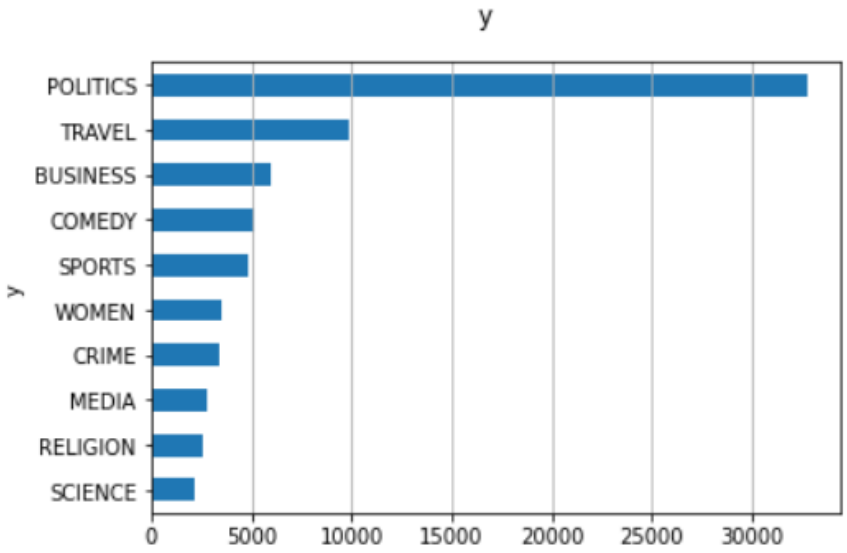


Figure 13: Ordered (max and min) frequency count of 10 categories

TF-IDF (Term Frequency - Inverse Document Frequency) is a numerical statistics technique that intend to reflect the most important words of a document in the corpus. Term Frequency measures how frequently a term appears in a document and Inverse Document Frequency measures how important a term is whereas term frequency is not sufficient to analyse the importance of words. Thus, each document and term would have its term frequency value. With this, when words are reduced to 0 have little importance whereas words with more importance have a higher value.

As with drawbacks of using BoW, represents simple count of word frequency. Word frequency means number of times a term is repeated in a text. The issues that are facing across new sentences are:

i. If new sentences contain new words, the vocabulary size may increase as well as length of the vectors may also increase too.
ii. The vectors would also contain many 0s, thereby resulting in sparse matrix.
iii. Retaining no information on grammar of the sentences nor on ordering words in text.

In this article, the most important words are analysed and processed using TF-IDF of unigrams and bigrams. Consequently, Unigram and Bigrams are also been plotted here. With this manifest or agenda, the importance of each word

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

is calculated with respect to TF-IDF which is one of the eminent algorithms used in mining the text. This algorithm calculates the inverse probability of finding a word in text using classic TF-IDF equation represents to do calculation of weights as follows:

In this formula, W_{ij} is the weight of the word i in the document j, N is the number of documents, tf_{ij} is the frequency of the word i in document j, and df_i is the number of documents containing the word i. In this article, we used above equation to compute TF-IDF scores for each dominant feature of different categories used (words with maximum TF-IDF score greater than a certain threshold).

$$W_{ij} = tf_{ij} * \log \frac{N}{df_i} \longrightarrow \text{Eq 1}$$

	feature	score	y
86	247	1.000000	BUSINESS
87	247 wall	1.000000	BUSINESS
418	amazon	1.000000	BUSINESS
788	bank	1.000000	BUSINESS
989	billion	1.000000	BUSINESS
...	...	...	...
8117	soar	0.993008	BUSINESS
7632	say	0.992893	BUSINESS
4923	largest	0.992848	BUSINESS
7118	regulation	0.992646	BUSINESS
8014	simple way	0.992597	BUSINESS

Figure 14: TF-IDF scores

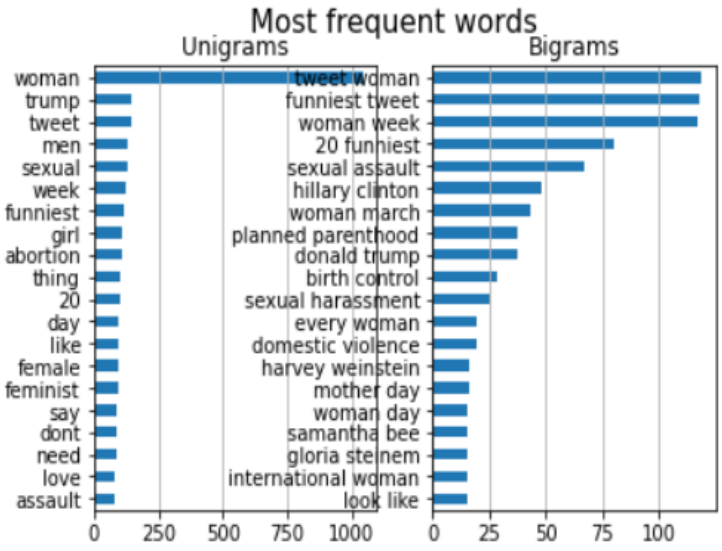


Figure 15: Unigrams and Bigrams for Woman Category with Most frequent words.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

V. News Category Sentiment Analysis

The news sources classification in categories gathered all news information which is not correctly labelled due to huge flow of information, it is impossible for a human to label each one manually. So, among news generators there is no common ontology and neither agencies are organised in same manner. This text labelling is the true classification for the recommender system who are aggregating news to have its own classification system. Using machine learning we classify short news and informs which label that news is from.

Sentiment analysis is one of the demanding task, used to interpret and classify emotions such as positive and negative values. This may be helpful in analysing the people's emotions on various news expertise. In this concept, tells whether the word occurs in a sentence or not as well as its frequency. While predicting the polarity of new news information, finds out the probability of occurrence of all the words in news category which gives the polarity index. Polarity index evaluates the inclination of the news category interms of positive (indicated by 1) and negative (indicated by -1) values.

VI. Result Analysis

Analysis I: Assigned tags to the text are specified for how many number of times the particular tag occurs.

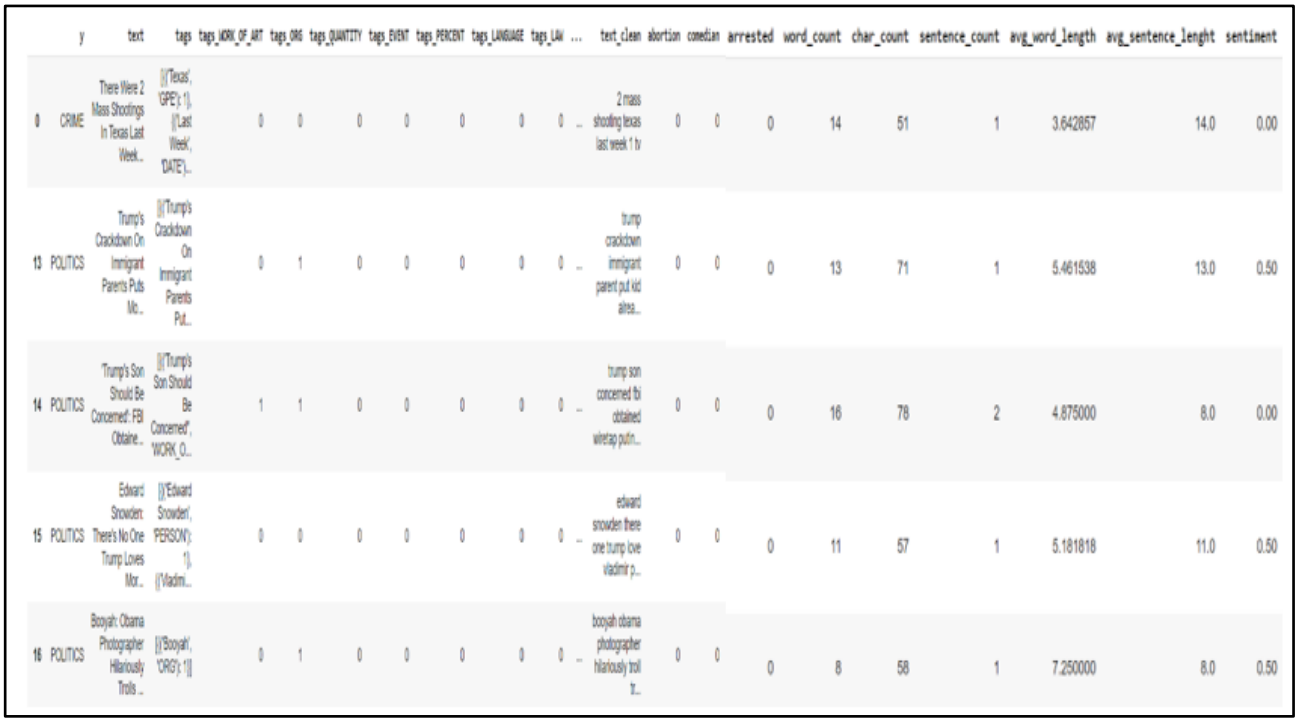
	y	text	tags	tags_WORK_OF_ART	tags_ORG	tags_QUANTITY	tags_EVENT	tags_PERCENT	tags_LANGUAGE	tags_LAW	...	tags_DATE	tags_PERSON	tags_NORP	tags_MONEY	tags_GPE	tags_LOC	tags_FAC	tags_ORDINAL	tags_PRODUCT	tags_CARDINAL
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	[('Texas', 'GPE'): 1), [('Last Week', 'DATE')...	0	0	0	0	0	0	0	...	1	0	0	0	1	0	0	0	0	1
13	POLITICS	Trump's Crackdown On Immigrant Parents Puts Mo...	[('Trump's Crackdown On Immigrant Parents Put...	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
14	POLITICS	'Trump's Son Should Be Concerned': FBI Obtaine...	[('Trump's Son Should Be Concerned', 'WORK_O...	1	1	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0
15	POLITICS	Edward Snowden: There's No One Trump Loves Mor...	[('Edward Snowden', 'PERSON'): 1), [('Vladimi...	0	0	0	0	0	0	0	...	0	2	0	0	0	0	0	0	0	0
16	POLITICS	Booyah: Obama Photographer Hilariously Trolls ...	[('Booyah', 'ORG'): 1)]	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

Analysis II: After text cleaning, compute the number of characters, words, sentences and also compute average word as well as average sentence length of text analysis.

	y	text	tags	tags_WORK_OF_ART	tags_ORG	tags_QUANTITY	tags_EVENT	tags_PERCENT	tags_LANGUAGE	tags_LAW	...	tags_GPE	tags_LOC	tags_FAC	tags_ORDINAL	tags_PRODUCT	tags_CARDINAL	text_clean	abortion	comedian	arrested	word_count	char_count	sentence_count	avg_word_length	avg_sentence_lenght
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	[('Texas', 'GPE'): 1), ('Last Week', 'DATE')...	0	0	0	0	0	0	0	...	1	0	0	0	0	1	2 mass shooting texas last week 1 tv	0	0	0	14	51	1	3.642857	14.0
13	POLITICS	Trump's Crackdown On Immigrant Parents Puts Mo...	[('Trump's Crackdown On Immigrant Parents Put...	0	1	0	0	0	0	0	...	0	0	0	0	0	0	trump crackdown immigrant parent put kid alrea...	0	0	0	13	71	1	5.461538	13.0
14	POLITICS	'Trump's Son Should Be Concerned': FBI Obtaine...	[('Trump's Son Should Be Concerned', 'WORK_O...	1	1	0	0	0	0	0	...	0	0	0	0	0	0	trump son concerned fbi obtained wiretap putin...	0	0	0	16	78	2	4.875000	8.0
15	POLITICS	Edward Snowden: There's No One Trump Loves Mor...	[('Edward Snowden', 'PERSON'): 1), ('Vladimi...	0	0	0	0	0	0	0	...	0	0	0	0	0	0	edward snowden there one trump love vladimir p...	0	0	0	11	57	1	5.181818	11.0
16	POLITICS	Booyah: Obama Photographer Hilariously Trolls ...	[('Booyah', 'ORG'): 1)]	0	1	0	0	0	0	0	...	0	0	0	0	0	0	booyah obama photographer hilariously troll tr...	0	0	0	8	58	1	7.250000	8.0

Analysis III: After Length Analysis, compute average sentiment score of each news headlines for the selected category either positive or negative.



	y	text	tags	tags_WORK_OF_ART	tags_ORG	tags_QUANTITY	tags_EVENT	tags_PERCENT	tags_LANGUAGE	tags_LAW	...	text_clean	abortion	comedian	arrested	word_count	char_count	sentence_count	avg_word_length	avg_sentence_lenght	sentiment
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	[('Texas', 'GPE'): 1), ('Last Week', 'DATE')...	0	0	0	0	0	0	0	...	2 mass shooting texas last week 1 tv	0	0	0	14	51	1	3.642857	14.0	0.00
13	POLITICS	Trump's Crackdown On Immigrant Parents Puts Mo...	[('Trump's Crackdown On Immigrant Parents Put...	0	1	0	0	0	0	0	...	trump crackdown immigrant parent put kid alrea...	0	0	0	13	71	1	5.461538	13.0	0.50
14	POLITICS	'Trump's Son Should Be Concerned': FBI Obtaine...	[('Trump's Son Should Be Concerned', 'WORK_O...	1	1	0	0	0	0	0	...	trump son concerned fbi obtained wiretap putin...	0	0	0	16	78	2	4.875000	8.0	0.00
15	POLITICS	Edward Snowden: There's No One Trump Loves Mor...	[('Edward Snowden', 'PERSON'): 1), ('Vladimi...	0	0	0	0	0	0	0	...	edward snowden there one trump love vladimir p...	0	0	0	11	57	1	5.181818	11.0	0.50
16	POLITICS	Booyah: Obama Photographer Hilariously Trolls ...	[('Booyah', 'ORG'): 1)]	0	1	0	0	0	0	0	...	booyah obama photographer hilariously troll tr...	0	0	0	8	58	1	7.250000	8.0	0.50

VII. Model Evaluation

In this model, pre-processing and the exploratory data analysis steps have been done. The next step is to split the dataset into training and testing subsets. Thus, the classification model is apted with training data and predictions are obtained with test dataset. For this dataset, the Multinomial Naïve Bayes Classifier are used to give best performance when compared to other classifiers. According to this documentation, the classifier is advisable for text classification with word counts.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

To train the classifier on feature matrix and test it on transformed test set. We build Scikit-learn pipeline: an application with list of transformations and final estimator. By combining TF-IDF vectorizer and the Naïve Bayes classifier in a pipeline allows to transform and predict test data in single step.

The evaluation metrics of our model for 10 different categories are analysed and computed with Accuracy (0.70), Confusion Matrix, Receiver Operating Characteristics, Area Under Curve (0.94), Precision, f1 score and Recall.

```
Accuracy: 0.7
Auc: 0.94
Detail:
                precision    recall  f1-score   support

      BUSINESS       0.79      0.34      0.47      1774
        COMEDY       0.81      0.46      0.58      1525
         CRIME       0.81      0.45      0.58      1004
         MEDIA       0.82      0.14      0.24       853
      POLITICS       0.65      0.97      0.78      9810
      RELIGION       0.90      0.23      0.36       756
       SCIENCE       0.94      0.31      0.47       669
        SPORTS       0.86      0.59      0.70      1499
        TRAVEL       0.79      0.84      0.81      2957
         WOMEN       0.83      0.19      0.31      1073

      accuracy                           0.70     21920
     macro avg       0.82      0.45      0.53     21920
  weighted avg       0.74      0.70      0.66     21920
```

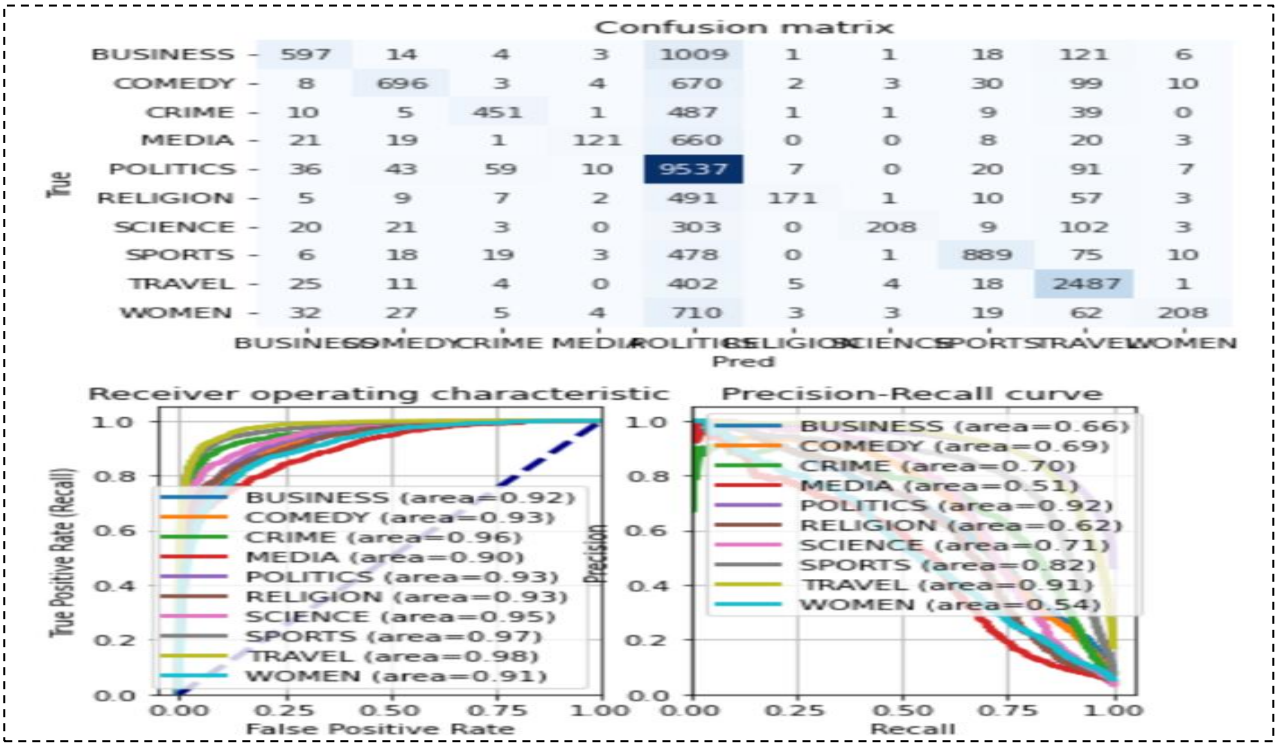


Figure 16: Evaluation metrics plotted with ROC, AUC and Precision-Recall curve

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.5485-5498
https://publishoa.com
ISSN: 1309-3452

VIII. Conclusion

We proposed a machine learning model for multi-label news category text classification as presented in this article. The main idea behind our proposed machine learning model is to use Multinomial Naïve Bayes Classifier with weights estimated to maximize evaluation scores computed as macro-averaged, Weighted average, Precision, f1 score and Recall. The results obtained are well performed over the conventional methods of news text classification with Accuracy 0.70), Receiver Operating Characteristics, Area Under Curve (0.94), Precision, f1 score and Recall are computed for each category of news headlines. The future scope of the article includes to investigate the work in the directions of web scraping to search information in any given website. To further perform analysis on the various evaluation parameters in "scraping" the news headlines.

References:

1. Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., & Han, J. [2020]. Text classification using label names only: A language model self-training approach. *arXiv preprint arXiv:2010.07245*.
2. Asad, M. I., Siddique, M. A., Hussain, S., Hassan, H. N., & Gul, J. M. [2020]. Classification of News Articles using Supervised Machine Learning Approach. *Pakistan Journal of Engineering and Technology*, *3*(03), 26-30.
3. Singh, G., Kumar, B., Gaur, L., & Tyagi, A. [2019, April]. Comparison between multinomial and Bernoulli naïve Bayes for text classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 593-596). IEEE.
4. Saigal, P., & Khanna, V. [2020]. Multi-category news classification using Support Vector Machine based classifiers. *SN Applied Sciences*, *2*(3), 1-12.
5. Singh, D., & Malhotra, S. (2018). Intra News Category Classification using N-gram TF-IDF Features and Decision Tree Classifier. *IJSART*, *4*, 508-514.
6. Katari, R., & Myneni, M. B. [2020, March]. A survey on news classification techniques. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-5). IEEE.
7. Fanny, F., Muliono, Y., & Tanzil, F. [2018]. A comparison of text classification methods k-NN, Naïve Bayes, and support vector machine for news classification. *Jurnal Informatika: Jurnal Pengembangan IT*, *3*(2), 157-160.
8. Hui, J. L. O., Hoon, G. K., & Zainon, W. M. N. W. [2017]. Effects of word class and text position in sentiment-based news classification. *Procedia Computer Science*, *124*, 77-85.
9. Fuks, O. [2018]. Classification of News Dataset. *Standford University*.
10. Song, S., & Meng, Y. [2015, May]. Classifying and ranking microblogging hashtags with news categories. In *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)* (pp. 540-541). IEEE.
11. Qu, H., La Pietra, A., & Poon, S. S. [2006, March]. Automated Blog Classification: Challenges and Pitfalls. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 184-186).
12. Showrov, M. I. H., Dubey, V. K., Hasib, K. M., & Shameem, M. A. [2021, February]. News classification from microblogging dataset using supervised learning. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 53-57). IEEE.
13. Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., & Pollak, S. [2020]. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, *10*(17), 5993.
14. Hussain, A., Ali, G., Akhtar, F., Khand, Z. H., & Ali, A. [2020]. Design and Analysis of News Category Predictor. *Engineering, Technology & Applied Science Research*, *10*(5), 6380-6385.
15. Basha, S. R., & Reddy, T. B. [2021]. Design and Implementation of NEWS Classification Predictor using Mechine Learning.