Journal of Advanced Zoology

ISSN: 0253-7214

Volume 44 Issue S-2 Year 2023 Page 3838:3846

SPECIFIC WORD EMBEDDINGS FOR SENTIMENT ANALYSIS IN TWITTER DATA ANALYSIS

[1]GOLLA CHAKRAPANI, [2]K.VENKATAKRISHNA, [3]Mr. E SHRAVAN KUMAR

[1]Assistant Professor, Department of CSE, Malla Reddy Engineering College and Management Sciences, Kistapur (V), Medchal (M), Telangana, India

[2]Assistant Professor, Department of CSE, Malla Reddy Engineering College and Management Sciences, Kistapur (V), Medchal (M), Telangana, India

[3]Assistant Professor, Department of CSE, Malla Reddy Engineering College and Management Sciences, Kistapur (V), Medchal (M), Telangana, India

Article History

Received: 12 Jan 2023

Revised: 10 Apr 2023

Accepted:01 Jun 2023

ABSTRACT: Social media sites are one of the platforms where a lot of people interact in the present, expanding world. Twitter is one popular social media platform. Tweets are shared with the general public through Twitter. Currently, significant amounts of continuous data word representation learning algorithms do not take into consider the sentimental relationships between words and instead focus only on the text's syntactic information. Specific Word Embeddings for Sentiment Analysis in Twitter Data Analysis is presented in this analysis. Weighted average word embeddings is a method that uses an adapted version of the delta Term Frequency-Inverse Document Frequency (TFIDF) a method to integrate sentiment data into continuous word representations. Sentiment Analysis experiment model makes use of a classifier described as Tailored Random Forest, which was trained the training sample. An analysis utilizes tokenization sentiment, stemming, and the removal of stop words. In this study, the development of multiplication polarity-based sentiment analysis is the main focus. In comparison to un-weighted embeddings, experiments have shown promising results. Experimental results demonstrate that described classifier gives very high predictive Accuracy, macro average Recall and Precision. Finally, they can enhance the sentiment analysis model's performance.

KEYWORDS: Sentiment analysis, Twitter, weighted word embeddings, TFIDF, accuracy, NLP.

I. INTRODUCTION

Twitter, a well-known micro-blogging service that is used by people all over the world [1], has been shaped and changed the way people get information from people or organizations they are interested.

Tweets are status updates that users can post on Twitter to share with their followers thoughts, actions, or current events. Additionally, users can respond to or repost other users tweets to interact in conversation with them. Twitter has grown to be one of the world's top online social networking sites from its beginning in 2006. Because to its various applications and the ever-increasing collection of information provided in Twitter, extracting users emotion polarities conveyed in Twitter messages has become as popular research topics [2].For example, a number of systems have been generated to keep providing political election techniques by analyzing the

polarization of Twitter user's opinions regarding political parties and candidates. Twitter sentiment analysis is also used by businesses to quickly and effectively monitor how people feel about their brands and products [3].

Sentiment can be completely absurd or even completely passionate depending on the individual. When attempting to evaluate sentiment, it is essential to mine a substantial and pertinent case of data. In general, no individual data point is relevant [4]. Any number of indirect factors can affect how someone feels about a brand or product; it is possible for someone to have a horrendous day and then tweet negatively about something about which they generally had extremely unprejudiced assumptions. In order to determine whether a person's opinion on a certain topics, products, etc. is neutral, negative, or positive, sentiment analysis is a technique for computing and stratifying a person's perspective presented in a piece of text. Worldwide, people use social media platforms to publicly express their feelings through images and text messages.

With the rise of online businesses in recent years, sentiment analysis now plays a crucial role in industrial growth. These online marketing companies' analysts keep track of the days, kinds of products, and quantities sold [5]. They keep increasing or decreasing the prices of the products on a regular basis, depending on the results. The company's success and expansion would be enhanced by accurate analysis.

Some sentiment-specific issues are still hardly addressed by deal with sentiment analysis systems. Indeed, difficult issues with negations and intensifiers in natural language processing, attempting to resolve anaphoras to communicate examinations to their objectives, sources, expressing sentiment with metaphors and considering the existing sentiment analysis approaches and systems that utilize semantic rules in particularly adding context, it is difficult to confirm the sense of opinion that has been detected.The Bag-of-Words (BoW) model is typically used in supervised approaches. This model ignores context, grammar, and even word order, has a high dimensionality, is lacking and unable to fully express the complex linguistic characteristics of words. The meanings of words are encoded into low-dimensional vector spaces using dense word embeddings in recent works.

A subfield of artificial intelligence is Natural Language Processing (NLP) [6]. The goal of this field is to establish a way to interact with individuals using natural language. It required computational and linguistic technology to create this kind of intelligent
system, which processes natural language similarly to humans. When processing a text for the first time, this NLP method is used. Tokenization, stemming, and the removal of stop words are the methods used in sentiment analysis. Using tweeter data sets, In this study, they will discuss about sentiment analysis. Then, at that point, text processing is finished utilizing a few NLP methods.

Generating continuous word embeddings is a productive feature learning technique to capture both the syntactic structure and contextual information [7]. Word embeddings are gaining popularity as a technique for sentiment analysis among researchers. They provide a weighted average word embeddings approach that utilizes a modified delta TFIDF measure to encode sentiments data in the continuously word representations. According to our knowledge, this is the

first study that contains weighted word embedding vectors utilizing the Delta TFIDF measurements.

The following is the structure of the remainder of the analysis: The literature review is given in Section II, the sentiment analysis approach is detailed in Section III, the experimental results are shown in Section IV, and the study is concluded with Section V.

II. LITERATURE SURVEY

Singh, Prabhsimran, Ravinder Sawhney Singh, and Karanjeet KahlonSingh et al, [8] analyzed sentiment using data from Twitter, focusing on the government's "Demonetization" policy from the perspective of the common citizen.The cloud and cataloging use of the sentiment analysis API (Application Programming Interface) is divided into six groups: “sad”, "no data", "very sad", "neutral" and "very happy".

A. C. E. S. Lima, L. N. De Castro, and J. M. Corchado, et. al. [9] describes the two primary methods for developing a polarity analysis framework: the algorithm was either knowledge-based (lexical dictionaries) or machine learning. As a result, they present a Twitter-specific polarity analysis framework that integrates both techniques. This model makes use of techniques designed specifically to deal with tweets and other short messages. The automated knowledge-based classifiers and machine learning algorithms work together on developing the classifications in a two-stage machine learning algorithm. This framework is modular, with different approaches for each module that can be set up according to the application domain.

Fang, Xing, and Justin Zhan, et. al. [10] worked on Amazon Online Product Review Sentiment Analysis using data from Amazon.com. Using Machine Learning Algorithms, they solved the issues with categorical sentiment polarity. In this example, they utilized numerous Random Forest, SVM (Support Vector Machine), and Naive Bayesian libraries. N. F. Da Silva, E. R. Hruschka, and Jr. E. R. Hruschka, et. al. [11] demonstrates the possibilities for tweet sentiment analysis using classifier ensembles consisting of multiple components. Additionally, they compared and demonstrate the advantages and disadvantages of feature hashing-based and bag-of-words representation methods for tweets. Finally, they demonstrate how classifiers ensembles can be created by combining bag-of-words, lexicons, feature hashing, and emoticons. Min SongMeen Chul Kim, Yoo Kyung Jeong, et.al. [12] Utilize the temporal Latent Dirichlet Allocation (LDA) to examine and verify the connection between topics obtained from tweets and associated activities. To keep track of temporally related terms and address LDA's limitations, they created the terms cooccurrence retrieval approach. Finally, authors discover a thematic coherence among users who were identified when sending and receiving mentions. D. Tang, M. Zhou,N. Yang, F. Wei,T. Liu, and B. Qin, et al. [13] create three neural networks with loss functions that effectively incorporate supervision from sentiment polarity. They acquire the sentiment embeddings from a large categorization of remotely supervised tweets that contain both negative and positive emotions.

A. Montejo-Ráez, E. Martínez-Cámara,T. Martín-Valdivia,M. and L. A. Ureña-López, et. al. [14], presents a novel method for scoring posts based on the degree to which the opinions are communicated in the text are positive or negative. SentiWordNet scores and a random walk analysis of the concepts that are presented over the WordNet graph could be combined to address

the issue with polarity classification. Many experiments would be done with the goal of examining the main problems with their system and to compare it to alternative ways, adding simple SentiWordNet scoring or supervised machine learning techniques like Support Vector Machines.

A. Hassan, D. Zeng,and A. Abbasi, et. al. [15] provides a Twitter-based text analytics platform for sentiment analysis. To address problems with class imbalance, sparsity, and representational richness, the system utilizes a complex bootstrapping ensemble. the system utilizes a complex bootstrapping ensemble. Results from the experiments show that, in comparison to other comparison tools and algorithms, in its predictions across sentiment classifications, the presented approach is more balanced and accurate. The results have significant implications for analytics on social media and social intelligence considering Twitter is one of the most important platforms for social media.

III. SPECIFIC WORD EMBEDDINGS FOR SENTIMENT ANALYSIS

Figure 1 demonstrates the architecture of Specific Word Embeddings for Sentiment Analysis in Twitter Data Analysis.

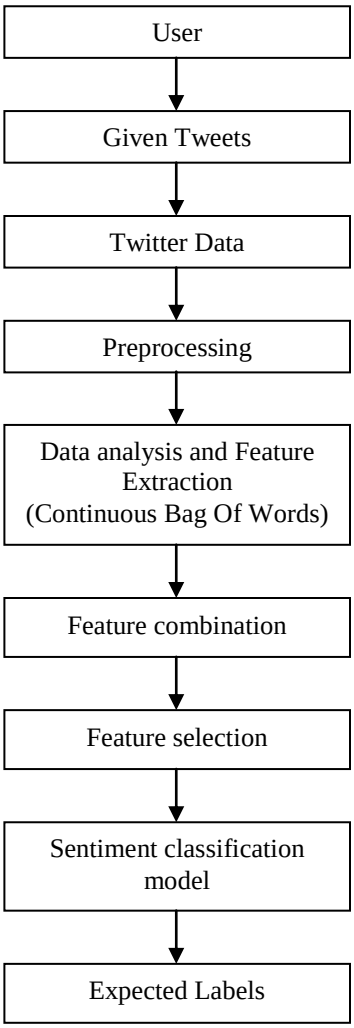

Fig. 1: ARCHITECTURE OF SENTIMENT ANALYSIS

In order to find tweets that are relevant to the topic, at first, a user enters it. The topic was sent to the server through the Twitter API, where some tweets with words related to the topic were found. The tweet dataset is produced by the users directly. For the purpose of training, they have gathered 40,000 datasets from Kaggle andGithub.

Sometimes tweets are not in a format that can be used. A number of preprocessing techniques for cleaning tweets are used to convert them into a format that can be used. The tweets are removed of all userID, twitterId, and user information. The tweets no longer contain any hyperlinks or special characters. The training dataset is also removed from duplicate tweets. Re-tweets are not included in the training data set. Text that consists exclusively of tweets remains after applying all cleaning techniques. Preprocessing stage involves Tokenizing, stop word removal, stemming and Slang removal.

During the tokenizing stages, the collected tweets are divided into different words. Accordingly, the URL (Universal Resource Locator) address and punctuation that are included in tweets are removed. The slang removal stage seeks, in accordance with language rules, to substitute standard words for words with current terms (slang words). A database list of slang words and their synonyms is developed as part of this change process. After that, compare the words that are already in use to the database's list of slang terms.The synonym will take its place if it matches. The word is omitted if it does not match. There are number of words in a sentence that indicate that the significance of the word is irrelevant. A stop word list for the Twitter API has been provided by this study. Stemming stage aims to capture a tweet's fundamental word. Porter Stemming, which can be found in the Python library, is used to obtain the fundamental word. This is done to reduce the amount of data needed for tweet processing.

In order to create the sentiment analysis model, they must extract from the text input each and every characteristic that can be broadly classified as morphological features and word N-gram features. At the point when given a constantly distributing bag-of-words portrayal of the words in its unique situation, the continuous Bag-of-Words (CBOW) model predicts the central words. Each word embedding vector is weighted according to the Delta Term Frequency-Inverse document frequency (Delta TFIDF) of the word it represents in our word embedding weighting method. Delta TFIDF is a supervised weighting measure that shows how different the positive and negative IDF values of a term’s. Words that aren't evenly transmitted between the negative and positive classes are given more weight in Delta TFIDF than words.

A classifier called tailored random forest was developed using the training sample in sentiment analysis experiment model. Random forests are modified in tailored random forests by adding two additional parameters. Regression, classification, and other tasks utilizes random forests as a collective learning source that aggregates decision trees created during training and produces the class as an output (classification) or computes the mean prediction of each tree (regression). The total number of features to be considered while determining the best split is max_features, where n_estimators represents the number of trees in the forest.

To enable the identification of sentiments, the polarity of the tweet must be processed. The evaluator feature of this sentiment analysis is present. A text bob library is used to accomplish

this, and it accurately predicts each tweet's sentiment. The number of times a tweet was re-tweeted was used to determine its popularity. A tweet was deemed popular if it acquired a higher proportion of re-tweets, and unpopular if it received less. Positive tweets gained more traction with the general public and were more likely to be re-tweeted, according to our findings. Positive sentiment was expressed in nearly half of the tweets that were widely re-tweeted. Nevertheless, it is evident that neutral tweets are less likely to be re-tweeted. The labeling is done as positive, negative, or neutral based on this polarity value.

IV. RESULT ANALYSIS

The present experiments make use of a dataset that is accessible to the public and is called Twitter sentiment. This dataset contains 40,000 tweets and is labeled as positive, negative, or neutral with fine-grained sentiment alignments. Each dataset was first shuffled as part of the analysis method that is being presented. The training phase of the dataset uses 80% of the data, and the testing phase uses 20% of the residual data.

As the most widely used measurements in the sentiment analysis task, we use the macro-averaged Recall, Accuracy, and Precision to evaluate the performance of our system. The three categories of positive, negative, and neutral macro averaged recall are as follows:

$$R_{macro} = \frac{R_{pos} + R_{neg} + R_{neu}}{3} \dots (1)$$

Where the following formula is used to calculate the recall for each polarity:

$$Recall = \frac{TP}{TP + FN} \dots (2)$$

The proportion of correct guesses to total predictions is referred to as accuracy. The capacity to accurately predict a situation's outcome is known as accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \dots (3)$$

A classification algorithm's positive predictive value, also known as precision, is determined by dividing the amount of accurate positive scores by the number of positive scores predicted by the algorithm (3)

$$Precision = \frac{TP}{TP + FP} \dots (4)$$

According to the texts FP, FN, TP, and TN, there will be in order a certain number of relevant identified features, relevant non-identified features, irrelevant identified features, and irrelevant non-identified features.

The performance of described sentiment analysis with Tailored Random Forest model is evaluated by preparing 40000 data tweet. Following that, those data provide a label with the appropriate sentiment. The polarity multiplication and word embedding weighting approach is used to analyze each tweet in those data, using the Delta Term Frequency-Inverse Document

Frequency (Delta TFIDF) to weight each word embedding vector.Sentiment analysis evaluation is represented in below Table 1.

Table 1: SENTIMENT ANALYSIS EVALUATION

Testing category	Sentiment polarity or label			Average
	Positive	Negative	Neutral	
macro-averaged Recall	0.97	0.86	0.61	0.813
Precision	0.97	0.94	0.63	0.846
Accuracy	0.98	0.97	0.64	0.863

The comparative performance analysis of described Specific Word Embeddings for sentiment analysis with Tailored Random Forest (SWE-SATRF) model and sentiment analysis with Linear SVM(SALSVM) is represented in below Table 2 in terms of macro-averaged Recall, Accuracy and Precision.

Table 2: COMPARATIVE PERFORMANCE ANALYSIS

Parameters	SWE-SATRF	SALSVM
macro-averaged Recall	98	85
Precision	97	84
Accuracy	98	86

Below Fig. 2 shows the graphical representation of comparative performance analysis of two models.

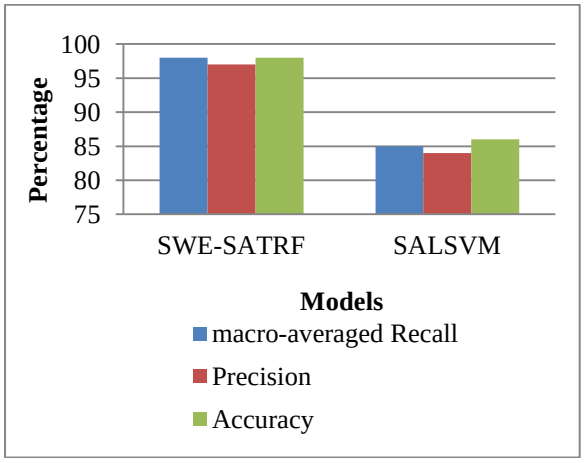

Fig. 2: COMPARATIVE PERFORMANCE ANALYSIS

From results it is clear that the performance of described Specific Word Embeddings for sentiment analysis with Tailored Random Forest (SWE-SATRF) model is better than Sentiment Analysis with Linear SVM (SALSVM). Successfully implemented is intelligent analysis for evaluating the sentiment label of the available input stream. When compared to other weighted embeddings, the performance of our sentiment-weighted word embeddings is better because they are capable of distinguishing words with opposite sentiment polarity. Obtained performance parameters of described sentiment model are macro-averaged Recall as 98%, Accuracy as 98% and Precision as 97%.

V. CONCLUSION

This analysis describes Specific Word Embeddings for Sentiment Analysis in Twitter Data Analysis.By weighting the vectors using a modified Delta TFIDF, they acquire the dense word representations. In this paper, the term "fine grained sentiment analysis" was defined, in relation to one of the most widely used and widely used social networking platforms Twitter. The estimation of extracting features from twitter data has been finished successfully and can be utilized for further analysis and decision-making in order to mining the sentiments or opinions.A classifier named Tailored Random Forest was generated using the training sample in Sentiment Analysis experiment model. They can utilize the macro-averaged Recall, Accuracy, and Precision parameters to assess the performance of our system. Obtained performance parameters of described sentiment model are macro-averaged Recall as 98%, Accuracy as 98% and Precision as 97%.The results of the studies demonstrate that the presented technique is promising and has the ability to significantly enhance sentiment analysis results.

VI. REFERENCES

[1]Siddhartha Sangwan, Swati Jain, Keshav Gupta, SandeepRao, “Social Media Sentiment Analysis- A Relative Study on Twitter Dataset”, 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Year: 2022

[2]SuyashGatkal, DhwaniPanjwani, SanketBarhate, RitikaMangla, FarukKazi, “Community Detection and Impact of Bots on Sentiment Polarity of Twitter Networks”, 2021 Asian Conference on Innovation in Technology (ASIANCON), Year: 2021

[3]Moin Ahmed, MohitGoel, Raju Kumar, ArunaBhat, “Sentiment Analysis on Twitter using Ordinal Regression”, 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Year: 2021

[4]ZhigangXu, Kai Dong, Honglei Zhu, “Text sentiment analysis method based on attention word vector”, 2020 International Conference on Modern Education and Information Management (ICMEIM), Year: 2020

[5]GuanlinZhai, Yan Yang, Heng Wang, Shengdong Du, “Multi-attention fusion modeling for sentiment analysis of educational big data”, Big Data Mining and Analytics, Volume: 3, Issue: 4, Year: 2020

[6]Md. RakibulHasan, MaishaMaliha, M. Arifuzzaman, “Sentiment Analysis with NLP on Twitter Data”, 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Year: 2019

[7]Liang-Chih Yu, Jin Wang, K. Robert Lai, Xuejie Zhang, “Refining Word Embeddings Using Intensity Scores for Sentiment Analysis”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume: 26, Issue: 3, Year: 2018

[8]. Singh, Prabhsimran, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government" ICT Express (2017)
[9] A. C. E. S. Lima, L. N. De Castro, and J. M. Corchado, "A polarity analysis framework for twitter messages", Applied Mathematics and Computation, 270, pp. 756-767, 2015.
[10]. Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data" Journal of Big Data 2.1 (2015)
[11] N. F. Da Silva, E. R. Hruschka, and Jr. E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles", Decision Support Systems, 66, pp. 170-179, 2014.
[12] Min SongMeenChulKim ,Yoo Kyung Jeong, "Analyzing the Political Landscape of 2012 Korean Presidential Election in Twitter ",Intelligent Systems, IEEE (Volume:29 , Issue: 2), 2014 IEEE.
[13] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2014, pp. 1555–1565.
[14] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López, "Ranked WordNet graph for sentiment polarity classification in twitter", Computer Speech and Language, 28(1), pp. 93- 107, 2014.
[15] A. Hassan, A. Abbasi, and D. Zeng, "Twitter sentiment analysis: A bootstrap ensemble framework", IEEE International Conference on Social Computing (SocialCom 2013), pp. 357-364, 2013.