**Abstract**

Twitter is a platform used by people to express their opinions and display sentiments on different occasions. Twitter is one of the non-traditional data sources with unlimited potential. It contains a large reserve of data sets which are easier to access and collect when compared to others. One of the available tools is twitter API which lets us collect data and various other information about the tweets. The tweets are collected through Tweepy and the tweets are being labeled through Text-Blob.The process of analyzing the tweets and the texts and emojis in them to classify the opinions of the people and their reaction towards the tweet, along with the viewpoint the owner of the tweet into positive, negative or neutral. To increase the processing of the tweets are preprocessed during collection of data itself. For accuracy we will create a vocab using the W2R (word to vector) model containing all the related words. The words are vectorized with the relationship among them. This makes the prediction easier. Then through implementation of CNN (conventional neural network) we insert the embedding layer which is formed with vectorized words and a LSTM (long-short-term-memory) layer. The accuracy value of the model is created and we can analyze the tweets of the particular domain. The Dataset for each text will be chosen separately by using Entity Identification. The major challenges such as Domain Dependency, World Knowledge and Pragmatics can be handled through selection of major entities and obtaining data related to it. This process can give us the current condition or status of the topic involved so that it will be easier to find out the anomalies. Another major challenge Thwarted Expectations can be handled using emoji and hashtags present in the obtained text. The more detailed preprocessed data helps in better end results and the model is a hybrid approach which will outperform the baseline.

**Keywords:**

Tweet’s Text Preprocessing, Collection of Tweets, Sentiment of Tweets, Polarity of Tweets, Model Creation, Model Testing.

**Introduction**

Twitter is a free data collection site where people update about their status in their social networking platforms.The tweets may contain more advanced features which can be researched in the future like the involvement of videos, links, images, etc. This gives us the quality data which minimum cost as the only being the server for collecting the data. It is one of the fastest growing research areas in computer science. Twitter is a service for social networking including friends, family, coworkers to communicate and keep others updated their status.

These huge loads of tweets can be used wisely to train our Natural Language Processing models. To extract the polarity of tweets sentiment analysis is used. The values obtained from the prediction can be used for various fields such as the opinion of a brand or view of implementing new models or policies. Through sentiment analysis we can determine the tweet is positive, negative or neutral. We derive a sentiment of a statement or a text in sentiment analysis.

Sentiment analysis can be briefly divided into two types: lexicon based modeling and machine learning based models. Lexicon based models are unsupervised as the scoring method is done through lexicons to calculate the opinions. Whereas machine learning approach involves use of feature extraction and training the model using feature set and some dataset.

Aim of twitter sentiment analysis is that it allows you to keep track of what’s being said about your product or service on social media, and can help you detect angry customers or negative mentions before they escalate. Doing sentiment analysis is hard to do in Twitter as we have seen the tweets and json file in it.

The challenges faced during the extraction of tweets from twitter include the following:

* There is only a limited size of texts in the tweet which can go up to a maximum of 140 characters which lead to limited statements and leading to lack of features.
* The native way of speaking of a particular region can be different from other regions and the way we speak is constantly changing and we need to implement ways to handle them.
* Sometimes links and hashtags and references are present in the tweets. We must use them carefully or else it will affect the polarity of our predictions.
* Expressing opinions cannot be standardized as they are various ways of representing them. As they are not standard for expressing them, features obtained may be hard.

**Existing System**

There are various methods present for handling sentiment analysis; we must use the optimum for each case. By training on a known dataset, the machine learning method employs different learning algorithms to ascertain the sentiment. The lexicon-based method involves determining a review's emotion polarity based on the semantic orientation of its words or sentences. A text's subjectivity and viewpoint are measured by its "semantic orientation".

The rule-based technique scans a document for opinion words before classifying it according to the proportion of positive and negative words. It takes into account a variety of classification criteria, including dictionary polarity, negation and boosting words, idioms, emoticons, and mixed viewpoints, among others.

Each review is represented by statistical models as a combination of latent features and ratings. In order to cluster head words into aspects and sentiments into ratings, it is believed that aspects and their ratings can be represented by multinomial distributions. Another categorization is focused primarily on the organization of the text classification.

At the document, phrase, or word level. Sentence- or word-level classification can express a sentiment polarity for each sentence in a review or even for each word, as opposed to document-level classification, which seeks to identify a sentiment polarity for the entire review.

According to our research, the majority of approaches concentrate on the document-level. Additionally, we may distinguish between techniques that aim to score a review globally versus techniques that gauge the strength of sentiment for various parts of a product. Most approaches to global review classification that depend on machine learning simply take into account the polarity of the review (positive/negative) into account. More linguistic variables, such as intensification, negation, modality, and discourse structure, are used in solutions that aim for a more precise classification of reviews (Such as three- or five-star ratings).

They are 4 main models used in sentiment analysis. They are:

1. Rule based or Lexicon based approach :

It counts the number of individual values with positive and negative polarity to calculate the score. This method is not preferred when combinations of words are not proper and negation words are more and it is not used as it needs quick but constant maintenance.

1. Automated approach or Machine Learning approch :

It is one of the most standardized ways for sentiment analysis. Here we will collect labeled data with their respective polarities and we will process our data and train the algorithm.

1. Deep Learning Models :

These include Naive Bayes sentiment analysis approach and Deep learning algorithms.

Sentiment analysis using Deep learning Natural language processing methods are able

able to find the insights through the layers from the dataset collect from the twitter API to

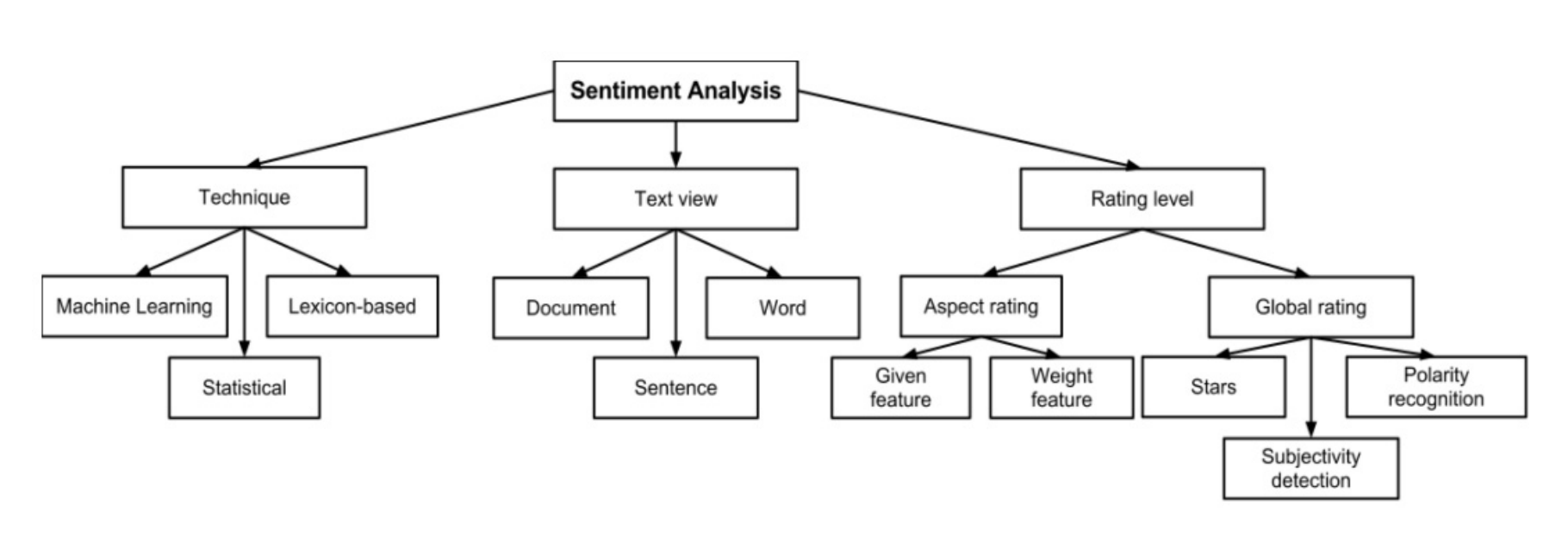
perform sentiment analysis.

1. Hybrid Approach :

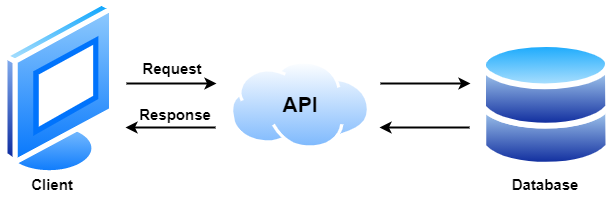
It includes both the traditional way and the newly improved neural network learning

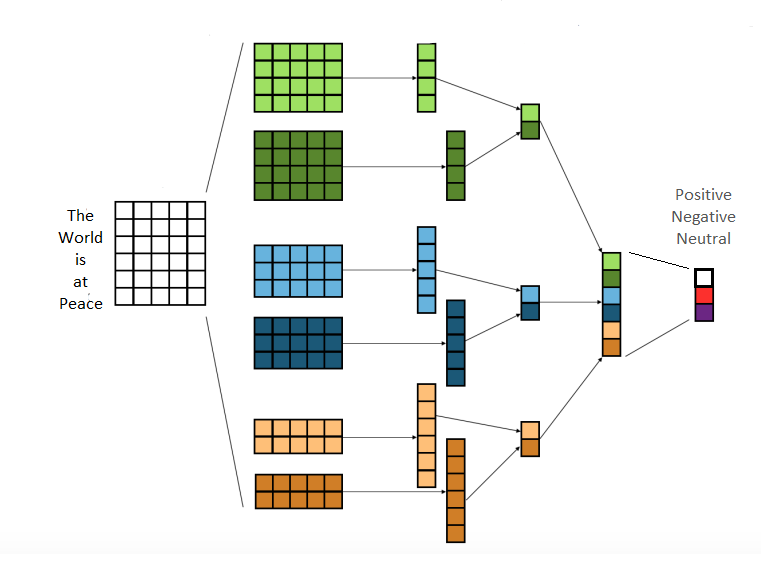
methods. This process gives us an efficient and optimum way of processing the dataset.

To the present day there are various other methods which are being introduced. We will use the proved theorem and methods to find an optimal solution for finding the polarity of our tweets.



**System Diagram**





**System Modules**

Here we will implement our system from the data collection process itself. We will be using twitter API for data collection for sentiment analysis of twitter data. Major obstacle in twitter analysis is domain dependence and world knowledge. To overcome these we will collect data directly from the topics where the tweets are tested.

We will be collecting 1.6 M data from the twitter API. These datasets are divided into training datasets and testing datasets. Using the Online TextBlob library we will append the texts along with their respective polarity of the texts. Before we Use the Textblob we will convert the emojis in the text and convert them into the appropriate meanings. This is one of the reasons why twitter is considered one with the most potential to develop the Natural Language processing where we can include other information present in the tweets.

The Preprocessing of data does not end here as they are various steps involved in making the data adaptable to our model that is being created. The first step always practiced in NLP is Removal of Stop words which are commonly used words to make sure phrase structure does not get affected. Then Tokenization is done where the text is splitted to array of words. Then Stemming is done to reduce each word in the array to reduce it to their respective root words.

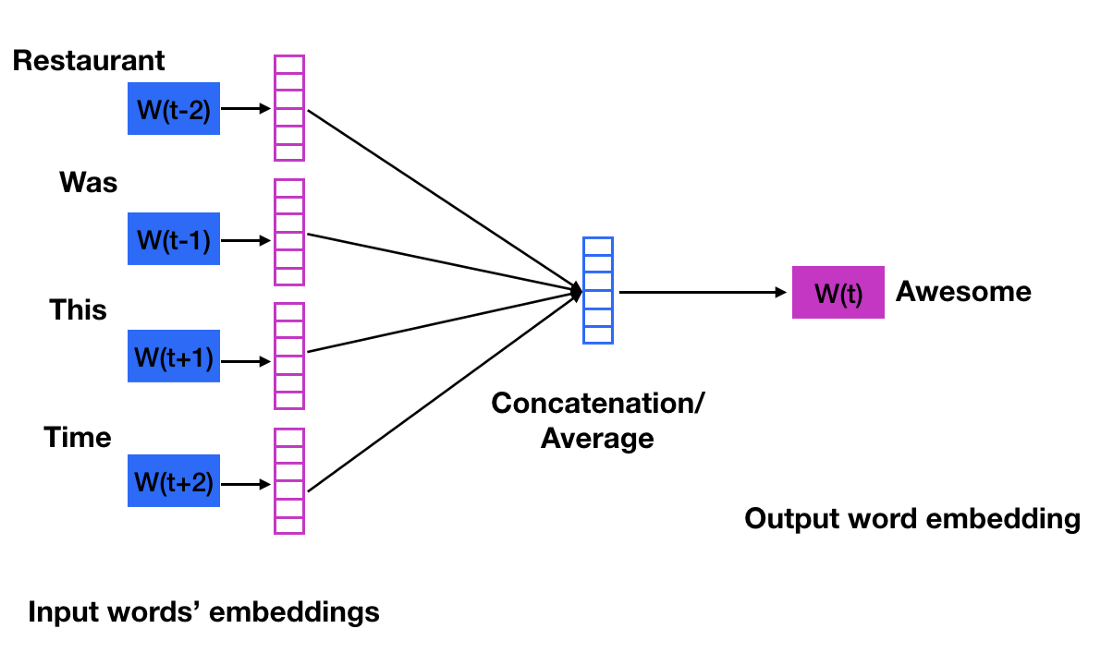
After the preprocess is done the dataset is divided into testing dataset and training dataset.

1. **Word2Vec Model**
2. **Word Embeddings**

Word embeddings are words that have been translated into real number vectors in a way that preserves their semantic meaning. To obtain the word techniques we use bag of words and term frequency inverse document frequency instead, they treat each word separately as a feature. This model considers the words around a given word inside a specific window size when training it. The term "embedding vectors" can be derived in a variety of ways. One such approach is Word2vec, which makes use of a neural embeddings model to learn that.

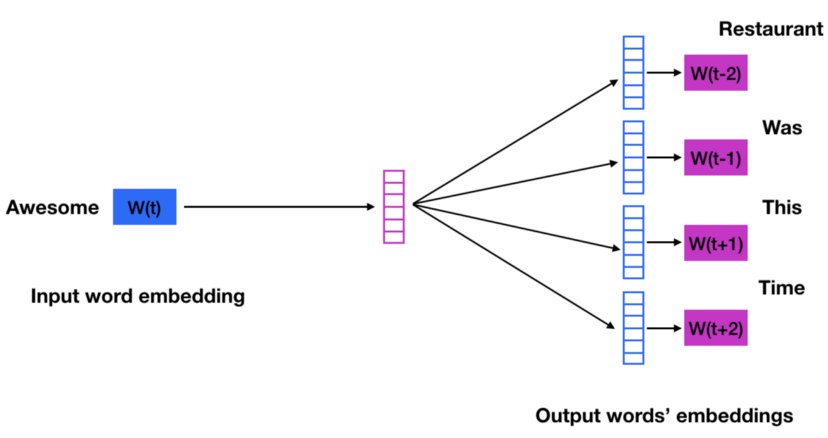
1. **Continuous Bag of Words**

In this the model gets the words under various parameters and constraints with the given content of words of a specific boundary or window. There are a number of dimensions in the hidden layer in which the output layer represents the current word. This makes us and gives us the content of the words and to proceed to the next process.



1. **Skip-Gram**

The skip-gram is direct from the Continuous Bag of Words. In this method we will predict embeddings of the window of the world bag created when we give a certain word to the model. The input layer will have the current word and the output for that in skip-gram will be list of similar words.



1. **Word2Vec Vectors**

By traversing through the dataset at each iteration the Word2Vec will be generated which will be later used. We can get the embedding layer through the embedding matrix from the training dataset by simply using the gensim model. We will be applying the mean values of words in the tweets which are in the form of sentences.

**2**. **Convolutional Neural Network**

Here after creating the embedding layer using the word2vec vectors it is added as one of the layers in the CNN then a Dropout layer will be added with a value of 0.5. The dropout layer is added which is set to either 1 or 0 so that the overall sum of the input is unchanged so that overfitting is prevented. Finally an LSTM layer is being introduced along with a dense layer. The dense layer used to reduce the dimensionality of the output. In the end of the CNN modeling along with the Word2Vec vectors word embeddings we get the polarity or the sentiment of the tweets.

Model: "sequential"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

embedding (Embedding) (None, 300, 300) 12659700

dropout (Dropout) (None, 300, 300) 0

lstm (LSTM) (None, 100) 160400

dense (Dense) (None, 1) 101

=================================================================

Total params: 12,820,201

Trainable params: 160,501

Non-trainable params: 12,659,700

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This classification and labeling of tweets makes it more polarized towards the specific domain and the respective world knowledge through our suggestion and the way of preparation of our data.

**Conclusion and Future Work**

In the present work we provided an implementation of sentiment analysis using distributed presentation of words and sentences and fit them into the Convolutional Neural Network through gensim Word2Vec for Sentiment Analysis based on tweets. There are many research done such as considering the Recursive Neural Tensor Network, Attention-Based Neural Networks, Multi-Task Learning Based Models, Unsupervised Sentiment Neuron and Non-Neural Networks Based Models these models are still in research and their strong points and advantage of using these particular models in specific situations are clearly defined.

We have applied a specific way to collect data from the twitter database so that it helps us overcome the domain specific limitations and implement one of the major concepts which is applying world knowledge to the Natural Language Processing techniques. We will also include the emojis and convert them to the related meaningful texts. As twitter is a social networking platform it will provide us with quality data and to help us build models to directly connect with the people or the audience.

With use of twitter as a Database we can explore various clusters of data and we can do proper collection of data and remove the other domain texts at the collection process itself. This process can be improved with having more advanced access to data in the twitter developer options. The speed of collection may help us build models of each domain separately and using hierarchical methods we can classify the polarities of tweets of specific domains

**References**

[1] CNN for situations understanding based on sentiment analysis of twitter data by Shiyang Liaoa, Junbo Wangb, Ruiyun Yua, Koichi Satob, Zixue Cheng at 8th International Conference on Advances in Information Technology, IAIT 2016, 19-22 (December 2016)

[2] A sentiment analysis study for twitter using the various model of convolutional neural network by Sartini, Subiyanto and M F Alim at Journal of Physics: Conference Series, ICMSE 2020 (2021)

[3] Twitter sentiment analysis using distributed word and sentence representation by Dwarampudi Mahidhar Reddy, Dr. N V Subba Reddy and Dr. Prema K V at arXiv :1904.12580v1 [cs.IR] (1 April 2019)

[4] Using Word2Vec to Process Big Text Data by Long Ma and Yanqing Zhang conference held at Big Data (Big Data), 2015 IEEE International ConferenceAt: San Jose, CA

(October 2015)

[5] Sentiment Analysis of Twitter Data: A Survey of Techniques by Vishal A.Kharde and S.S.Sonawane held at International Journal of Computer Applications 139(11): 5-15

(April 2016)

[6] Deep Learning applied to NLP by Marc Moreno Lopez and Jugal Kalita at arXiv:1703.03091 [cs.CL] (9 March 2017)