

Identifying Multiple Categories for News Headlines using Text Analysis

Shilpa Patil

Department of Studies in Computer Science,
Vijayanagara SriKrishnadevaraya University,
Ballari, Karnataka, India.
Email: shilpavskvsc@gmail.com

V Lokesha

Department of Studies in Mathematics,
Vijayanagara SriKrishnadevaraya University,
Ballari, Karnataka, India.
Email: v.lokesha@gmail.com

Anuradha S G

Department of Computer Science and Engineering,
Rao Bahadur Y. Mahabaleswarappa Engineering College,
Ballari, Karnataka, India.
Email: anuradhasuresh13@rymec.in

(Received Feb 21, 2023, Revised Dec 30, 2023, Accepted Jan 10, 2024)

Abstract: As entire world is witnessing this digital transformation giving rise to change in every aspect of human life. At Present, news classification is one of the valuable tools to improve the accessibility, and efficiency of news articles, due to the availability of huge digital data demanding accurate classification for users easy. The focus of our article is to build a context classifier for the news categories and evaluate the sentiments of text-based headlines. The Experimental study is carried out on 15 News categories available in the ‘**News_Category_Dataset_v2**’ dataset from Kaggle. The results of our study reveal the sentiment of the text associated with these news categories using a polarity index, with values ranging from positive and negative.

Keywords: News Categories, context classifier, Kaggle, polarity index, dataset, headlines.

1. Background

In digital transformation, the Internet allows for communal interactions among people through sharing or exchange of information and ideas in virtual

communities. The most spectacular development in communication for the innovative and astonishing different information that has given access to good quality. The present era is, fortunately, witnessing the most amazing technological developments which have become the rage of this age. The idea of web-based technologies is highly interactive and heralded a revolution on the world. Some of the leading media technologies that have an enormous collection of data in social media like Facebook, Twitter, Instagram, WhatsApp, Facebook Messenger, etc., for different entity-based. Thus, social media is a phenomenal source for distinct topics that have a big deal to be up-to-date with the most recent worldwide news for automated text classification using Natural Language Processing (NLP) which acquires vogue among researchers. Many researchers have a significant impact on the increasing growth of information in today's evolution.

NLP can be used in the processing and analysis of news information based on class. This text classifying method is to categorize text-based entities into predictive categories. A label may either identifies or extract important people, places, and organizations mentioned in news article. This may help us to build a knowledge graph or for tracking the relationships between entities with more labels or exactly single label or no label at all. By classifying news category, it automatically summarizes news to determine the sentiment (positive or negative) of a news article or report.

In this article, we utilize Natural Language Processing and Python-based approach to classify textual data and assign predicted categories based on its content. we will explore different methods for analyzing textual data and extracting features to build a classification model for 15 classes such as *Crime, Travel, Women, Business, Comedy, Politics, Science, Sports, Media, Religion, World News, Impact, Queer Voices, Tech, and Education* which are part of a news category dataset. Some of the literatures studied are related to our research work that has been discussed. Taj *et al.*¹ had focused on sentiment analysis of news articles on lexicon-based approach, using BBC dataset from 2004-2005. The study found that positive articles were more prevalent in business and sports categories, while negative articles dominated entertainment and tech categories. Isnaini *et al.*² a study of identifying classes in each multi-label news article using K-NN. The performance are measured by the comparison between manhattan distance, euclidean distance and supremum distance using K=11 parameters, have evaluated in hamming loss value of 11.16%. O. Fuks³ a dataset of 1,25,000 news articles from Kaggle through HuffPost for over 5 years. This dataset has 31 topics, to be pre-processed with machine learning Techniques: NB, Multinomial Logistic Regression, Kernel SVM, and Rf and achieved 68.85% accuracy on dev set

for LR and confusion matrix for LR with TF-IDF that has obtained overall accuracy. Fanny *et al.*⁴ have studied on how to classify the news category in English using K-Nearest Neighbours, Naïve Bayes and SVM. In this study, they classify six categories like entertainment, health, sports, economy, politics, and technology. 90 samples are taken as per the category from fox news, newyork times, ABC news, and BBC datasets. The focus is to find the quality and improvisation of methods for categorizing news Katari *et al.*⁵ have contributed the merits and demerits of different algorithms for multi-news category based on machine learning models. This observation has reviewed for the finer choice of algorithms on datasets. Thus, many research studies and techniques used to categorize news headlines with accuracy (using K-NN, NB, SVM, Hierarchical Multi-label, and clustering). Pintas *et al.*⁶ paper study was on organizing a systematic literature review of 1,376 unique papers from journals and conferences published from 2013-2020. All reviewed papers have contributed several proposed methods and also mapping to new categorization, the characteristics of experiments, catalogue of datasets, languages, machine learning algorithms, and validation methods that are evaluated with most recent techniques. Mapping crisis and evaluation settings, help researchers to study and develop effective results. Singh *et al.*⁷ analyzes feature-based multi-news categories on the BBC dataset with inter and intra-news. As certainly, the sports category for Intra and politics, entertainment, sports, technology, and business for Inter. Pre-processing and feature extraction with TF-IDF using single, binary and 3-grams are evaluated. By comparing different classifiers, decision tree gives more efficient results than others. With an achievement of 96% accuracy in intra-class and 98% on inter-class news classification datasets. Pozzi *et al.*⁸ have discussed a chapter on some topics about sentiment analysis such as Background details, the latest research approaches in social media, characteristics, levels, comparisons, applications, etc. Harjule *et al.*⁹ article is on existing datasets i.e., 'Sentiment140' and 'Crowd flower's Data for Everyone library' that have already been expressed with sentiments. But, some classifiers like TextBlob, sentiwordnet, MNB, LR, SVM, and RNN apply on both datasets and compute distinguished results from sentiment classifiers in the form of positive or negative. It also proved that among all classifiers RNN obtained highest accuracy. Sabri *et al.*¹⁰ compared different classification algorithms using two Arabic datasets 'cnn_arabic' and 'osac_uft8'. Methods used are: TF-IDF, wordcount and word2vec. Then, preprocessing is done for Arabic datasets. As results finds that SVM and LR classifiers are best one among five models. HaCohen *et al.*¹¹ explored the impact of all combination of 5 or 6 basic preprocessing methods on four

benchmark text corpora using 3 ML methods and trained and tested. The novelty of this paper is to improve accuracy using bag-of-words representation. Wang *et al.*¹² propose an improved TF-IDF model with word2vec to weigh word vectors. This proposed method is to revamp the traditional String To Word Vector algorithm for preprocessing. And finally, this improved algorithm has tested for four various datasets (WEBO_SINA and 3 standard UCI datasets) to show high accuracy, recall, and F1 values than traditional text classification models using word2vec or TF-IDF.

The paper is organized in 4 sections as: Section 1, introduces web-based technologies and discusses relevant literature. Section 2, discusses proposed framework of the paper. Subsection 2.1 discusses dataset view, Subsection 2.2 outlines the steps of text preprocessing. Subsection 2.3 explains about feature extraction of news data, and Subsection 2.4 is about word cloud. Subsection 2.5 analyzing length of the text extracted from news source, Subsection 2.6 describes challenging endeavour of performing sentiment analysis using TextBlob. Section 3 shows statistical analysis of news headlines. Finally, Section 4 concludes by discussing the multi-label news categories for classifying and evaluating the polarity index, as well as describing potential future enhancements of our study.

2. Proposal for a Frame work

News category dataset considers a group of articles with labels that are used to describe the content of news. This helps us to classify and organize news articles by topic, that yields people to easily find and read articles that are of interest to them. The datasets are used for miscellany purposes, such as building a news recommendation system, improving the organization of news articles in a database, or creating a summary of the day's news for a particular topic. News Category dataset is to train machine learning models for the classification of news articles into discrete categories Some of the common categories are: Politics are about political events, parties, and leaders, Tech news category is about new and emerging technologies that can develop skills, Entertainment is all about celebrities, movies, music, etc., the Science category collects the information of scientific research and discoveries, and so on.

As our proposed dataset has a substantial volume of data for the process of text classification and also analyzes the performance of entity-based. In our article, we emphasized only 15 classes: *Crime, Travel, Women, Business, Comedy, Politics, Science, Sports, Media, Religion, World News Impact, Queer Voices, Tech, And Education* from news category dataset as in

Figure1. These datasets includes with fields-like Category, Headline, Author, Link, Short_Description, Date. Hence, we combine the ‘Headline’ and ‘Short Description’ attributes into a single ‘Text’ attribute as input for our model.

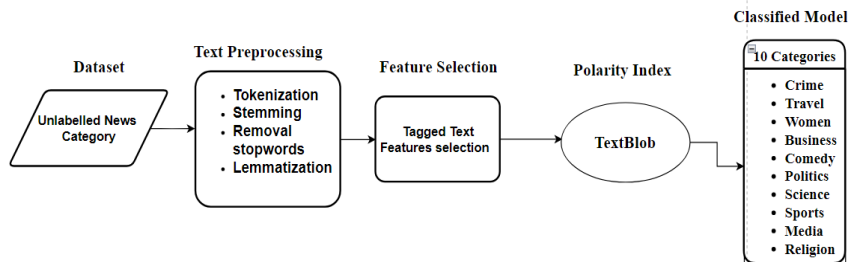


Figure 1: Proposed Data Flow of News Headlines Text Category

From this dataset, we also labelled bar graph for selected categories as shown in Figure 2 or 3.

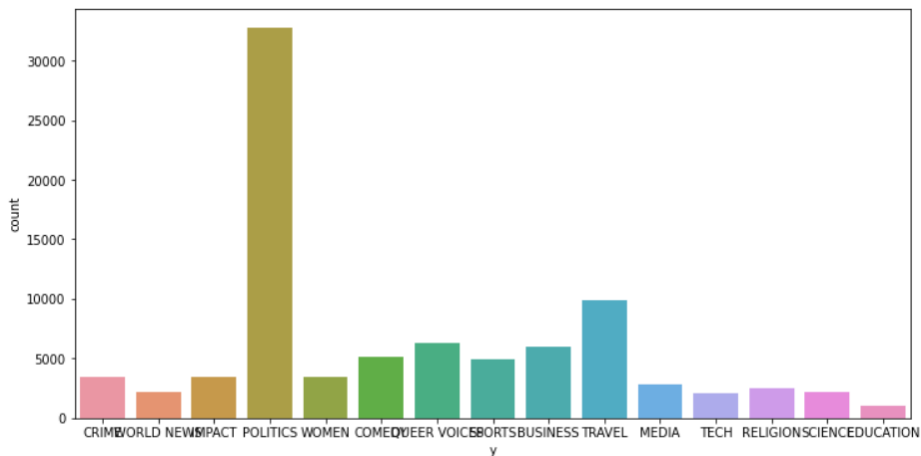


Figure 2: Labeled Frequency of 15 categories

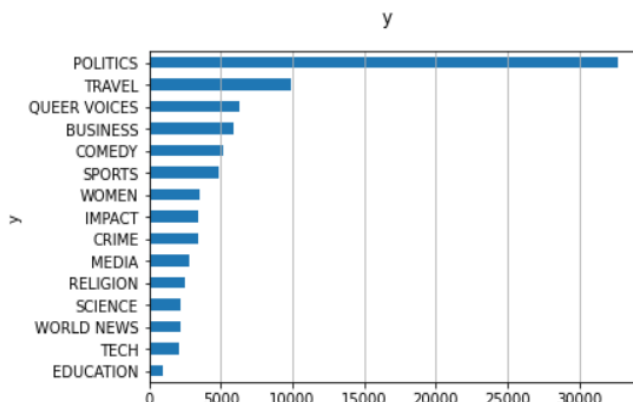


Figure 3: Maximum and Minimum Category Based on 15 Categories

2.1 Dataset

Our study analysis on ‘*News_Category_Dataset_v2*’ from Kaggle, which includes news headlines from Huffpost from 2012 to 2018, is organized into various categories. The dataset comprises 30 categories and a total of 200,853 records, exceeding the 5MB limit. For the purposes of our article, we consider 15 categories, resulting in 88,102 records with a memory usage of 2.0MB. This creates an imbalanced dataset.

2.2 Data Preparation Steps

The method of data preparation is to categorize entire sentences into prespecified entities for organizing and filtering innumerable data. Text categorization is trained to categorize text automatically using labeled training data. To explore sophisticated methods on datasets, that is text extraction and text preprocessing. Text pre-processing is one of the prominent steps in the NLP pipeline, as it aids to transform unprocessed data into a more suitable form for further analysis. It also facilitates by lowering the dimensionality of data and removing noise that may interfere learning process of NLP models.

Table I: Text Preprocessing steps involved

Text Cleaning	An unstructured data is <u>preprocessed</u> with collections of significant and insignificant data cleaned from perverted data. And also meaningless information such as <u>as !, :, , ' , "</u> etc, irrelevant sentences, dates etc. By transforming all characters either lower and upper case into lower case. To eliminate homologous words <u>interns</u> of their case.
Removal <u>stopwords</u>	Filtering <u>stopwords</u> 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 <u>stopwords</u> for English vocabulary using <u>NLTK</u> (Natural Language Toolkit) libraries.
Tokenization	First step of <u>preprocessing</u> 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 <u>preprocessing</u> is stemming, as it reduces the complexity to some extent by storing single word for various forms present in it.
<u>Lemmatization</u> (Wordnet <u>lemmatizer</u>) –	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.

This Text preprocessing involves text cleaning, stop word removal, tokenization, stemming (Porter Stemmer), and lemmatization (Wordnet lemmatizer). The text preprocessing steps are as follows in the Table I. Text extraction is the process of extricating relevant information to represent quirky patterns from a piece of text. This text extraction includes named entities (such as people, organizations, locations), key phrases and terms, and structured data(such as dates, prices, and addresses).

2.3 Text Tagged with Feature Extraction

A Named Entity Recognition (NER) model is a machine learning technique that categorizes named entities in unstructured text into prespecified categories, as in Figure 4, for the feature selections such as people, org, places, date, expressions, quantity, etc.

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

Figure 4: Feature Selection for Label Text

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 5: Tagged Attributes

2.5 Length analysis

To extract word frequencies as a feature from textual data for use in a machine learning model, you can use CountVectorizer from Scikit-learn, which converts a collection of text documents into a matrix of token counts, where each row represents a document and each column represents a unique token. The tokenization process is to split a text into a list of individual ‘tokens’, such as words, characters, and sentences. CountVectorizer can also be used to analyze the length of the text, such as the number of words, characters, and sentences, and to find the average length. This information can be used to build a predicted classified model, especially if one category in the dataset is systematically larger than the other. Table II, defines some of the terminologies used in this research.

Table II: Some Terminologies of Length Analysis

Length Count	Definitions
Word Count	Counting tokens in a sentence that is separated by a space.
Character Count	A number of characters are to be counted for each token in a sentence.
Sentence Count	Counting the number of sentences <u>in</u> text
Average WordLength	The total sum of words is divided by the number of words counted (either character/word)
Average sentencelength	The total sum of sentences will be divided by the <u>number of</u> sentence (either word /sentence count)

2.6 Polarity index

Sentiment analysis is one of the challenging endeavors, used to determine the emotional tone or attitude in a piece of text, such as a sentence or paragraph. Regardless of the approach used, this challenging task gives the meaning of words and phrases that may vary sentiment, on its content. When predicting the polarity of news information, sentiment analysis involves evaluating the probability of occurrence of words in the news category. This can give a polarity index, which indicates the inclination of the news category towards positive or negative sentiments. Positive sentiments are usually indicated by a value of 1, while negative sentiments are indicated by a value of -1.

In our model, we use the TextBlob library to evaluate the sentiment index of our data. This library is a pre-trained model that is able to adapt to the context of our data. We also train our dataset to evaluate the sentiment index of the full-length text data, estimating an average sentiment score. Our evaluation results reveal that most of the news headlines have a neutral sentiment, except for political news, which tends to have a negative sentiment, and education news which tends to have a positive sentiment. The

polarity index is represented in a histogram and density map, as shown in Figure 8.

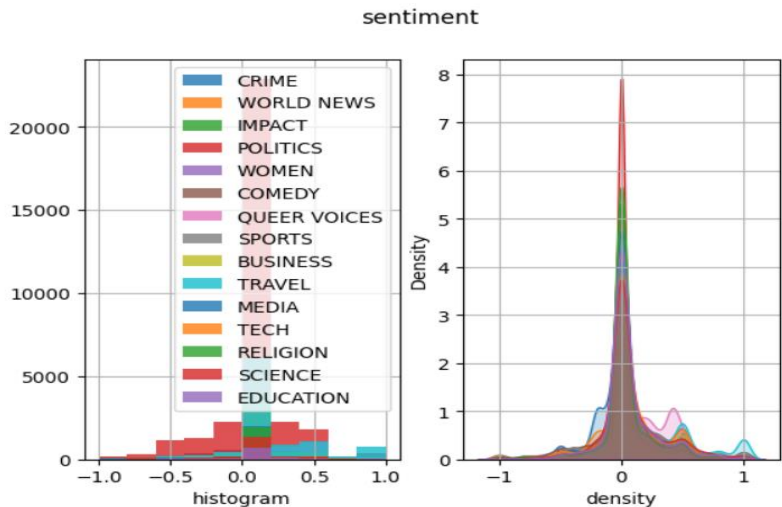


Figure 8: Proposed Method Result

And also the state of the art result is also shown in Figure 9 which refers from Taj *et al.*¹.

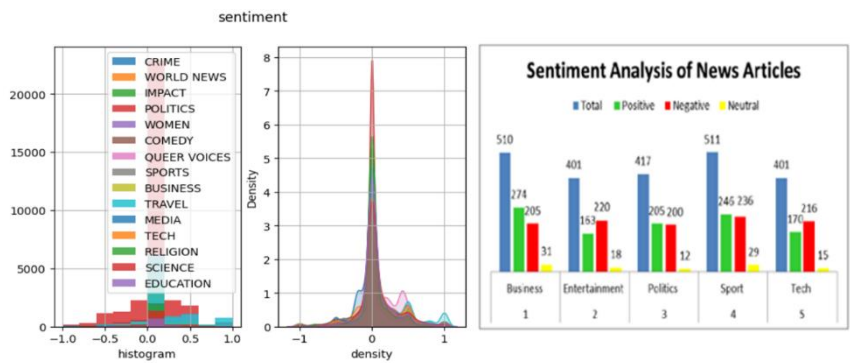


Figure 9: State of the Art Method Result

3. Statistical analysis of News Headlines Text Category

Our dataset includes statistical interpretation of data-including a number of samples per category and the average number of words per news description. From this structure, we have selected 15 variant categories to show Univariate and Bivariate distributions. This is a probability distribution of one or two variables represented by a most labeled frequency samples. If the distributions are variant then each category is predictive with different news patterns. And the word frequency refers to the number of times a term

appears in a text. We also analyze and process the most important words using N-grams, which are contiguous sequence of n items from a given sample of text. In this case, we have analyzed unigrams and bigrams of a particular category “World News” as shown in Figure 10.

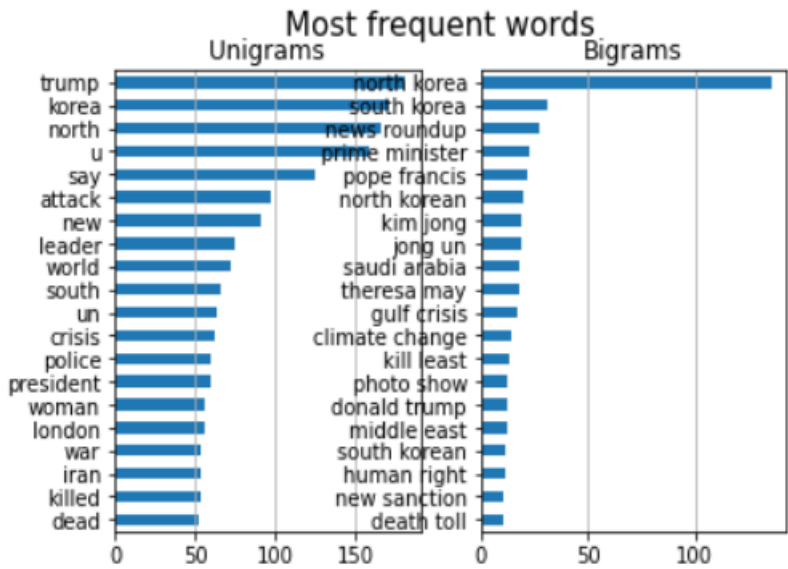


Figure 10: Unigrams and Bigrams for ‘World News’Category

By analyzing entire dataset, we categorize the sentences by assigning corresponding tags with appropriate attributes as shown in Figure 11.

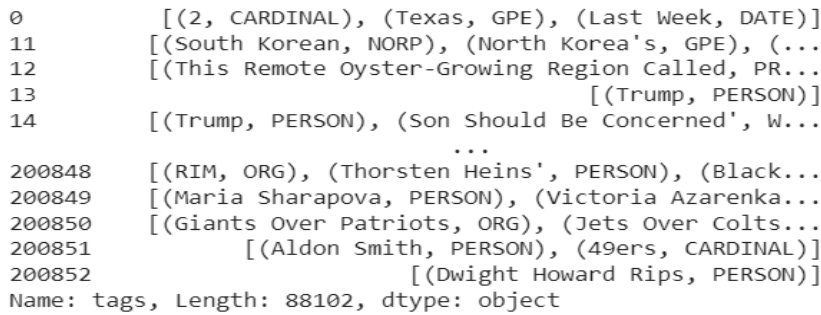


Figure 11: Preferred Tag Attributes for the Dataset

Then, we use preprocessing steps which are applied to the samples of the dataset. Tagging text with entities enables a length analysis by computing word count, character count, sentence count, average word length, and average sentence length. In addition, visualize the results by generating histograms and density maps of character count for specified categories, as shown in below Figure 12.

Acknowledgement: The authors appreciate the unknown referee's valuable and profound comment.

References

1. S. Taj, B.B. Shaikh and A.F. Meghji; Sentiment analysis of news articles: a lexicon-based approach. In *2019 2nd international conference on computing, mathematics and engineering technologies (iCoMET)*, IEEE (2019) 1-5.
2. N. Isnaini, M.S. Mubarak and M.Y.A. Bakar; A multi-label classification on topics of Indonesian news using K-Nearest Neighbor. In *Journal of Physics: Conference Series*, IOP Publishing, 1192-1 (2019), 012027.
3. O. Fuks; *Classification of News Dataset*, Stanford University, 2018.
4. F. Fanny, Y. Muliono and F. Tanzil; A comparison of text classification methods k-NN, Naïve Bayes, and support vector machine for news classification, *Jurnal Informatika: Jurnal Pengembangan IT*, 3-2 (2018), 157-160.
5. R Katari and M B Myneni; A survey on news classification techniques. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* IEEE, 2020.
6. J.T. Pintas, L.A.F. Fernandes and A.C.B. Garcia; Feature selection methods for text classification: a systematic literature review, *Artif Intell Rev*, 54 (2021), 6149–6200.
7. D.Singh and S Malhotra; Intra News Category Classification using N-gram TF-IDF Features and Decision Tree Classifier, *IJSART*, 4 (2018), 508-514.
8. F.A. Pozzi, E. Fersini, E. Messina and B. Liu; Challenges of sentiment analysis in social networks: an overview, *Sentiment analysis in social networks*, (2017) 1-11.
9. P. Harjule, A. Gurjar, H. Seth and P. Thakur; Text classification on Twitter data. In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)* IEEE (2020) 160-164.
10. T. Sabri, O. El Beggar and M. Kissi; Comparative study of Arabic text classification using feature vectorization methods, *Procedia Computer Science*, 198 (2022), 269-275.
11. Y. HaCohen-Kerner, D. Miller and Y. Yigal; The influence of preprocessing on text classification using a bag-of-words representation, *PloS one*, 15-5 (2020), e0232525.
12. Y. Wang and L. Zhu; Research on improved text classification method based on combined weighted model, *Concurrency and Computation: Practice and Experience*, 32-6 (2020), e5140.