Vol. 3 No. 2 (2025)

Spectrum of Engineering Sciences

Online ISSN
3007-3138

Print ISSN
3007-312X

Enhancing News Tweets Classification Through Pre-Processing Techniques

Rabia Latif[1]
Department of Computer Science, HITEC University, Taxila Cantt
Taxila, Pakistan. rabialatif313@gmail.com
Muhammad Khalid[2]
Department of Computer Science, HITEC University, Taxila Cantt
Taxila, Pakistan. mkmalghani@gmail.com
Samrin Fatima[3]
Department of Computer Science, HITEC University, Taxila Cantt
Taxila, Pakistan. samreenfatimah048@gmail.com
Dr. Saima Shaheen[4]
Department of Computer Science, HITEC University, Taxila Cantt
Taxila, Pakistan. saima.shaheen@hitecuni.edu.pk
Abdullah Asif[5]
Department of Computer Science, HITEC University, Taxila Cantt
Taxila, Pakistan. am6117184@gmail.com

Abstract

Today in the era of technology, social media platforms have reshaped the dissemination of news Twitter emerged as a main source for real-time news updates. As a large number of Twitter news is generated every second there is a need for a system that accurately classification of news content for better real-time media monitoring. In this research, a machine learning based approach to enhance the classification of news tweets through preprocessing techniques is introduced. A combination of different preprocessing is implemented on Wall Street Journal twitter news tweets. This preprocessing especially design for twitter includes removing URLs, removing mentions,

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

removing emoticons along with basic text preprocessing. The pre-processed text corpus is evaluated with different machine-learning models. Support Vector Machine (SVM) outperforms others with an accuracy of 95%.

Keywords: Wall Street Journal; Tokenization; Vectorization; Machine Learning Models; Deep Learning Models; Text Classification; Twitter; Preprocessing; Data Mining; Feature Engineering

Introduction

The internet and social media are becoming increasingly pervasive, as evidenced by the proliferation of well-known social media platforms like Twitter and Facebook. There, users generate a variety of text documents, including reactions, news updates, and responses. Undoubtedly, one of the most significant challenges encountered by researchers working in the domain of Natural Language Processing (NLP) is the classification of text [1]. This is because individuals employ words differently when attempting to categorize texts. A prevalent challenge encountered in Twitter news classification pertains to the evaluation of information reliance on tone and sentiment. The ability to recognize news by utilizing accumulated and stored information in context. Comprehension and interpretation of intricate and nuanced news sentences require an understanding of the linguistic complexity between ambiguous and complicated sentences. Effective documentation and the implementation of high-quality labeling can pose significant obstacles in the classification of Twitter news. The process of human classification is labor-intensive, and there are circumstances in which the volume of generated data may render manual classification unfeasible [2]. NLP is increasingly recognized and implemented across various domains due to its exceptional prowess in

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

decision-making, data processing, and human-machine interaction. By employing the processes of deciphering, comprehending, and generating human language, NLP approaches the intersection of machine cognition and human communication. Natural Language Processing (NLP) plays a significant role in various critical domains, including healthcare informatics [3], machine translation [5], personal interest recommendations [6], and legal document analysis [7]. NLP goes beyond merely extracting information from vast quantities of texts, which constitutes data for organizations and individuals. By fostering innovation and advancement, NLP facilitates the transition of the world into the digital age.

The research aims to employ various pre-processing techniques to the Twitter news dataset, the research primarily seeks to enhance the dataset's suitability for classification tasks with NLP models [4]. Traditional techniques such as tokenization and stop word removal can effectively structure data to facilitate its analysis.

However, for tweets, specialized text preprocessing remains necessary. The application of NLP models such as RNN and CNN, as well as transformers including BERT, to preprocessed text will facilitate the accurate categorization of news items into their relative classes.

The dataset is published on Kaggle, where it can be examined in a practical setting; it is one of the most comprehensive news datasets categorized on Kaggle. Following the implementation of natural language processing models incorporating machine learning algorithms including Support Vector Machines (SVM), Naive Bayes, Gradient Boosting , Logistic Regression and Random Forest, we executed various pre-

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

processing techniques pertinent to the model. The contribution of this research work can be summarized as follows:

• To introduce a news tweets dataset from Twitter.

• To develop pre-processing techniques for twitter text classification, especially for twitter news data.

• To evaluate the different machine learning models on processed text dataset.

The arrangement of the paper is as follows: The literature review is given in section II. Section III presents the dataset description, section IV presents the suggested system. Moreover, section V and section VI, respectively, present the results and conclusions.

Literature Review

To recover estimation (Sentiments) from such assets, experts have been motivated to experiment with different assumptions investigation models by the abundance of data available on long-distance interpersonal communication locales where people are expected to express their emotions, conclusions, and viewpoints. Nevertheless, experts must contend with the challenge of the language's constant evolution as it appears in content supplied by clients; the words that surround us always have an impact on word choice. There haven't been many studies done to describe vernaculars because of the manner they're usually spoken rather than written [5].

A further study was conducted focusing on the analysis of brief and informal content such as tweets. This art is obsessed with how casual sentences are constructed. In order to identify the conclusion of content, the scientists formulated an estimation highlight generator employing a hybrid approach. The developed approach was assessed to be immune to the majority of common

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

techniques in the area, including Maximum Entropy, Naive-Bayes, and SVM. The results demonstrate that sentiment feature extraction can provide higher accuracy than traditional text analysis [6]. Multi-class classifiers are used to solve problems involving outcomes with more than two class labels, and various model selection and assessment procedures are used to gauge the performers' performances [7].

In recent works Data from Twitter is used in real-time to classify spam. Preprocessing is done using text mining techniques and machine learning methods like backpropagation. As classifiers, Naive Bayes and neural networks are employed. Real-time datasets from publicly accessible Twitter data are gathered using the Twitter API. Research reveals that naive Bayes outperforms backpropagation neural networks [1]. A system that classifies tweets using user input and tweet-based attributes is proposed. One advantage of these tweet text features is that it can identify spam tweets even if the sender tries to register for a new account. Four distinct machine-learning methods were used for the evaluation, and the results showed how accurate they were [2]. The solution was created for Twitter situations that operate in real-time or very close to it. The process is used to extract the most essential features from a tweet. The Social Honeypot Dataset and 1KS are the two datasets that were used. The integration of many feature sets raises the probability of capturing various categories of spam and makes it hard for the spammers to exploit all the available feature set in the spam detection system. The tweets are categorized as spam using SVM. The Sequence Minimal Optimization Algorithm and the Waikato Environment for Knowledge Analysis were used. A dataset of tweets from Twitter was used to train the model. Based on the accuracy of

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

the system, this model has a high degree of reliability in comparison to other spam models. On Twitter, spam is identified using the KNN algorithm, the naive Bayes algorithm, and the decision tree induction method. By randomly selecting 25 active Twitter users and collecting tweets from the publishers they follow, the researchers created a data set Compared to other currently used approaches, the proposed solution has the advantage of being more practical and producing considerably better categorization results. The lengthier training times for models and the potentially costly feature extraction process are two issues with the suggested approach [5]. Utilizing the account-based and tweet-based functionalities has the benefit of increasing the accuracy rate even further [6].

Deep learning algorithms have recently shown amazing outcomes in the field of natural language processing. They use sequential layers to represent data. They are capable of automatically extracting syntactic features from sentences without the need for additional feature extraction approaches, which are time- and resource-consuming. This is the rationale behind NLP academics' interest in using deep learning models to investigate sentiment classification. CNN can learn sophisticated, multi-layered, non-linear categorization by utilizing a multi-layer perceptron structure in deep learning. CNN is therefore utilized in a wide range of applications, including speech recognition, image processing, and computer vision. The Dynamic Convolution Neural Network (DCNN) model was created by Blunsom and Klatch Brenner for text processing. To achieve sentence level classification, Kim et al. proposed classifying English text using word vectors as CNN input. Although CNN performs well in text classification, it mostly concentrates on

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

extracting local features and ignores word context, which has a significant impact on the accuracy of text classification findings. This drive for the effort led to the proposal of an integrated CNN/Bi-LSTM model. Neural network features that can be automated allow RNN to efficiently incorporate nearby locations of information in natural language processing. One RNN model called LSTM creates a large-scale neural network structure by effectively utilizing memory to prevent gradient issues in RNNs.

Unlike CNN and LSTM, RNN retains sequential information while paying greater attention to the context of feature information and fitting into non-linear relationships. Another family of neural network models that are often used in text classification is the bidirectional RNN. To enhance the performance of the RNN neural network model, bidirectional RNN combines the features of two RNN models: the forward and backward hidden layers.

This method efficiently learns word information because word semantics are categorized and connected with information about words that come before and after them. CNN is a multi-layer feed-forward neural network that decreases the computation time and complexity of the backpropagation network (BP) while improving the error in BP. Because it can identify local features using a convolution kernel and learn these features automatically for a classification solution, it has been utilized recently for sentiment classification. The convolution layer, pooling layer, and fully connected layer make up the three primary layers of the CNN model. Sentences are entered into the convolutional layer after being transformed into a numerical matrix. Words or tokens make up each phrase, and each token on the matrix table corresponds to a row or vector.

Vol. 3 No. 2 (2025)
Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

Word2Vec and GloVe models are two embedding techniques that are commonly used to generate these vectors. Using filters, the CNN model accepts vector input and extracts local features. The most crucial layer of CNN, the convolutional layer, is where the majority of feature computations are done. A function known as the convolution kernel is used by the convolutional layer to create feature maps. The pooling layer extracts the most significant information following the convolution procedure. Locally sufficient statistics are computed by the pooling layer. This technique keeps the model from overfitting, enables the pooling layer to lower feature dimensions, and reduces computing time and cost for CNN. Ultimately, a probability distribution is generated by the fully connected layer to categorize sentiment outcomes.

Several scholars have contributed to the sentiment analysis on Twitter, as shown by the literature review. Different machine learning and deep learning techniques have been used by the researchers, along with diverse datasets. The lack of datasets for multiple machine learning and deep learning models to compare on Twitter news classification is the primary research gap found. It has also been demonstrated that the suggested work analyzes tweets in real time to classify news. In light of the kind of dataset utilized, the classification algorithms employed, and the numerous studies performed on the outcomes, we thus think that the suggested methodology offers a distinctive contribution to the categorization of Twitter news.

Dataset Description

The dataset used in this research is being obtained from the tweets of a news publisher, The Wall Street Journal, published by Dow Jones: The Wall Street Journal is the worldwide leading daily

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

newspaper focused on business and financial information and news, together with the US and world news, politics, art, culture, business & technology, sports, and more. The news in the dataset was collected from the past 3 month's tweets of The Wall Street Journal and then categorized into various news categories. As a public domain, Twitter users can make use of data for research and following Twitter's Terms of Services [23]. Twitter also has an API [22] for extracting tweets in XML format but not in a bigger format. We avoided any sensitive areas or possibly harmful content to stay within ethical standards.

The data has a total 1059 documents, each of which having two fields NEWS and CATEGORY. In this data set, the "Category" field means the category news belong to. We created labels with proper names of the categories that the news belongs to. The dataset contains around 20 categories, which span a wide range of newsworthy topics, helping achieve a sample that is varied and representative of the different types of news. What groups and their prevalence in the dataset mean in terms of understanding the proportions and diversity of data is central. The "News" field is where the news is the actual tweet is put. The news is a summarizing or short cut of the news tweet. It extracts the main messages from the original long news tweets. The length, style, and content of these stories can vary, just like real news articles which can have different lengths, style and subject matter. The dataset has been constructed for news classification, a core problem in NLP that involves the assigning of predefined categories to textual data according to the content. The role of different unique categories and news texts as the main platform for assessing the efficiency of diverse NLP models is hard to overestimate.

Vol. 3 No. 2 (2025)

Spectrum of Engineering Sciences

SPECTRUM OF ENGINEERING SCIENCES

Online ISSN
3007-3138

Print ISSN
3007-312X

Through this study, we apply the dataset for both training and assessing several NLP models. It has got really easy to train a model of supervised learning where we don't need to provide pre-processed and transformed data as techniques, like TF-IDF are created to do this. This transition is an indispensable link in a line that can teach machines to learn from and understand textual information. This data set also this dataset was uploaded to the Kaggle Data set.

Proposed System

The aim of this study is to forecast the direction of assessment conveyed in tweets by analyzing tweets gathered from the Twitter dataset. To ascertain which tweet or tweets were the most successful, many techniques were applied to the collection.

Figure 1 illustrates the process by which the tweets were arranged, starting with the stage of Tweet assortment (Input Dataset), going through pre-processing, feature extraction, and classification phases, and ending with the assessment stage (wherein the outcomes of multiple classification systems will be examined).

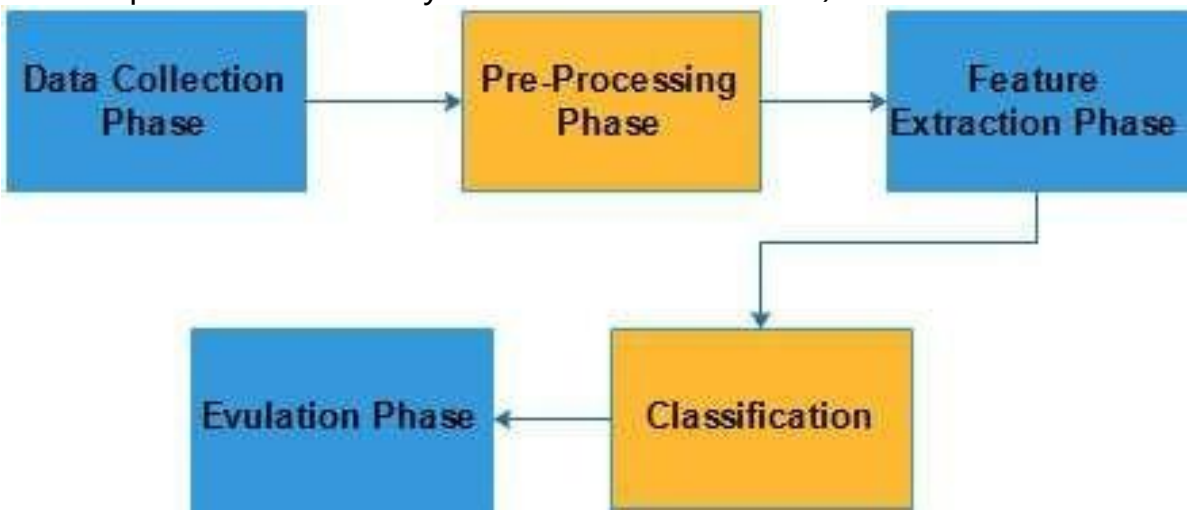

Fig.1: Outline of the Proposed Methodology

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

The main focus of this paper is English-language tweet mining. Any business might benefit greatly from deriving insights about people's sentiments and opinions about products and services from their tweets via social media.

After the tweets are gathered, the dataset is prepared and split into training and testing datasets. Test data is used to calculate the model's performance measure, whereas training data is used to train the model. By doing this, the model is able to categorize the freshly collected tweets that were obtained via the Twitter.

Data Collection

The vast quantity of tweets gathered from The Wall Street Journal Twitter account of past 3 months. We collected the tweets manually from Twitter. The collected data is categorized as health, lifestyle, finance, technology, local news, politics, travel, life and arts, markets, sports, entertainment, world news, business, science, religion, crime, Ted talks, education, and auto mobiles. This set of classified tweets serves as the training dataset.

Pre-Processing

Pre-processing is a fundamental step that takes a central role in all the Natural Language Processing (NLP) projects as a preliminary action that sets the stage for data refinement before analysis and modeling. Its operations are in filtering the data that is spoilt by HTML tags, punctuations, and unworldly symbols. However, its output is more refined and relevant only by having the essential information. Stem and lemmatize forms of words as normalization processes preserve the consistency of representation. Thereby, plurality of spelling, casing, and forms of words is being removed. Tokenization can be seen as the representation which is divided by meaning into the pieces and that is exactly what feature extraction in tasks such as sentiment analysis is made easier by it. Additionally,

Vol. 3 No. 2 (2025)
Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

pre-processing allows for language-specific processing, plus dimensionality, which significantly boosts the model's efficiency and optimizes computing system. Data preprocessing not only helps the models to perform best in the classification phase but also facilitates the users of the text data to do the exploratory data analysis and visualization which then enables researchers and practitioners to mine for valuable insights. Consequently, preprocessing is an essential precursor to NLP, as it involves normalizing and formatting the texts for better processing and dissection.

When scraping raw tweets from Twitter, one is most likely to end up with a noisy collection of data. This is a result of people's casual attitude toward using social media. Certain unusual features of tweets, such as emoticons, client references, and retweets, need to be kept reasonably apart [11]. Therefore, it is necessary to standardize raw Twitter data in order to create a dataset that various classifiers can use to learn from. To reduce the amount of the dataset and standardize it, numerous pre-processing techniques have been used [12]. The Twitter news dataset underwent multiple steps of preprocessing procedures in order to prepare the text data for analysis. Tokenization, which divided the content into discrete words and phrases, was the first step in the process. Initially, tweets undergo the following general pre-processing steps:

- Convert all the tweets' characters to lower case..
- Put the spaces in place of two or more dots.
- Remove quotation marks and spaces from the tweet's end.
- Avoid multiple spaces instead use one space.
- Eliminate any unique elements.
- This enhanced equation will also account for intermediate processes such as normalization, tokenization, removing stop words,

Vol. 3 No. 2 (2025)
Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

lemmatization, and vectorization. Breaking down the preprocessing steps in more detail:

$$P(d) = Vectorize(Stem(StopwordsRemove\left(Normalize(Tokenize(d))\right)))$$

Special twitter features are as follows:

URL (Links): Users frequently include URLs to other webpages in their tweets [13]. For content categorization, a specific URL is not important because it would result in few highlights. In keeping with this, replace every URL in tweets with just URL. When it comes to URLs, the regular expression is ((https?:// [\S] +)| (www\.[\S]+)).

User References: Every Twitter user has a handle allocated to them. Those that tweet using @handle commonly make reference to other users. The term USER_REFERENCE is used in place of all user mentions. The regular expression that corresponds to the user references is @ [\S]+.

Emoticons: Customers frequently use a variety of emoticons to convey different emotions in the tweets. Since the number of emoticons used on internet-based networking materials is constantly growing, it is challenging to fully coordinate them all. Regardless, group together a few common basic emoji's that are used frequently. Depending on whether the coordinated emoji conveys a good or negative emotion, EMO_POS or EMO_NEG replaces it.

Retweet: Retweets are tweets shared by several clients that have recently been sent by someone else. The initial letters of a retweet are RT. Since RT is not at all important for content arrangement, we remove it from the tweets. \brt\b is the standard articulation used to coordinate retweets. Following the implementation of tweet-level pre-planning and individual tweet expressions, as follows:

- Remove all punctuation from the word, such as "?!,.():;".
- Reduce more than two letter repetitions to only two letters. A few people use different characters to emphasize key words in their

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

tweets, such as "I'm sooooo happy," This is how such tweets are handled by altering them to say "I'm so happy."

- Take out "and ". By replacing terms like "t-shirt" and "theirs" with the more inclusive "tshirt and theirs," this is done to cope with them.
- Verify whether the word is legitimate and accept it if it is. A word is considered legal if it begins with a set of letters and progresses via letter sets, digits, or a single dot (.) and underscore (_).

After that, cleaning methods were used, such as punctuation removal, stop word removal, and word reduction to base form using stemming and lemmatization. Additionally, sophisticated methods were used, such as managing imbalanced classes to correct for any unequal label distribution in the dataset. Ultimately, vectorization was carried out to transform the text input into a numerical format, possibly representing the text for machine learning models using techniques like TF-IDF or word embeddings.[3][4][5][9].

Feature Extraction

A critical stage in the machine learning process is feature engineering, which entails turning unprocessed data into a set of features that can be utilized to train models. This section will look at the feature engineering approach used on the Wall Street Journal dataset, focusing on using integrations that go beyond conventional techniques.

We have chosen TF-IDF for feature engineering based on their ability to extract contextual information and semantic similarity of text input. It is be for conversion of textual into numerical form, Aims to power ML models for classification task. Here is a brief description of the method used:

- TF-IDF: TF-IDF, or Term Frequency-Inverse Document Frequency, is a classic technique used for text analysis and information retrieval.

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

It determines the importance of a word in a document by considering its frequency in a set of papers. This method helps identify words that are more relevant to the content of a text and is useful for tasks like document categorization, information retrieval, and keyword extraction. Mathematical equation for TF-IDF (combining both Term Frequency (TF) and Inverse Document Frequency (IDF)) in the context of preprocessing techniques can be written as follows:

$$TF - IDF(t, d, D) = \left(\frac{f_{t,d}}{\sum t^{'} \in df_{t,d}}\right) \times \log\left(\frac{N}{1 + DF(t, D)}\right)$$

Where:

- **t**: The term (word) for which we are calculating the TF-IDF.
- **d**: A specific document from the document corpus.
- **D**: The entire document corpus.
- $f_{t,d}$: The raw frequency count of term t in document d.
- $\sum t' \in d \quad f\, t',d$: The total count of all terms in document d.
- **N**: The total number of documents in the corpus D.
- **D F(t,D):** The document frequency of term t, i.e., the number of documents in D that contain t.

Classifiers

Naives Baye

The Naive-Bayes model is a simple and effective method for text classification and categorization. The Multinomial NB classification method in scikit-learn naive_bayes package is commonly used for this purpose. By using a smoothing parameter of 1, the Laplace smoothed version of Naive Bayes can be employed for better results. Presence features are found to be more effective than frequency features in this model because Naïve-Bayes works better with integer features rather than floats.

We used Naive Bayes model and achieved an accuracy of 0.85.

Vol. 3 No. 2 (2025)
Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

N-GRAM

N-gram is a text classification technique often used with pre-trained word embeddings like GloVe, Word2Vec, and BERT to enhance machine learning model performance. These word embeddings capture contextual and semantic information in text, improving the representation of the classification model. Tensor Flow and Keras are common platforms to integrate pre-trained word embeddings into text classification models. Embeddings can also represent various data types beyond words, like images, used in fields like fraud detection and recommendations. Using an N-Gram model, we achieved accuracies of 0.94, 0.90, and 0.93 for 1-Gram,2 Gram and 3 Gram models respectively.

Random Forest

Random forest is a gathered learning approach for regression and classification. Coined as the Random Forest methodology, the approach results in a staggering number of decision tree models, all of them based on the decisions of all those trees, though. Again same random example (Xb, Yb) is accomplished by packing for large number of tweets x1, x2, ...xn and individual assessment marks (sentiment labels) y1,y2...yn. Each arrangement tree fb is prepared to use different arbitrary example (Xb, Yb), in which b is the number between 1 and B. Last of all, the forecasts of a majority of these B trees are voted. As a classification model, we applied Random Forest to analyze our tweets data and we got the accuracy of 0.90.

Support Vector Machine

A supervised machine learning technique called Support Vector Machine (SVM) is used to solve problems with regression, outlier identification, and classification. Based on statistical learning frameworks, SVMs can use the kernel trick to do both linear and non-linear classification. Speech and image recognition, signal processing,

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

natural language processing, healthcare, and other industries all make extensive use of SVMs. Finding a hyperplane with the largest margin that divides the data points of different classes is the aim of support vector machines (SVM). SVM is a strong and sophisticated algorithm that requires less code to achieve notable accuracy. We used SVM model to classify our tweets data and we achieved the accuracy 0.95.

Gradient Boosting

Gradient boosting is a kind of machine learning that uses boosting in a functional space using pseudo-residuals as the aim instead of the usual residuals that are used in standard boosting. The Gradient Boosting algorithm combines weak prediction models, such as decision trees, to optimize a loss function. It is used in both classification and regression tasks because of its ability to detect non-linear patterns in data. In this study using tweet data, the model achieved an accuracy of 0.92.

BERT

Most commonly used NLP techniques include BERT. BERT is a pre-trained transformer-based model and demonstrates notable performance gains across a range of NLP tasks. It is especially effective at capturing contextual information. BERT has become famous for its ability to understand and convey the context of words, making it an invaluable tool for many NLP applications. We used BERT models to classify our tweet data and achieved an accuracy of 0.24.

Results

For the classification of twitter data collected from The Wall Street Journal twitter account we use the models Navies Baye, SVM ,Random Forest, N-Gram, extensions of RNN and BERT, and extensions of CNN and LSTM. For all these models we have calculated accuracy, F1-Score, precision and recall.

Vol. 3 No. 2 (2025)

Spectrum of Engineering Sciences

SPECTRUM OF ENGINEERING SCIENCES

Online ISSN
3007-3138
Print ISSN
3007-312X

Table 1: Model's Accuracy And Result

Model	Accuracy	F1-Score	Precision	Recall
Navies Bayes	0.85	0.93	0.90	0.97
SVM	0.95	0.96	0.92	1.00
Navies Baye with N-Gram 1-Gram	0.94	0.94	0.89	1.00
2-Gram	0.90	0.95	0.90	1.00
3-Gram	0.93	0.89	0.89	1.00
Random Forest	0.90	0.92	0.86	1.00
BERT	0.24	0.39	0.24	1.00
Gradient Boosting	0.92	0.94	0.88	1.00

Table 1 Indicates that SVM has demonstrated high accuracy in classifying tweets.

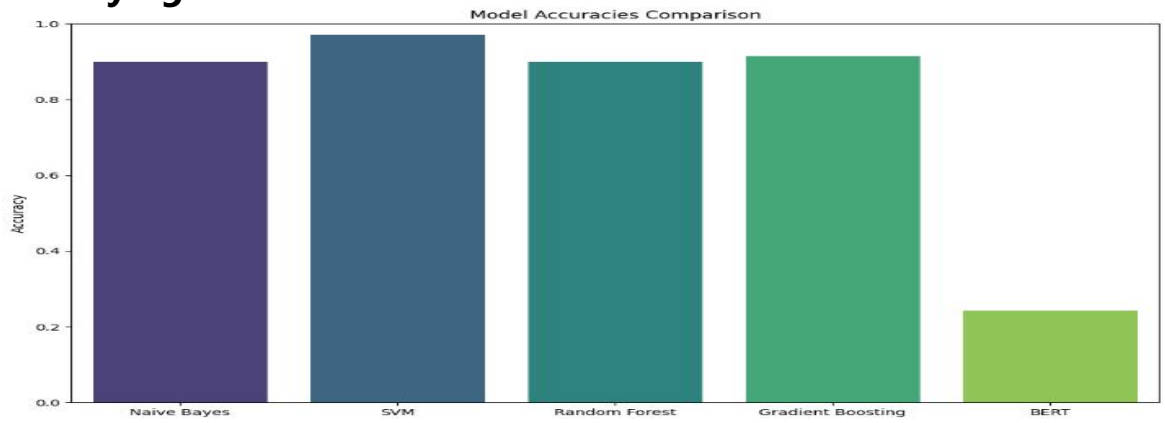

Fig 02: Models Accuracy Comparison Graph

Fig 01. Illustrates that by comparing all the models we can clearly say that SVM Support Vector Machine has shown the highest accuracy among all models.

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

Conclusions

With a total of one thousand tweets over the previous three months, the Wall Street Journal Twitter dataset is thoroughly examined in this study. In order to fully define the corpus and highlight significant data characteristics, we undertook a rigorous preparation procedure that included tokenization, cleaning, and vectorization. Then, complex topic modeling methods were used and various integrations including TF-IDF were tested Support vector machines (SVMs), logistic regression, Naive Bayes, and more complex models are among the machine learning models whose performance we compare. We tested models based on transformer topologies (e.g., BERT) and neural networks (RNN), in addition to testing multiple N-Gram models as classifiers. At last, the model is evaluated the performance of various models by comparing the results of the values of accuracy, precision, F1 Score, and recall. SVM Support Vector Machine model demonstrated the highest accuracy, i.e. 95 %, which illustrates good performance of our model and indicates that our dataset is very good. We have done an extensive testing of The Wall Street Journal Twitter dataset and the results of have proved and showed that how different machine learning models helps us classifying text. Our approach highlights the importance of pre-processing, feature engineering, and model selection in order to achieve good performance for natural language processing (NLP) issues.

Acknowledgements

I **Rabia Latif** would like to express our sincere gratitude to our supervisor, **Dr. Saima Shaheen**, for her continuous guidance, valuable insights, and encouragement throughout this research. We also extend our appreciation to **HITEC University, Taxila Cantt**, for

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

providing the necessary resources and a supportive environment. Special thanks to our colleagues, **Muhammad Khalid**, **Samrin Fatima**, and **Abdullah Asif**, for their collaboration and technical assistance. Finally, we are grateful to our families for their unwavering support and motivation during this project.

Refrences

[1] B. Wang, A. Zubiaga, M. Liakata, and R. Procter, "Making the most of tweet-inherent features for social spam detection on twitter," 2015,

[2] Sahar A. El_Rahman, Feddah Alhumaidi AlOtaib, and Wejdan Abdullah AlShehri, " Sentiment Analysis of Twitter Data", 2019 ICCIS, ISBN: 978-1-5386-8125-1, 2019.

[3] H. Gupta, M. S. Jamal, S. Madisetty, and M. S. Desarkar, "A framework for real-time spam detection in Twitter," in *Proceedings of the 2018 10th International Conference on Communication Systems & Networks (COMSNETS)*,pp. 380–383, IEEE, Bengaluru, India, 3-7 Jan. 2018.

[4] K. Subba Reddy and E. Srinivasa Reddy, "Spam detection in social media networking sites using ensemble methodology with cross validation," *International Journal of Engineering and Advanced Technology (IJEAT) ISSN*, vol. 9, no. 3,

[5] Rohit Joshi, and Rajkumar Tekchandani, "Comparative analysis of Twitter data using supervised classifiers", 2016 International Conference on Inventive Computation Technologies (ICICT), ISBN: 978-1-5090-1285-5, 2016.

[6] Asst. Prof. A Kowcika, Aditi Gupta, Karthik Sondhi, Nishit Shivhre, and Raunaq Kumar, " Sentiment Analysis for Social Media", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume: 3, Issue: 7, 2013.

[7] Geetika Gautam, Divakar Yadav. (2014). Sentiment Analysis

Spectrum of Engineering Sciences
SPECTRUM OF ENGINEERING SCIENCES
Online ISSN
3007-3138
Print ISSN
3007-312X

of Twitter Data Using Machine Learning Approaches and Semantic Analysis. IEEE 2014.

[8] Pooja Kumari, Shikha Singh, Devika More, and Dakshata Talpade," Sentiment Analysis of Tweets", IJST E - International Journal of Science Technology & Engineering, ISSN: 2349 -784X, Volume:1, Issue: 10, pp: 130-134, 2015.

[9] Pennacchiotti, M., & Popescu, A. M. (2011). A machine learning approach to twitter user classification. In *Proceedings of the international AAAI conference on web and social media* (Vol. 5, No. 1, pp. 281-288).

[10] Ajitkumar Shitole and Manoj Devare, "TPR, PPV and ROC based Performance Measurement and Optimization of Human Face Recognition of IoT Enabled Physical Location Monitoring", International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume: 8, Issue: 2, pp. 3582-3590, July 2019.

[11] Neethu M S, Rajasree R. Sentiment Analysis in Twitter using Machine Learning Techniques. IEEE 2013.

[12] Gautam, G., & Yadav, D. (2014, August). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In *2014 Seventh international conference on contemporary computing (IC3)* (pp. 437-442). IEEE.

[13] David Alfred Ostrowski," Sentiment Mining within Social Media for Topic Identification",2010 IEEE Forth International Conference on Semantic Computing, ISBN: 978-1-4244-7912-2, 2010.

[14] B. Gokulkrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perera,. Opinion Mining and Sentiment Analysis on a Twitter Data Stream. IEEE 2012.

[15] Dilrukshi, I., De Zoysa, K., & Caldera, A. (2013, April). Twitter news classification using SVM. In *2013 8th International Conference on*

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

Computer Science & Education (pp. 287-291). IEEE.

[16] Alrence Santiago Halibas, Abubucker Samsudeen Shaffi, and Mohamed Abdul Kader Varusai Mohamed, "Application of Text Classification and Clustering of Twitter Data for Business Analytics", 2018 Majan International Conference (MIC), ISBN: 978-1-5386-3761-6, 2018.

[17] Qasem, M., Thulasiram, R., & Thulasiram, P. (2015, August). Twitter sentiment classification using machine learning techniques for stock markets. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 834-840). IEEE.

[18] Phillip Tichaona Sumbureru. Analysis of Tweets for Prediction of Indian Stock Markets. IJSR 2013.

[19] Adyan Marendra Ramadhaniand Hong Soon Goo, "Twitter Sentiment Analysis Using Deep Learning Methods", 2017 7th International Annual Engineering Seminar (InAES), ISBN: 978-1-5386-3111- 92017.

[20] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang, " Sarcasm as Contrast between a Positive Sentiment and Negative Situation", 2013 Conference on Emperial Methods in Natural Language Processing, pp:704 -714, 2013.

[21] Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband," Machine Learning-Based Sentiment Analysis for Twitter Accounts", Mathematical and Computational Applications, ISSN: 2297-8747, Volume: 21, Issue: 1, 2016.

[22] Prakruthi V, Sindhu D, and Dr S Anupama Kumar "Real Time Sentiment Anlysis Of Twitter Posts" 3rd IEEE International Conference on Computational System and Information Technology for Sustainable Solutions 2018, ISBN: 978-1-5386-

Spectrum of Engineering Sciences
Online ISSN
3007-3138
Print ISSN
3007-312X

6078-2,2018.

[23] Monireh Ebrahimi, Amir Hossein Yazdavar, and Amit Sheth, “Challenges of Sentiment Analysis for Dynamic Events”, IEEE Intelligent Systems, ISSN: 1941-1294, Volume: 32, Issue: 5, pp: 70-75, 2017.

[24] Sidra Ijaz, M. IkramUllah Lali, Basit Shahzad, Azhar Imran, and Salman Tiwana, “ Biasness identification of talk show's host by using twitter data”, 2017 13th International Conference on Emerging Technologies (ICET), ISBN: 978-1-5386-2260-5, 2017.