IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 1, December 2022

W2V and Glove Embedding based Sentiment Analysis of Text Messages

Mrs. R. Priya[1] and Dr. S. Sujatha[2]

Associate Professor, Department of Information Technology[1]
Professor and Head, Department of Computer Applications[2]
Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Tamilnadu, India[1]
BIT Campus, Anna University, Tiruchirappalli, Tamilnadu, India[2]
rpriya.it2017@gmail.com[1] and sujathaaut@gmail.com[2]

Abstract: *Social media may be a capable communications medium, with a broad impact over cities as well as farther regions. Computerized change has not as it were affected businesses and made the world more available. It has had a long-lasting effect on the way individuals communicate and has become a necessary portion of their lives. For occurrence, WhatsApp has re-imagined the culture of IMs (moment informing) and taken it to a whole new level. The chat message will be a critical source of data to know the intention of an individual. In this consideration, it is being analyzed whether individuals are beneath stretch, and what almost their attitude towards others is. This could be surveyed utilizing W2V and Glove embedding, which can discover hidden semantic structures within the text body. The embedding learns the relationship between the words to build the representation. This is achieved by various methods like co-occurrence matrix, probabilistic modeling, and neural networks. The proposed framework clearly portrays the inserting of words, classification of extricated words, and giving an alarm to the concerned individual beneath crisis cases. It too gives nonstop monitoring of the client.*

Keywords: Sentiment Analysis, W2V, Glove

I. INTRODUCTION

Data mining helps businesses extract unknown insights from data. This information supports business decision-making processes, helping them make data-driven decisions. This is the key to the growth of any organization. When most people hear the term “Data Mining”, they think of digital data. The reason is that in most cases, data mining involves extracting information from digital data. However, this does not mean that data mining does not apply to text and speech data. The fact that both text and voice data contain a wealth of information can be useful for decision-making. For example, social media comments about your brand can help you learn what your customers and potential customers think about your brand.NLP (Natural Language Processing) helps data engineers extract information from natural languages like English.

Sentiment analysis is a machine learning tool that examines texts for extremity, from good to pessimistic. Using textual examples of emotions as training material, machine learning tools learn to automatically recognize emotion without human intervention. Simply put, machine learning enables computers to acquire new skills without being explicitly programmed to do so. It is possible to train sentiment analysis models to read beyond just definitions to comprehend things like context, sarcasm, and misused words. For instance: The interface is extremely intuitive. Yes, exactly. A degree in engineering would be helpful.

The phrases "super user-friendly" and "helpful" could be interpreted positively when taken out of context, but this is clearly a negative comment. Computers can automatically process and comprehend text data using sentiment analysis, saving hundreds of employee hours.Generally, by collecting the person's chat messages, we can predict their intentions or mood. In order for an automated system to learn and predict a person's mood, textual information must be represented numerically. Applying data mining techniques to a person's text messages is the best way to learn about their feelings. When the elements are chosen classifiers for the framework are prepared to utilize a preparation set. Using data mining techniques to the text messages that a person sends, one can determine a person's positive, neutral, or negative sentiment. Activities that inform the expressive portion of any text message convey the true essence of the

DOI: 10.48175/IJARSCT-7677

conversation between the counterparts, even if a person exhibits negative or miss behavior. Therefore, analyzing emoticons in text messages is crucial to understanding the message's true meaning. The objective is to collect data from the user's text messages for various purposes, including sentiment analysis.

II. PRELIMINARIES

They suggested a method for selecting plans and extracting data from a large amount of data. Data sentiment analysis is used to extract knowledge from a large amount of data. Opinion mining is another name for sentiment analysis. This is the field of study that looks at how people feel, think, and feel about things. These things could be things like surveys of people, issues, or movies. Text messages from users of social media platforms like Facebook and Twitter display their status or sentiments. People post comments on their Facebook accounts about any relevant topic that interests them [1].

Facebook, Twitter, Linked In, and Instagram are just a few of the Web 2.0 platforms that have made it possible for citizens to voice their opinions on a wide range of topics, from entertainment to education. Blogs, tweets, posts, status updates, and other types of content can be found on these media platforms, which store a lot of data. The purpose of sentiment analysis is to ascertain the polarity of feelings like happiness, sorrow, grief, hatred, rage, and affection as well as opinions from online text, reviews, and posts. Based on the text that has been provided, opinion mining identifies the sentiment in the text.

The use of native language, slang, forthcoming emoticons, characters' repetition, misspellings, and the short form of words have all contributed to its complexity. Thus, the meaning of each word is appropriately identified. Data mining devotes a lot of time and effort to sentiment analysis, which has grown into a hugely popular field of study. Because opinions play a crucial role in human activities and behaviors [2], this method is widely used in every social and business setting.

They suggested in this paper that the scope of this research includes sentiment analysis. Users' opinions can be classified as positive, negative, or neutral using sentiment classification. Only the text is used to analyze this process. The text messages are analyzed using data mining classification techniques like naive Bayes, KNN, and neural networks. The way it approached a number of issues set it apart from traditional thematic research thanks to the sentimental analysis. In this paper, near investigations of opinion examination from various explorers were talked about. Comparative analysis using various machine learning techniques. This study examines in greater detail the pre-processing steps, which are particularly challenging when the comments are written in unstructured language [3] and are crucial to the success of the sentiment analysis process.

In this paper, sentiment regarding users' medical conditions is examined using information from social media. Benchmark is set up for the examination is finished. To this end, the medical forum website "patient.info" has been crawled for user-generated opinions about medical conditions. We restricted our efforts to a few of the areas, including allergy, asthma, depression, and anxiety. In order to predict the sentiment, medical sentiments are identified. The medical condition, treatment, and medication of the user are inferred. After that, for the purpose of evaluation, medical sentiment analysis is carried out using a deep convolutional neural network. The LRE map makes the resources available to the community for further study [4].

Users of social media explain the roles social media play in tourism in this chapter, which provides a comprehensive picture of the landscape. There is a significant discussion of the consumer and business impacts of social media, as well as suggestions for social media research opportunities and significant difficulties in terms of methods, data quality, and ethics. It concludes by urging more refined and methodical approaches to the study of social media in order to continuously enhance theoretical and practical comprehension of their nature and effects [5].

For sentiment analysis, the Lexicon approach is used to calculate sentiment based on the semantic orientation of words or phrases in text messages. 6].

A collection of positive and negative words is compiled for this strategy. Both positive and negative values are assigned sentimental values. Manual [7] and automated [8] approaches to the creation of dictionaries have both been proposed.

In lexicon-based approaches, a text message is typically used as a bag of words. All of the positive and negative words are given sentiment values from the dictionary based on how the messages are shown. To improve the prediction regarding the message's overall reception, a mathematical function like sum or average is used. A local context aspect value, such as intensification or negation, is taken into consideration in addition to sentiment value.

In recent studies [10], word embedding has been found to be one of the most effective methods for text modeling and feature extraction in sentiment analysis. Word embedding turns words into vectors with real numbers of a certain length. This vector's dimensions each represent a word's fundamental characteristics. A low-dimensional and non-sparse word vector representation is essential for effective sentiment classification.

III. PROPOSED FRAMEWORK

In the event of an emergency, the proposed framework can notify users. Then, it can clearly convey messages that are positive, neutral, or negative.

3.1 Methodology for Proposed Work

From the above literature review, most of the works have been carried out using emoticons and context-dependent to find the polarity of text messages conversed. In the proposed work, W2V algorithms employ a neural network model that, once trained, can either suggest a word to complete an incomplete sentence or identify synonyms and antonyms. Glove word is a combination of two words- Global and Vectors. In-depth, the Glove is a model used for the representation of the distributed words. This model represents words in the form of vectors using an unsupervised learning algorithm. This unsupervised learning algorithm maps the words into space where the semantic similarity between the words is observed by the distance between the words. These algorithms perform the Training of a corpus consisting of the aggregated global word-word co-occurrence statistics, and the result of the training usually represents the subspace of the words in which our interest lies. For which, chats of each person are retrieved for evaluation. It provides a novel hybrid-based approach using emoticon and contextual word identification to classify the sentiment words of chat messages and also to establish a classifier model to produce better accuracy than the existing method.

3.2 System Architecture

The user's text messages are the source of the data. Preprocessing should take place prior to receiving the data from the user. The preprocessed information is then given as a contribution to the Word vectorization unit in which the vector for each word is determined for assessment as AI calculations such as order, and relapse require numbers as opposed to strings.

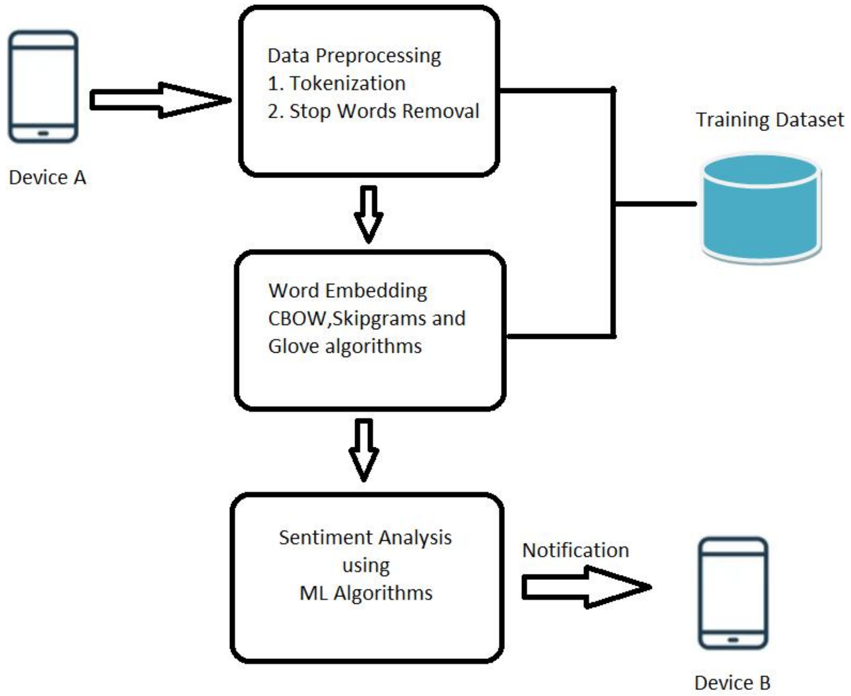


Figure 1: Proposed system architecture

DOI: 10.48175/IJARSCT-7677

After that, the cleaned data are fed into the sentimental analysis region along with vector values. Unsupervised learning algorithms are utilized in this region. Training the classifiers for classification results in their being stored in the test set for future use. For classifications, the training dataset is compared to new test data. In view of the classifiers, instant messages are assessed and expected to be as sure, negative, or unbiased. The alert messages can be created and sent to the intended recipients if any negative warnings have been received.

3.3 Modules and Descriptions

Data preprocessing after the user's text messages have been gathered. Preprocessing of the data takes place. The bulk of the project's time is spent getting the data ready. The following actions ought to be taken.

1. Data cleaning is the process of removing tags and punctuation from text messages.
2. Tokenization is the process of converting sentences into words.
3. Removing stop words — frequent words such as” the”, ”is”, etc. that do not have specific semantic
4. Shortened words: Chats also include shortened words like gud,mrng,enjoy etc.
5. Recognition of Emoticons: The emoticon list includes commonly used emoticons like happy or sad faces, just like the vocabulary list.
6. Stemming is the process of reducing a word to its root by omitting inflectional characters, typically a suffix.
7. Lemmatization: This method removes inflection by identifying the part of speech and utilizing the language's extensive database.

3.4 Feature Extraction

Feature extraction is the process of mapping textual data to real-valued vectors. After pre-processing, Word2Vec uses a large corpus of text messages as its input. It processes the data and generates a vector space in which each distinct word is given a vector from the space. Word vectors in the vector space are positioned so that they are in close proximity to one another. Sentiment prediction using Machine learning algorithms

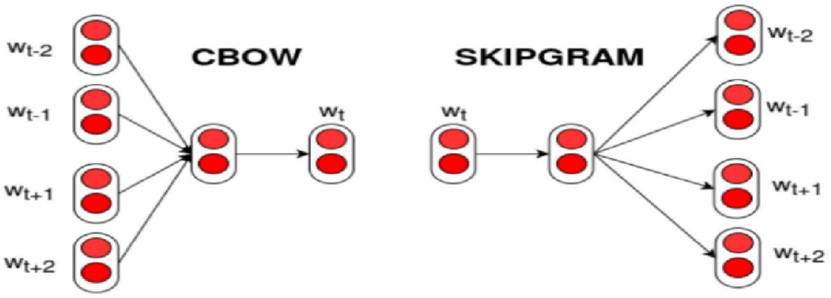
3.5 Procedure for Training the Word2vec Model

The Word2Vec family of models and optimizers aids in the acquisition of word embeddings from a substantial corpus of words. Word2Vec can be used to represent words in two primary ways.

The Continuous Bag of Words (CBOW) Method predicts the words that can be inserted into the middle of a sentence based on the surrounding context to help complete an incomplete sentence. The few words that come before and after the predicted word determine the context of the prediction. Because the order of the words in the context is irrelevant, these techniques are referred to as "bag-of-words" methods.

Skip-Gram Method: Given a current word in the same sentence, this method is used to predict the context words or surrounding words. Using embedding weights, the hidden or embedding layer of the Skip-gram model predicts the context words for each word in the extensive corpus.

The CBOW and Skip-Gram models' fundamental architectures are depicted below.



$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0} logp(w_t|w_{t+j})$$

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0} logp(w_{t+j}|w_t)$$

3.6 Procedure for Training the Glove Model

The word embedding on the word-context matrix of the glove model is accomplished through the use of the matrix factorization method. The process begins with the construction of a substantial matrix containing the co-occurrence information for the words. The basic purpose of this matrix is to use statistics to determine the relationship between the words. The information regarding the words' occurrence in various pairs is provided by the co-occurrence matrix.

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

3.7 Result Aggregation and notify the user

The system can receive the actual input after the classifiers have been trained. Users can actively participate in a number of chats during a specific time period. The individual results are then combined into a final, and the user's overall sentiment analysis is measured. Every conversation's emotions will be predicted by the system. In the end, all of these results are combined into a single sentiment value. The sentiment analysis of the user is concluded by these accumulated values of emotion, and an alert for a negative message is sent to the associated individuals.

IV. CHALLENGES AND CONCLUSION

4.1 Challenges

- One of the challenges is to extract useful data from the user.

4.2 Conclusion and Future Enhancement

The proposed model takes input from the data set that was created by collecting all of the text messages that the user sent and received. all of WhatsApp's messages. After pre-processing the messages, keywords are retrieved from the data sets. The word embedding method is used to relate words semantically after pre-processing. The context and co-occurrence of the words are what are used to learn the Word2Vec embeddings. The vectors keep the semantic and syntactic relationships. Man, woman, the king and queen, the sun, and the day all have similar vectors, for instance. The overall co-occurrence of the words in the corpus is the basis for glove embeddings.

The glove is based on the co-occurrence of words across the entire corpus, whereas Word2vec tries to capture co-occurrence one window at a time.

Classifying algorithms work better when weights are attached to the data set. Using the unsupervised algorithm, the next step is to use the classifying algorithms to classify the conversations as "positive," "neutral," or "negative." It proves to be extremely efficient for such computations and has a significant increase in efficiency.

The sentiment lexicon and strings that are utilized similarity functions can be translated into other languages in the future, and an algorithm may be developed to support these languages.

REFERENCES

[1]. Farhan Laeeq ,MD.TabrezNafiz and MizraRahilBeg “Sentimental Classification of Social Media using Data Mining,” June,2017.

[2]. PalakadBaid,ApporvaGupta,NeelamChaplot “Sentimental Analysis of Movie Reviews using Machine Learning Techniques”December,2017.

[3]. Sandeep Nigam, Ajit Kumar Das, Rakesh Chandra Balabantaray “Machine Learning Based Approach To Sentiment Analysis“October,2018.

[4]. Shweta Yadav, Asif Ekbal, SriparnaSaha, Pushpak Bhattacharyya “Medical Sentiment AnalysisusingSocial Media: Towards building a Patient Assisted System”2018.

[5]. Ulrike Gretzel “Tourism and Social Media” January,2018.

[6]. RupinderKaur,Dr.Harmandeepsingh, Dr.GauravGupta“Sentimental Analysis on Facebook comments using Data Mining Technique”, August,2019.

[7]. Taboada M, Brooke J, Tofiloski M, Voll , Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist J 267–307

[8]. Tong RM (2001) An operational system for detecting and tracking opinions in on-line discussions. In: Working Notes of the SIGIR Workshop on Operational Text Classification, pp 1–6

[9]. Turney P, Littman M (2003) Measuring praise and criticism: inference of semantic orientation from association. ACM Transact Inform Syst J 21(4):315–346

[10]. B. Naderalvojoud and E. A. Sezer, "Sentiment aware word embeddings using refinement and senti-contextualized learning approach," Neurocomputing, vol. 405, pp. 149–160, 2020.