International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 646-650 www.ijaem.net ISSN: 2395-5252

Sentiment Analysis using Machine Learning

Nikhil Kirar[1] Mohit Pratap Singh[2] Lakshay Arora[3] Ved Prakash[4]

[1,2,3]Student [4]Assistant Professor
[1,2,3,4]Department of Information Technology
[5]Department of Computer Science & Engineering
[1,2,3,4,5]Dr. Akhilesh Das Gupta Institute of Technology & Management, India

Submitted: 01-06-2022 Revised: 05-06-2022 Accepted: 08-06-2022

ABSTRACT
In today's world, everyone expresses themselves in some way. Numerous social media sites and Android programmes, such as Facebook, WhatsApp, and Twitter, have advanced significantly, and the current world is brimming with new ideas and information. Twitter is one of the most well-known and widely used stages on the planet. This is regarded as the primary source of opinions, as almost any active or sociable individual will invariably express their viewpoints in the form of remarks. Individuals are expressed as well as their mindset is understood through these words. Because the messages on these media are unstructured, we must first pre-process them. Six pre-handling processes are used, and then includes are extracted from the pre-handled data. There are numerous component extraction processes, for example, Bag of Words, TF-IDF, word installation, and NLP (Natural Language Processing) based highlights such as word count, item count, and so on. We investigated the impact of two highlights on the SS-Tweet dataset of opinion investigation in this paper: TF-IDF word level and N-Gram. We discovered that TF-IDF word level (Term Frequency-Inverse Document Recurrence) gave a accuracy of 1 % higher than N-GRAM and SVM model gave an accuracy of 2-3% higher than Logistic Regression.

Keywords— Decision Tree, Logistic Regression, Random Forest, Sentiment Analysis

I. INTRODUCTION

Because of the growth of content on websites like Twitter, Facebook, and Trip consultant, where people can share their opinions on products, services, and management strategies, among other things. Twitter, which has 336 million1 monthly active users, is now a major source of criticism for government, private sector, and other professional groups. Every day, over 500 million tweets are sent out on Twitter2, resulting in a massive amount of unstructured text data. Text grouping is a mechanism for handling text information generated by web-based entertainment for a variety of purposes, including email classification, web search, point displaying, and data recovery. The understanding data from the tweets posted by clients is recovered using feeling research (assessment Mining). The tweets are categorised as neutral, favourable, or bad using Twitter opinion. Order procedures in sensation examination have been introduced by a number of scientists [19, 20]. Preprocessing the message is the first step in feeling grouping; this cycle will organise the unstructured information on the web that contains disturbance into a structure that can be used for grouping. Tokenization, stop word expulsion, lowercase transformation, stemming, number elimination, and other errands are included in preprocessing. The extraction of highlights is the next step. Count vectors, pack of words, TF-IDF, word embeddings, and NLP-based text highlights are just a few examples. The next step is to select the elements; most commonly, common data, data gain, chi-square, and Ginni records are used. For order, the final stage is to use AI calculations such as support vector machines, choice trees and counterfeit brain networks. A few experts have looked into the impact of pre-handling procedures on feeling analysis from Twitter data. The effect of several highlights (TF-IDF and Bag of Words) on the exhibition of feeling exploration will be discussed in this study. Following the use of six pre-handling processes and the extraction of two types of features (TF-IDF and BOW) from the text,

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 646-650 www.ijaem.net ISSN: 2395-5252

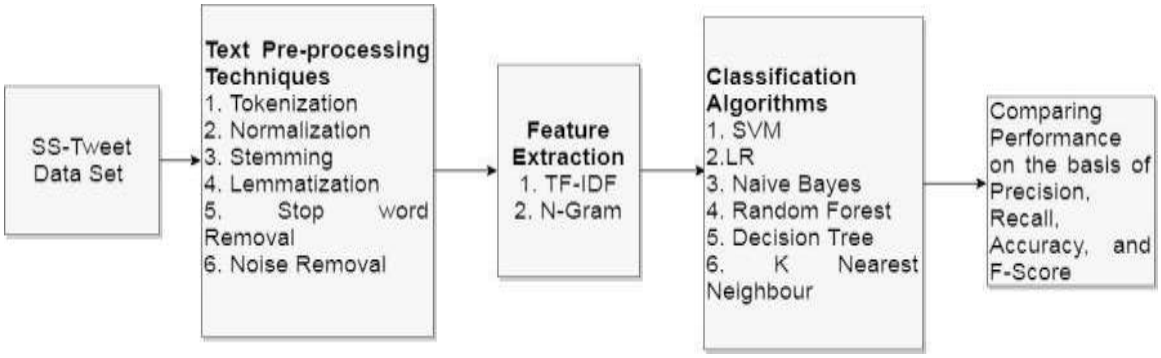
six arranging strategies are used to determine which highlights are superior. The remainder of the paper is organised as follows: Section 2 contains existing words from the feeling examination, Section 3 depicts proposed techniques, Section 4 depicts characterization calculations, Section 5 introduces execution boundaries, Section 6 contains results, and Section 7 concludes with end and future work.

II. RELATED WORK

The influence of pre-handling is investigated in the work [19]. Images, unidentified words, and contraction abound in the tweets under consideration. URLs, accentuation, client specifications, and stop words were removed, and they determined the meaning of shoptalk phrases and used an SVM classifier to perform their analysis. Vector representations were employed in this paper [2] for Natural Language processing tasks. The creators here are focused on utilising the efficiency of word vector representations to provide solutions to emotional investigation challenges. The importance of three errands has been assigned to them: recovering opinion words, determining the extremity of feeling words recognisable proof, and evaluating message opinion. They looked at the intensity of vector representations over notable text data and the nature of vectors based on different spaces. The depictions have also been utilised to compute various vector-based highlights in order to provide and truly examine adequacy. They claim that for the message feeling examination for APP surveys, they achieved F1 score and exactness of 85.77 percent and 86.35 percent, respectively. The influence of preprocessing on a dataset of cinema surveys was investigated in paper [3]. They considered removing stop words, removing refutations, removing non-English letters, stemming, and the prefix 'NOT_' while using SVM classifiers. The authors of the research [4] looked at four datasets: HCR, Sentiment140 (just 3000 Tweets), Sanders, and Stanford 1k, and looked at Bag of Words highlights, vocabulary-based elements, and Part of Speech-based highlights. They used three AI strategies: SVM and Logistic Regression as well as a combination of these. Old methodologies for feeling examination of short text miss the mark on reliance on feeling words and modifiers and essentially gather the opinion of the sentence to look for the feeling of a short message, but they figured out how to moderate the problems through feeling structure and opinion calculation standards in this paper [5]. The proposed approach in the research demonstrates how dependence parsing closes the opinion structure with social relocation and modified distance, demonstrating a strong commitment to understanding the feeling of brief text. Short text opinions are formed based on the undeniable impact of mappings between the modifier and the inclination word. Their examined findings support the practicality of the methods they developed for resolving challenges through the use of opinion structure. The authors of paper [6] considered text (e-learning critique) in Greek and separated parts of discourse elements and text-based highlights, assessing the impact of these highlights on the execution of opinion grouping. Joseph D. Prusa used 10 different component determination methods and four classifiers in his study [7]. They discover that employing a highlight selection technique will help with the presenting of feelings research. The authors of study used three levels of element extraction processes. They used classifiers such as SVM, J48. The authors of paper looked at several order techniques and component selection tactics with little to no success. In their study, the authors looked at a Twitter dataset (all 1000 comments) and used a distinct AI approach as well as a troupe approach (the majority voting) to group the comments. They used Twitter explicit elements as a contribution to the grouping classifier. In their work, the authors used API to analyse tweets regarding the Samsung universe phone and classified them as favourable, negative, or impartial. In their study, the authors used LEM2 uncomfortable set calculation to create a prompted choice rule based on Samsung universe G5 item surveys on the Samsung website. This will aid the business examiner in grasping goods in numerous aspects, each with its own set of features and relationships.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 646-650 www.ijaem.net ISSN: 2395-5252

III. PROPOSED MODEL / METHODOLOGY



We have taken right off the bat SS-Tweet dataset than we applied six pre handling procedures on the dataset and extricated highlights utilizing N-grams and TF-IDF methods. In the subsequent stage we applied seven arrangement methods and assessed four boundaries.

3.1 Dataset tweet

SS - TWEET represents Sentiment Strength Twitter Dataset. This dataset is explained physically. It contains 40000 tweets, 20000 are negative tweets, 20000 are nonpartisan tweets, and 1252 are positive tweets.

3.2 PRE-PROCESSING TECHNIQUES

3.2.1 Tokenization

This progression breaks the enormous passages called lumps of messages broken into tokens which are really sentences. These sentences can additionally be broken into words. For instance, think about the sentence, before tokenization the it is? PhD is a tuff task to take care of and after tokenization it comes: {'?', ' PhD',' is', 'tuff', 'work', 'to',' do'}

3.2.2 Normalization

There are many undertakings performed at the same time to accomplish standardization. It incorporates the transformation of all text to either upper or lower case, disposing of accentuations and change of numbers to their comparable words. This builds the consistency of preprocessing on every text.

3.2.3 Stemming

The stemming system is utilized to change various tenses of words to its base structure this interaction is hence useful to eliminate undesirable calculation of words. For instance: fishing, fish, fisher to fish, Argue, contending, contends to contend

3.2.4 Lemmatization

Lemmatization is the method involved with combining at least two words into single word. This breaks down the word morphology and takes out the completion of the word like stunned to stun, got to get and so on.

3.2.5 Removing Stop Words

Stop words allude to most familiar words in the English language which has no commitment towards opinion investigation. A portion of the stop words are "are", "of", "the", "at" and so forth. So these should be disposed of.

3.2.6 Noise removal

The datasets taken come in crude structure. We have applied manual cleaning of crude information alongside the utilization of standard articulation in NLP used to dispense with commotions. The commotion evacuation is done cautiously as it now and again kills a couple of quantities of columns of the dataset which prompts diminished exactness. The standard articulation utilized on datasets cleaning had the option to eliminate superfluous void areas and acquire information legitimate sections.

IV. RESULT AND ANALYSIS

In this paper, we thought about two elements TF-IDF (word level) and N-Grams (worth of n=2) on the Twitter sentiment investigation dataset (SS-Tweet). Table 1 shows the output(four execution boundaries i.e exactness, accuracy, review, also, f-score) of six arrangement methods (Random Forest, Decision Tree, SVM, Logistic Relapse, and KNN) utilizing TF-IDF highlight. Table 2 shows the result (four execution boundaries i.e exactness, accuracy, review, and f-score) of six arrangement methods (Random Forest, Decision Tree, SVM, Calculated Regression, and KNN) utilizing N-gram highlights. As it very well may be seen from both the tables calculated relapse is performing better in both the cases and our occupation is figure out which elements is performing better when contrasted with other, this has been displayed.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 6 June 2022, pp: 646-650 www.ijaem.net ISSN: 2395-5252

Tweets Dataset Analysis		
	Accuracy (%)	
ML Algorithms	**TF-IDF**	**N-GRAM**
KNN	71.1	65.4
Decision Tree	67.9	70.6
SVM	80.5	80
Logistic Regression	78.3	78.5
Random Forest	77.1	76.9

V. CONCLUSION

In this paper, we have consequently applied 6 distinct calculations of arrangement on the SS-Tweet dataset thinking about two highlights (TF-IDF and N-Grams). Consequently subsequent to doing opinion investigation of these tweets, that's what we established, TF-IDF

highlights are giving improved results (2-3%) when contrasted with N-Gram highlights and SVM model gave an accuracy of (2-3%) more than Logistic Regression. Accordingly we can reason that assuming we are going to utilize AI calculation for the text grouping then TF-IDF is the most ideal selection of elements as thought about by N-Gram. By generally speaking correlation of AI calculations, we figured out that strategic relapse gave best expectations of feelings by giving greatest result for every one of the four examination boundaries to be specific accuracy for both element extraction strategies in particular - N-Gram and word-level TF-IDF. Thus we concluded that TF-IDF is the best choice of features and SVM gave best accuracy.

REFERENCES

[1]. Sentimental analysis of social media content using N-Gram Graphs CSS, National Technical University of Athens, Greece

[2]. X. Fan, X. Li, F. Du, X. Li, and M. Wei, "Apply word vectors for sentiment analysis of APPreviews," 2016 3rd International Conference on Systems and Informatics(ICSAI), Shanghai,2016,pp.1062-1066.

[3]. Fouad M.M., Gharib T.F., Mashat A.S. (2018) Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble. In: Hassanien A., Tolba M., Elhoseny M., Mostafa M. (eds) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018). AMLTA 2018. Advances in Intelligent Systems and Computing, vol 723. Springer,Cham

[4]. Prusa, Joseph D., Taghi M. Khoshgoftaar, and David J. Dittman. "Impact of Feature Selection Techniques for Tweet Sentiment Classification." In FLAIRS Conference, pp. 299-304. 2015.

[5]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

[6]. İ. İşeri, Ö. F. Atasoy and H. Alçiçek, "Sentiment classification of social media data for telecommunication companies in Turkey," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 1015-1019

[7]. Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbour (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications, 3(5), 605-610.

[8]. R. M. Esteves, T. Hacker, and C. Rong, "Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets," 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Bristol, 2013, pp. 17-24.

[9]. Breiman, L., Random forests. Mach. Learn. 45(1):5–32, 2001.

[10]. Kamale, Mr Amit S., Pradip K. Deshmukh, and Prakash B. Dhainje. "A Survey on Classification Techniques for Feature-Sentiment Analysis." International Journal on Recent and Innovation Trends in Computing and Communication 3, no. 7 (2015): 4823-4829.

[11]. Das, Tushar Kanti, D. P. Acharjya, and M. R. Patra. "Opinion mining about a product by analyzing public tweets in

Twitter." In Computer Communication and Informatics (ICCCI), 2014 International Conference on, pp. 1-4. IEEE, 2014.

[12]. The Impact of Features Extraction on the Sentiment Analysis Ravinder Ahuja a , Aakarsha Chug a , Shruti Kohli a , Shaurya Gupta a , and Pratyush Ahuja