International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
Volume 9, Issue-6 November-December- 2022
Email: editor@ijermt.org
www.ijermt.org

Tweet Sentiment Analysis using Feature Extraction and Machine Learning

Jeetendra Kumar
Assistant Professor, Department of Computer Science and Application
Atal Bihari Vajpayee Vishwavidyalaya
Infront of Koni Thana, Ratanpur Road, Bilaspur-495009 (C.G.), India

Abstract

Twitter is a microblogging website where users can publish updates (tweets) to their followers. It has evolved into a vast database of emotions. Sentiment analysis entails the examination of a text's sentiments and opinions. Opinion Mining is another term for sentiment analysis. Sentiment analysis determines and categorizes a person's attitude toward a given document or text source. Sentiment analysis of this massively generated data is incredibly useful for expressing the opinions or thoughts of the masses regarding a specific topic or trend. This approach is beneficial for customers who can use sentiment analysis to search for products, for businesses that wish to record and analyze public sentiment regarding their brands, and for many other applications. Experiments can be conducted on a variety of public tweet sentiment datasets using ensembles of classifiers based on majority voting, including Multinomial Nave Bayesian, Support Vector Machine, Random Forest, and Logical. When used alone for sentiment analysis, regression can enhance classification accuracy and the ability to compare different algorithms. According to the precision of various datasets, the proposed research identifies the optimal algorithm for each.

Keywords: *Sentiment Analysis, Feature Extraction,TF-IDF, Machine Learning*

1. Introduction

In the current digital era, the proliferation of social networking, microblogging, and blogging platforms has led to an unprecedented increase in data production. The evolving landscape of the internet has had a significant impact on the way in which people express their ideas. Individuals now communicate through a variety of online channels, such as sharing blog posts and participating in online forums. Amid this digital revolution, sentiment analysis emerges as a pivotal discipline, primarily concerned with discerning and categorizing the opinions and emotions expressed in various forms of online content, notably tweets. Online platforms, such as social media, enable people from all aspects of life to express their views on a variety of topics. When this vast quantity of user-generated content is systematically collected, analyzed, and processed, it yields invaluable insights. Sentiment analysis facilitates the evaluation of events that elicit public sentiment, the interpretation of prevalent social issues, and the collective viewpoint on these matters. In addition, it permits the identification and monitoring of concerning trends, such as bullying, self-harm, and violence. In addition to these societal implications, sentiment analysis has vast applications in the domains of marketing and campaigns, allowing businesses to effectively leverage consumer opinions on brands and products[1]. Sentiment analysis encompasses the determination of opinions, emotions, and attitudes across a broad range of source materials, such as documents, brief texts, and sentences from reviews, blogs, and news articles. This multidimensional field is broadly divided into two categories: feature-based sentiment analysis, also known as aspect-based sentiment analysis, and objectivity-based sentiment analysis. Classifying text into predefined categories, which can be binary (consisting of two classes, typically positive or negative) or multi-class (involving three or more distinct classes), is its most common application[2].

The explosive development of the internet and social networks in recent years has ushered in an era in which individuals can freely express their opinions on various web platforms. Consequently, immense quantities of user-generated comments and opinions flood the Internet, making manual analysis a formidable challenge. In this era of big data, the use of artificial intelligence technology to extract emotional tendencies from textual content is crucial for gaining a quick understanding of the prevalent public sentiment. In order to interpret the sentiment trends within these comments, the study of sentiment analysis assumes a significant role. Sentiment analysis is, at its foundation, a form of text classification that draws from a variety of disciplines, including natural language processing, machine learning, data mining, and information retrieval. It concentrates on the orientation of sentiment within comment corpuses, classifying user expressions as positive, negative, or neutral in relation to products, events, or other topics[3]. The analysis encompasses a variety of domains, such as news commentary, product evaluations, and film reviews, and serves as a window into the collective opinionss of internet users. In the process of sentiment analysis, words play a crucial role in determining classification outcomes. A crucial phase in sentiment classification is the generation of low-dimensional, non-sparse word vector representations, which frequently employ Word2Vec to capture the semantic essence of words. Notably, these word vectors typically lack sentiment data. To bridge this divide, innovative approaches embed sentiment information into traditional algorithms like TF-IDF, resulting in weighted word vectors.

Tweets from various internet users can express opinions on a variety of topics, which, when recorded as data, processed, and analyzed, can assist in evaluating events that generate insecurity, interpreting social issues and the views of the masses on them, and tracing the records of bullying, suicides, and violence[4]. As consumers express their opinions on brands and products, direct marketing campaigns can also benefit from consumer feedback. Sentiment Analysis can play a significant role in predicting the polarity of political debates and the public's approval or rejection of politicians in a highly efficient manner. Text classification is the most prevalent application of Sentiment Analysis. Sentiment Classification can be either a binary, or two-class (positive or negative) problem, or a multi-class (three or more class) problem, depending on the dataset and the requirements. Sentiment Analysis is a classification problem. Similar to large documents, sentiments of tweets can be expressed in a variety of ways.

2. Literature Review

In their 2009 research paper titled "Sentiment Analysis: A Combined Approach," Prabowo and Thelwall provide a comprehensive examination of sentiment analysis methods[5]. The paper introduces a combined approach incorporating various methods for analysing sentiment in textual data. The authors discuss the difficulties and complexities of sentiment analysis, such as the need to take multiple linguistic and contextual factors into account. They propose a framework that combines rule-based and machine learning techniques to enhance sentiment classification accuracy. The research paper is a useful resource for comprehending the complexities of sentiment analysis. In 2016, Kiritchenko, Zhu, et al. concentrate on sentiment analysis, addressing the difficulties associated with analyzing short and informal texts, such as those found in social media posts and microblogs[6]. This paper investigates various methods and techniques for extracting sentiment from these types of text and provides insights into the complexities of comprehending sentiment in brief and frequently colloquial expressions. The authors discuss the development and evaluation of short-text-specific sentiment lexicons and classifiers. Joshi, A., et al. (2017) discuss the fundamental function of sentiment lexicons and datasets in sentiment analysis (SA) systems. Sentiment lexicons are compilations of sentiment-labeled words and phrases, whereas sentiment-annotated datasets include documents (e.g., tweets, sentences) with sentiment labels [7]. Sentiment Knowledge Enhanced Pre-training (SKEP) is introduced by Hao Tian et al. in 2020 to improve sentiment analysis using pre-training methodologies[8]. While pre-training includes sophisticated sentiment analysis, it typically disregards sentiment-related knowledge, such as sentiment terms and aspect-sentiment pairs used in conventional approaches. SKEP addresses this issue by incorporating sentiment information at the word, polarity, and aspect levels by means of sentiment masking and knowledge prediction objectives. It concentrates in particular on capturing word dependence in aspect-sentence pairs. Experimental results on multiple sentiment tasks reveal that SKEP substantially outperforms strong pre-training methods and achieves new state-of-the-art performance on the majority of test datasets. In 2021, Nandwani, P. et al. will publish a review that examines the various levels of sentiment analysis, the different emotion models, and the processes involved in sentiment analysis and emotion detection from text data[9]. In addition, it addresses the difficulties of sentiment and emotion analysis in the digital age. In 2021, Pandian AP proposed a study focussed on augmenting sentiment analysis through the application of deep learning algorithms[10]. This research introduces an automated feature extraction technique that is more efficient than traditional methods, which rely on labour-intensive manual feature extraction. Traditional methods provide a solid foundation for evaluating the efficacy of features and incorporating deep learning techniques. The research involves three essential steps: the development of deep learning-based sentiment classifiers, the application of ensemble techniques and information merging, and the categorization of various model ensembles. Their study conducted an experimental analysis to determine the model with the highest performance relative to the baseline for deep learning. In 2021, Jain et al. will discuss applications of machine learning that incorporate online reviews in sentiment categorization, predictive decision-making, and the detection of fake reviews[11]. In 2022, Wang, Y., et al. will present a comprehensive review of recent developments in Twitter Sentiment Analysis (TSA), encompassing various algorithms and applications[12].

3.3. Methodology:

3.1 Collecting and reviewing the target data.

3.1.1 API- Using the Twitter API, Twitter Data can be retrieved. Its developer platform provides the ability to create a developer account and generate keys and tokens, namely, Consumer key (identifies the client), Consumer secret (client password used for authentication), Access token (defines the client's privileges), and Access token secret (similar to the consumer secret) to access tweets data in the desired format.

3.1.2 Dataset- To train a model with adequate performance, we need a relatively large training dataset. In our investigation, we utilized the "Sentiment140" dataset, which can be downloaded from Kaggle at https://www.kaggle.com/datasets/kazanova/sentiment140[13]. There are no null entries in the "sentiment" column of the 1,600,000,000-item dataset used in this experiment. The dataset description mentions a neutral class, but there is no neutral class in the training set. The data is divided into two labels, with 50% of the data bearing the negative label and the remaining 50% bearing the positive label. There are six fields in the dataset: target, ids, date, flag, user, and text. However, only the target and text fields are useful, so removing the remaining fields from the dataset will allow us to focus on the pertinent information and reduce the data's size and processing time.

3.2Data Cleaning and pre-processing.

Data cleansing and preprocessing is a crucial phase for addressing incorrect, inaccurate, incomplete, or irrelevant data elements. In our experiments, we used HTML Decoding to remove html text such as '&amp;', '&quot;, etc., Decoding BOM (Byte Order Matrix) to remove odd character patterns such as "xefxbfxdd", and removed the @ symbol, URLs, Hash tags, and numbers.

3.3Tokenization andFeature Extraction.

Tokenization is the process of dividing a text into tokens, which can be individual words, phrases, or even characters, to facilitate the analysis and understanding of textual data, thereby enabling text to be processed more efficiently. An n-gram is a contiguous string of n items extracted from a specified text sample. A 2-gram (bi-gram) is a sequence of two words, a 3-gram (tri-gram) is a sequence of three words, and so on. In contrast, feature extraction involves transforming unprocessed text into numerical or categorical features that machine learning models can use. The BoW approach is one method for converting them into numerical form. The BoW model disregards both the grammar and the order of the words. Once text data is obtained, a vocabulary list is compiled based on the entire text. The data are then represented as numerical vectors based on the dataset's vocabulary. In our experiments, we analysed the Data using the following methods:

3.3.1 Count Vectorizer – Used to transform a set of text documents into a matrix of token counts. In addition to tokenizing text documents, it creates a vocabulary of known words. A single class implements both tokenization and occurrence count.

3.3.2 TF-IDF Vectorizer– Inverse Term Frequency-Document Frequency Vectorizer quantifies the significance of a document's terms relative to their frequency in a corpus of documents. This technique assigns a weight to each term based on the frequency with which it appears in a particular document (Term Frequency, TF) and its uniqueness across the corpus (Inverse Document Frequency, IDF).Each document is depicted as a vector of TF-IDF values for each term. This representation is useful for various NLP tasks, such as document retrieval, text classification, etc., because it emphasizes the significance of terms that are both frequent within a document and rare across the corpus, thereby capturing the essence of a document's content in a concise and meaningful manner.

We have evaluated unigrams, bigrams, and trigrams in our investigation. From the results, we will determine which logistic regression classifier yields the best results and implement it. The comparison of the achieved precision is depicted in figure 1.

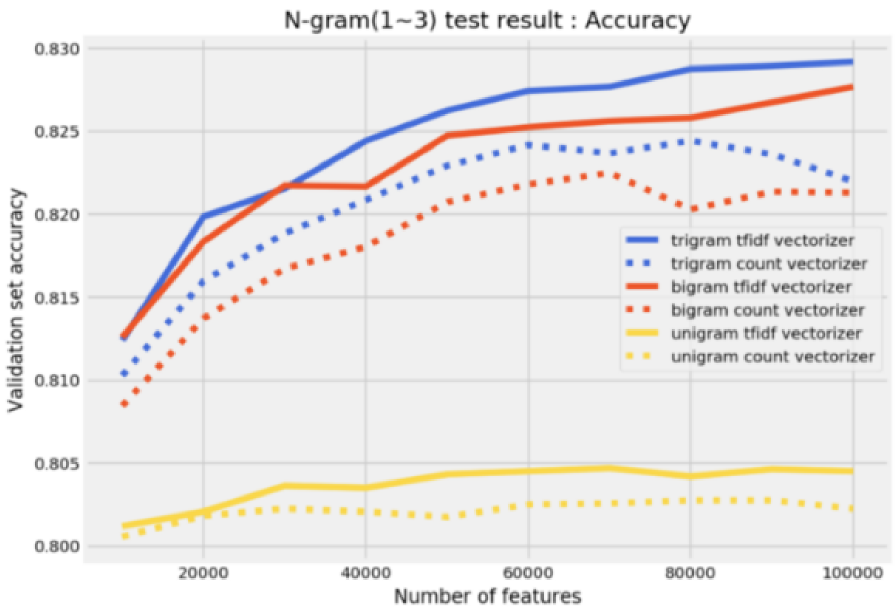


Fig. 1Comaprison of Feature extraction models

Bi-gram and tri-gram give the best results and improve the performance of the model regardless of whether Count or TF-IDF vectorizer is used, and when comparing Count and TF-IDF vectorizer in each case of uni-gram, bi-gram, and tri-gram, TF-IDF vectorizer provides more accurate results than Count vectorizer.

3.4 Classification Approach:

Classification in Sentiment Analysis can be done by using three different approaches:

3.4.1Machine Learning Approach– We have utilized a machine-learning technique and a variety of features to develop a classifier capable of identifying text documents that convey emotion. With the aid of feature selection techniques, machine learning algorithms reduce high-dimensional feature space. Feature Selection Techniques selects only essential features by eliminating the irrelevant and noisy features. Currently, sentiment analysis models based on machine learning are acquiring prominence in the field.

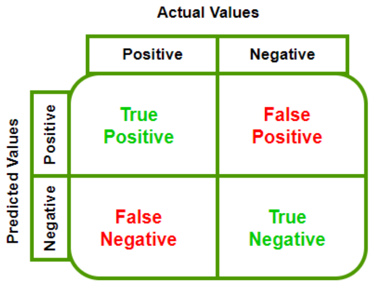
3.4.2 Lexicon – based Approach - In this approach, a variety of terms are annotated with a polarity score, which is then used to determine the overall content evaluation score. Lexicon-based Approach does not require any training data, which is regarded as a benefit. But it disregards a significant number of sentiment lexicon words and expressions.

3.4.3 Hybrid Approach– In comparison to the machine-learning approach and the lexicon-based approach, it yields the greatest results. It is rarely employed. It is a technique for sentiment analysis that combines machine learning and lexicon-based techniques.

In our experiment, we have used a Machine Learning Approach that includes a number of classifiers such as Logistic Regression, Naïve Bayesian, Support Vector machine, etc.

3.5 Performance Metrics:

3.5.1 Confusion Matrix, sometimes referred to as the error matrix, is a tabular representation utilized to evaluate the effectiveness of classification models or classifiers when applied to test data. The accuracy of a classification model is determined by dividing the sum of true positive and true negative predictions (TP + TN) by the total number of instances in the dataset (P + N).



Predicted Values	Actual Values: Positive	Actual Values: Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

3.5.2 Precision, also known as Positive Predictive Value, is the proportion of data that is correctly forecasted as positive by the employed classifier or model. The term "true positive rate" refers to the proportion of correctly identified positive instances within the total set of positively anticipated data.

$$\text{Precision} = \frac{TruePositive}{TruePositive+FalsePositive}$$

3.5.3 Recall pertains to the measurement of the proportion of positive data instances that are correctly classified as positive. Put otherwise, the true positive rate refers to the ratio of correctly identified positive instances in relation to the total number of positive instances. This metric is alternatively referred to as Hit Rate, True Positive Rate, or Sensitivity.

$$\text{Recall} = \frac{TruePositive}{TruePositive+FalseNegative}$$

3.5.4 The F1 Score can be defined as the harmonic mean of precision and recall. The harmonic mean is a distinct form of statistical measure employed for calculating averages pertaining to units, such as rates and ratios. By computing the harmonic mean of the two measures, one can obtain a reliable assessment of the model's performance in relation to both precision and recall. The formula is presented in the following manner:

$$\text{F1} = 2 * \frac{Precision.Recall}{Precision+Recall}$$

4. Result Comparison:

Before training any model considered for the experiments, we divide the dataset into three segments. The ratio of data distribution is 98:1:1, meaning that 98% of the data is the training set, 1% is the development set used for cross-validation, and the remaining 1% is the Test set. The null precision is the base precision calculated by the baseline algorithm, which has a value of 50.40 percent.

Model	Validation Set Accuracy	Time
Logistic Regression	82.92%	325.26s
Linear SVC	82.22%	358.16s
Multinomial NB	80.09%	316.77s
Bernoulli NB	79.20%	320.55s

Ensemble	82.46%	388.05s

5. Conclusion:

Sentiment analysis is a highly effective tool in the contemporary digital era, enabling the comprehension of human emotions and opinions as they are conveyed through textual content on diverse online platforms. The applications of this technology encompass a wide range of sectors, including social challenges and brand management. Furthermore, its development is influenced by ongoing breakthroughs in natural language processing and machine learning methodologies. The study of sentiment analysis on tweets is a dynamic and continuously developing area of academic inquiry. The inherent difficulties presented by the concise, casual, and dynamic nature of tweets have prompted the emergence of inventive methodologies and strategies. The ongoing prominence of social media in public discussions necessitates the utilization of sentiment analysis on tweets as a vital instrument for comprehending and effectively using the influence of online sentiment.

Reference

1. Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics, 1, 43-52.
2. Jing, L. P., Huang, H. K., & Shi, H. B. (2002, November). Improved feature selection approach TFIDF in text mining. In *Proceedings. International Conference on Machine Learning and Cybernetics* (Vol. 2, pp. 944-946). IEEE.
3. Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45-65.
4. Martineau, J., & Finin, T. (2009, March). Delta tfidf: An improved feature space for sentiment analysis. In *proceedings of the International AAAI Conference on Web and Social Media* (Vol. 3, No. 1, pp. 258-261).
5. Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2), 143-157.
6. Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research 50: 723–762.
7. Joshi, A., Bhattacharyya, P., Ahire, S. (2017). Sentiment Resources: Lexicons and Datasets. In: Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (eds) A Practical Guide to Sentiment Analysis. Socio-Affective Computing, vol 5. Springer, Cham. https://doi.org/10.1007/978-3-319-55394-8_5

8. Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H.& Wu, F. (2020). SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635.
9. Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(1), 81.
10. Pandian, A. P. (2021). Performance evaluation and comparison using deep learning techniques in sentiment analysis. Journal of Soft Computing Paradigm, 3(2), 123-134.
11. Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer science review*, *41*, 100413.
12. Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment analysis of Twitter data. *Applied Sciences*, *12*(22), 11775.
13. A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1(2009), p.12*.