COMPARATIVE TEXT BASED EMOTION RECOGNITION USING ARTIFICIAL NEURAL NETWORK

Miss.M.Vinodhini
dept. Computer science
Mangayarkasi college of Engineering
Madurai, India

Miss.M.Sangavi
dept. Computer science
Mangayarkarasi College of Engineering
Madurai, India

Mr.J.Stalin
dept. Computer science(Asst.professor)
Mangayarkarasi College of Engineering
Madurai, India

Mr.V.Ramachandran
dept. Computer science(Research scholar)
Kalasalingam University of Academy and Education
Virudhunagar, India

***Abstract*—Sentiment analysis deals with identifying, classifying and grouping opinions or sentiments expressed in the source text. Social media is creating an immense measure of emotions with rich information as tweets, announcements, blog entries and so forth etc. Sentiment analysis of this user-generated information is exceptionally valuable in knowing the opinion of people. The Twitter analysis is troublesome compared to general emotions due to the presence of slang words and incorrect spellings. The maximum limit of characters that are p e r m i t t e d in Twitter is 280. Knowledgebase approach and Deep learning approach are the two strategies used for analyzing emotions from the text. The ANNs are trained using words that describe the dictionary definition of emotions.Messages are collected from blogs and social networks. In this paper, we search in twitter posts about electronic products like mobiles, laptops etc using ANN mechanism. By doing such analysis in a particular domain, it is conceivable to recognize the impact of domain information in sentiment classification the response of internet users are evaluated from multiple artificial neural networks for different emotions and we classify the tweets as positive, negative and extract peoples opinion**

***Index Terms*—Twitter, Sentiment Analysis, Deep Learning Techniques, Artificial Neural Network.**

I. Introduction

The period of the Internet has changed the manner in which individuals express their perspectives. It is presently done through blog entries, online gatherings, item survey sites and so forth. Individuals rely on this client created substance all things considered. When somebody needs to purchase an item, they will look into its audits online before taking a choice. The measure of client created content is unreasonably huge for a typical client to break down. So to automate this, various sentiment analysis techniques are used. Knowledge base approach and Machine learning techniques are the two main techniques used in sentiment analysis. Knowledge base approach needs a huge database of predefined emotions and an efficient knowledge representation for identifying emotions. Deep learning approach makes use of a training set to develop a sentiment classifier that classifies sentiments. Since a pre-defined database of entire emotions is not required for a deep learning approach, it is rather simpler than the Knowledge base approach.

In this paper, we use different deep learning techniques for classifying tweets Sentiment analysis is usually conducted at different levels varying from coarse level to fine level. Coarse level sentiment analysis deals with determining the sentiment of an entire Document and Fine level deals with attribute level sentiment analysis. Sentence level sentiment analysis comes in between these two . There are many types of research on the area of sentiment analysis of user reviews. Previous researches show that the performances of sentiment classifiers are dependent on topics. Because of that, we cannot say that one classifier is the best for all topics since one classifier doesnt consistently outperform the other. Sentiment Analysis in twitter is quite difficult due to its short length. Presence of emoticons, slang words, and misspellings in tweets forced to have a preprocessing step before feature extraction. There are different feature extraction methods for collecting relevant features from the text which can be applied to tweets also. But the feature extraction is to be done in two phases to extract relevant features. In the first phase, twitter specific features are extracted. Then these features are removed from the tweets to create normal text. After that, again feature

extraction is done to get more features. Since no standard dataset is available for twitter posts of electronic devices, we created a dataset by collecting tweets for a certain period of time. By doing sentiment analysis on a specific domain, it is possible to identify the influence of domain information. Different classifiers are used to do the classification to find out their influence in this particular domain with this particular feature using an artificial neural network

II. RELATED WORK

A. Related Text Classification Techniques

In [9], two strategies are discussed for using neural networks for text classification. In the traditional method, documents are encoded into numerical vectors, and in a novel method, string vectors are used. In the traditional method, documents are concatenated into a long string. The strings are tokenized by white space and punctuation. Then, each token is stemmed into its root form: verbs in past tense are converted to root form, and nouns in plural form are translated into singular form. Then, stop words are removed. The remaining words are called the feature candidates. The number of features are usually too large, so feature selection methods are used to select a subset of the features. Even after feature selection, the dimension of the numerical vectors can still be large. This leads to a high cost time for processing and a decrease in categorization performance. In the novel method of [9], documents are encoded into string vectors instead of numerical vectors. A d dimensional string vector would contain d words in descending order based on their frequencies in the text. This string vector is used as input into a novel neural network. Tokenization, stemming, and stop word removal are common preprocessing steps used in text classification [9],[10],[11]. When using various classification algorithms for text classification, there are commonly a large amount of features due to there being many words in documents. Tokenization is the process of extracting words from the text to be classified. Stemming is reducing words that are not in their root form to their root form. Stop word removal is the removal of common functional words in order to improve performance of the classification. Even after these preprocessing steps, the dimensionality of the features can still be large. Four feature selection techniques were used in [10]: the document frequency (DF) method, the category frequency-document frequency (CF-DF) method, the term occurrence frequency inverse document frequency (TFxIDF method), and the principal component analysis method. Through experiments, it was found that principal component analysis was most effective.Besides feed forward neural networks, other kinds of neural networks called the self-organizing map (SOM) and the growing hierarchical self organizing map (GHSOM) were used in [11]. The reason for using the SOM is that it works well for mapping high-dimensional data into a two- dimensional representation space.

B. Sentiment Analysis Techniques

One approach to sentiment analysis is a recurrent neural network [3]. It has the advantage over traditional neural networks in that it gets better performance on structured data prediction on variable input. In the recurrent neural network architecture, the input layer contains the bag of words feature vector at time t. The input layer is connected to the hidden layer which contains the history of information. The hidden layer has recursive connections to itself, and also connections to the output layer. The hidden and output layers also feature neurons to store values at a time t. The recursion allows the network to be deeper than a traditional neural network [3]. Rong et al. [3] proposed a semi-supervised dual recurrent neural network for their sentiment analysis. It is

similar to a traditional recurrent neural network in that it can use time to cover a longer history memory. It is different in that the output layer also has recursive connections back to itself for better sentimental analysis. In Yao et al. [12], a recursive neural network is used for language understanding. In language understanding, the main goal is to label words that have semantic meaning [12]. The network is trained using semantic labels rather than the words Them selves. The language is modeled by using the output as the input shifted in time so that the next word is predicted. In [4], sentiment analysis techniques are discussed specific to Twitter. Focus was placed on the electronic products domain. Twitter sentiment analysis is different from traditional sentiment analysis in that more slang and misspelled words are used due to the 140 character limit. Because of this, preprocessing is required before extracting features. Preprocessing is done in two steps. First, Twitter features such as hash tags and emoticons are extracted. Emoticons are given positive or negative scores based on being positive or negative. Hashtags are also given positive or negative scores. After Twitter specific features were removed,a unigram approach is used and the plain text of the tweet is represented as a collection of words. Positive and negative tweets are used in the feature vector. Other classification techniques for text classification that are discussed in [4] include Naive Bayes, a Support Vector Machine, and a Maximum Entropy classifier. The Naive Bayes classifier considers all features in the feature vector as independent of each other. The Support Vector Machine uses a hyper plane to separate tweets into classes. The Maximum Entropy Classifier does not assume relationships between features, and in the feature vector, the relationships between the part of speech tag, emotional keyword, and negation can be utilized in the classification process, unlike in the Naive Bayes Classifier. Anjaria et al. also did sentiment analysis on Twitter with a focus on trying to predict election outcomes [8]. It is mentioned that n- grams and Part-of-speech tagging techniques are used in Twitter sentiment analysis. As part of the prediction method, retweets that each party generates are factored in.An event driven neural network was used in [1] for evaluating internet user39;s response to given events. A Mood engine is built from multiple artificial neural networks, one for each mood, and different sets for different cultures. The ANNs are trained using words that describe the dictionary definition of moods. Messages are collected from blogs and social networks. Socher and Perelygin [7] created a recursive neural

tensor network that tries to better understand the semantics of a group of words. It was created to better understand short messages, commonly found on Twitter. Compared with standard recursive neural networks, matrix-vector recursive neural networks, Naive Bayes, and support vector machines, the recursive neural tensor network showed the most accuracy.

C. Deep learning techniques

Deep learning techniques use a training set and a test set for classification. The training set contains input feature vectors and their corresponding class labels. Using this training set, a classification model is developed which tries to classify the input feature vectors into corresponding class labels. These features can be used to find out the semantic orientation of words, phrases, sentences and that of documents. Semantic orientation is the polarity which may be either positive or negative. Domingos et al. [10] found that Naive Bayes works well for certain problems with highly dependent features. This is surprising as the basic assumption of Naive Bayes is that the features are independent. Zhen Niu et al. [11] introduced a new model in which efficient approaches are used for feature selection, weight computation, and classification. The new model is based on Bayesian algorithm. Here weights of the classifier are adjusted by making use of the representative feature and unique feature. Representative feature is the information that represents a class and Unique feature is the information that helps in distinguishing classes. Using those weights, they calculated the probability of each classification and thus improved the Bayesian algorithm.Barbosa et al. [12] designed a 2-step automatic sentiment analysis method for classifying tweets. They used a noisy raining set to reduce the labeling effort in developing classifiers. Firstly, they classified tweets into subjective and objective

tweets. After that, subjective tweets are classified as positive and negative tweets. Celikyilmaz et al. [13] developed a pronunciation based word clustering method for normalizing noisy tweets. In pronunciation based word clustering, words having similar pronunciation are clustered and assigned common tokens. They also used text processing techniques like assigning similar tokens for numbers, HTML links, user identifiers, and target organization names for normalization. After doing normalization, they used probabilistic models to identify polarity lexicons. They performed classification using the BoosTexter classifier with these polarity lexicons as features and obtained a reduced error rate. Wu et al. [14] proposed an influence probability model for twitter sentiment analysis. If @username is found in the body of a tweet, it is influencing action and it contributes to influencing probability. Any tweet that begins with @username is a retweet that represents an influenced action and it contributes to influence probability. They observed that there is a strong correlation between these probabilities. Pak et al. [15] created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they built a sentiment classifier based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features. In that method, there is a chance of error since emotions of tweets in the training set are labeled solely based on the polarity of emoticons. The training set is also less efficient since it contains only tweets having emoticons. Xia et al. [16] used an ensemble framework for sentiment classification. Ensemble framework is obtained by combining various feature sets and classification techniques. In that work, they used two types of feature sets and three base classifiers to form the ensemble framework. Two types of feature sets are created using Part-of-speech information and Word-relations. Naive Bayes, Maximum Entropy and Support Vector Machines are selected as base classifiers. They applied different ensemble methods like the fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy. Certain attempts are made by some researches to identify the public opinion about movies, news etc from the twitter posts. V.M. Kiran et al. [17] utilized the information from other publicly available databases like IMDB and Blippr after proper modifications to aid twitter sentiment analysis in the movie domain.

III. PROPOSED SOLUTION

A dataset is created using twitter posts of electronic products. Tweets are short messages full of slang words and misspellings. So we perform a sentence level sentiment analysis. This is done in three phases. In the first phase preprossessing is done. Then Artificial Neural Network model is created using deep learning methodology. Finally using different classifiers, tweets are classified into positive and negative classes. Based on the number of tweets in each class, the final sentiment is derived.

A. Pre-processing of the data

- Step 1: Convert the data to lower case:
  the fraternity of doctors is special. during mannkibaat, i conveyed greetings for doctor's day. doctors from india are admired world over for their skills and excellence.
- Step 2: Word tokenization:
  In this step, we will split the entire sentence into individual words. We will get this ['the', 'fraternity', 'of', 'doctors', 'is', 'special', '.', 'during', 'mannkibaat', ',', 'i', 'conveyed', 'greetings', 'for', "doctor's", 'day', '.', 'doctors', 'from', 'India', 'are', 'admired', 'world', 'over', 'for', 'their', 'skills', 'and', 'Excellence', '.']
- Step 3: Removal of hash tags and usernames: In a tweet you may find the presence of hash tags which starts with a symbol and usernames which starts with @ symbol. Generally, we do not need this since it specifies a username and hash tag. So we will get this as a result. ['the', 'fraternity', 'of', 'doctors', 'is', 'special', '.', 'during', ',','conveyed', 'greetings', 'for', "doctors, day",'.', 'doctors', 'from', 'India', 'are', 'admired','world', 'over', 'for', 'their', 'skills', 'and', 'excellence', '.']
- Step 4: Removal of stop-words (a, was, the, etc):
  Stop words are generally meaningless pieces of data which is needed to be removed in order to mine the vital

information from the data. Here is the result after stop words are removed. You will not find is, was, the, a, etc., in the resulting words.
['fraternity', 'doctors', 'special', '.', ',', 'conveyed', 'greetings',"doctor's", 'day', '.', 'doctors', 'India' 'admired', 'world', 'skills', 'excellence', '.']

- Step 5: Remove anything which is not alphabetical:
After removal of non-alphabetical data, you will get this result. ['fraternity', 'doctors', 'special', 'conveyed', 'greetings', 'doctors day', 'doctors', 'India', 'admired', 'world, skills', 'excellence']
- Step 6: Perform stemming and Join into a sentence:
Stemming is the process of converting the words into its root form for example, The word, loving is converted into love and the word calling will be converted into the call. Important: The resulting root word needs not to be in proper English. This process is also called as lemmatization. This will be our final result. fraternal doctor special convey greet doctor day doctor india admire world skill excel Finally, we have mined the important information form the data.

B. Stemming and Lemmatization

The aim of both stemming as well as lemmatization is to scale down inflectional types mostly derivationally associated varieties of a phrase to a fashioned base kind. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of accomplishing this goal accurately more often than not, and quite often involves the removal of derivational affixes. Lemmatization often refers to doing matters competently with the usage of vocabulary and morphological analysis of phrases, in most cases aiming to eliminate inflectional endings only and to come back the base or dictionary type of a word, which is often called the lemma.

C. Artificial Neural Network (ANN)

An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it.

In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some nonlinear function of the sum of its inputs. The connections between artificial neurons are called 'edges'. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times. Artificial Neural Network (ANN) or known as the neural network is a mathematical technique that interconnects a group of artificial neurons. It will process information using the connections approach to computation. ANN is used in finding the relationship between input and output or to find patterns in data.

D. Natural Language Processing (NLP)

NLP techniques are based on deep learning and especially statistical learning which uses a general learning algorithm combined with a large sample, a corpus, of data to learn the rules Sentiment analysis has been handled as a Natural Language Processing denoted NLP, at many levels of granularity. Starting from being a document level classification task it has been handled at the sentence level and more recently at the phrase level, NLP is a field in computer science which involves making computers derive meaning from human language and input as a way of interacting with the real world.

E. Application Programming Interface (API)

Alchemy API performs better than the others in terms of the quality and the quantity of the extracted entities As time passed the Python Twitter Application Programming Interface (API) is created by collected tweets. Python can automatically calculate the frequency of messages being retreated every 100 seconds, sorted the top 200 messages based on there-tweeting frequency, and stored them in the designated database [12]. As the Python Twitter API only included Twitter messages for the most recent six days, collected the data needed to be stored in a different database.

F. Python

Python was found by Guido Van Rossum in Netherlands, 1989 which has been public in 1991Python is a programming language that's available and solves a computer problem which is providing a simple way to write out a solution]. Mentioned that Python can be called as a scripting language. Moreover, and also supported that actually, Python is a just description of language because it can be one written and run on many platforms. In addition, mentioned that Python is a language that is great for writing a prototype because Python is less time consuming and working prototype provided, contrast with other programming languages. Many researchers have been saying that Python is efficient, especially for a complex project, as has mentioned that Python is suitable to start up social networks or media steaming projects which most always are a web-based which are driving a big data. gave the reason that because Python can handle and manage the memory used. Besides Python creates a generator that allows an iterative process of things, one item at a time and allows the program to grab source data one item at a time to pass each through the full processing chain.

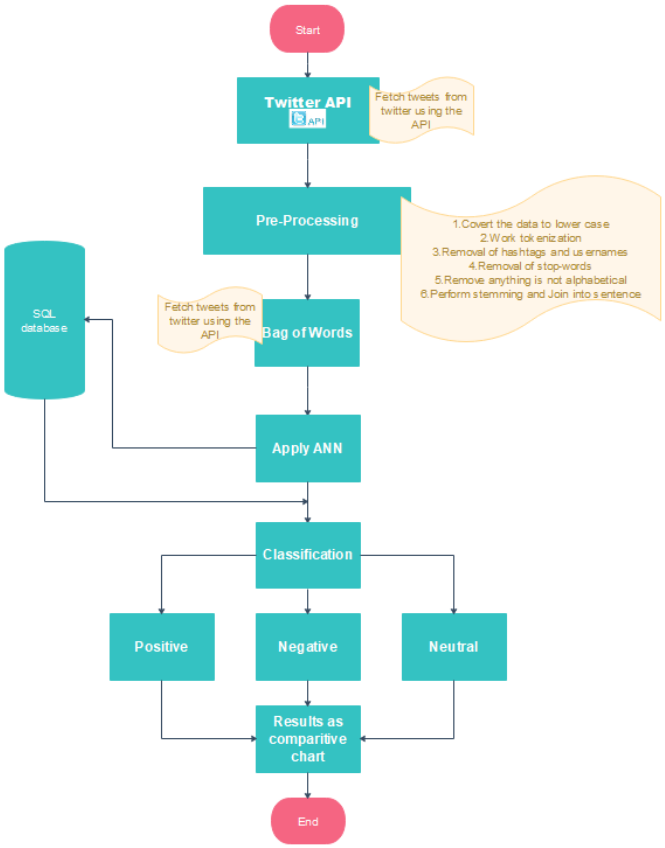

Fig. 1. Architecture diagram.

IV. SYSTEM DIAGRAM

In this block diagram the tweets are fetched initially, from twitter using the Application programming interface and pre-processing is done. However, the pre-processed data have some features like covering the data into lower case.

Word tokenization,Removal of hashtags and username, Removal of stop-words,Remove anything which is not alphabetical,Word tokenization will split the entire sentence into individual words and the hastags and usernamesxare removed as we don't need it to get the specified result.The stemming process helps to convert the words into root form and it is not necessary for the root form to be in proper english and this process is called lemmatization. Perform stemming and Join into a sentence,then the bag of words is applied into the pre-processed data and it is stored in the SQL database.Finally, the classification is done by using Artificial neural network model.(Fig.1)

V. RESULT AND DISCUSSION

A. Twitter Retrieved

To associate with Twitter API, the developer needs to agree in terms and conditions of development Twitter platform which has been provided to get an authorization to access a data. The output from this process will be saved in JSON file. The reason is, JSON JavaScript Object Notation) is a lightweight data-interchange format which is easy for humans to write and read. Moreover, stated that JSON is simple for machines to generate and parse. JSON is a text format that is totally language independent but uses a convention that is known to programmers of the C-family of languages, including Python and many others. However, output size depends on the time for retrieving tweets from Twitter. Nevertheless, the output will be categorized into 2 forms, which are encoded and un-encoded. According to the security issue for accessing a data, some of the output will be shown in an ID form such as string ID. Sentiment Analysis. The tweets will be assigned the value of each word, together with categorize into positive and negative word, according to the lexicon dictionary. The result will be in comparative chat.

B. Sentiment Analysis

Tweets from JSON file will be assigned the value of each word by matching with the lexicon dictionary. As a limitation of words in the lexicon dictionary which is not able to assign a value to a very single word from tweets. However, as a scientific language of python, which is able to analyze a sense of each tweet into positive or negative forgetting a result.

C. Information Presented

The result will be shown in a pie chart which is representing a percentage of positive, negative and null sentiment hashtags. For null hashtag is representing the hashtags that were assigned zero value. However, this program is able to list a top ten positive and negative hashtags. The pie chart is representing each percentage positive, negative and null sentiment hashtags in different color.

VI. CONCLUSION

Applying Sentiment analysis to mine a large amount of unstructured data has become an important research problem. Now business organizations and individuals are putting forward their efforts to find the best system for sentiment analysis. Some of the algorithms have been used in sentiment analysis to gives good results, but no technique can resolve all the challenges. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms, but it also has limitations. To overcome the limitation of some techniques, our study focus is on the Deep learning approaches and use of artificial neural networks (ANN) in sentiment classification and analysis. Our study suggests that the ANN implementations would result in improved classification, combining the best of an artificial neural network with fuzzy logic

REFERENCES

[1] Mejova, Y. (2009). Sentiment analysis: An overview. Comprehensive exam paper. Computer Science Department, 1-34.
[2] Boiy, E., Hens, P., Deschacht, K., Moens, M. F. (2007, June). Automatic Sentiment Analysis in On- line Text. In ELPUB (pp. 349-360).

[3] Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417- 424). Association for Computational Linguistics.
[4] https://nlp.stanford.edu/software/stanford-postagger-2013-04-04.zip
[5] http://www.cs.grinnell.edu/
[6] Zhou, Y., Fan, Y. (2013). A sociolinguistic study of American slang. Theory and Practice in Language Studies, 3(12), 2209-2214.
[7] Comesaa, M., Soares, A. P., Perea, M., Pieiro, A. P., Fraga, I., Pinheiro, A. (2013). ERP correlates of masked affective priming with emoticons. Computers in Human Behavior, 29(3), 588-595.
[8] Huang, A. H., Yen, D. C., Zhang, X. (2008). Exploring the potential effects of emoticons. Information Management, 45(7), 466-473.
[9] Boyd, D., Golder, S., Lotan, G. (2010, January). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In 2010 43rd Hawaii International Conference on System Sciences (pp. 1-10). IEEE.
[10] Carpenter, T., Way, T. (2012, January). Tracking Sentiment Analysis through Twitter. In Proceedings of the International Conference on Information and Knowledge Engineering (IKE) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
[11] Osimo, D., Mureddu, F. (2012). Research challenge on opinion mining and sentiment analysis. Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Btiment, 508.
[12] Pak, A., Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).
[13] Lohmann, S., Burch, M., Schmauder, H., Weiskopf, D. (2012, May). Visual analysis of microblog content using time-varying co- occurrence highlighting in tag clouds. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 753-756). ACM.
[14] Sarlan, A., Nadam, C., Basri, S. (2014, November). Twitter sentiment analysis. In Proceedings of the 6th International Conference on Information Technology and Multimedia (pp. 212- 216). IEEE.
[15] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38).
[16] Huang E.H, Socher R, Manning C.D. and Ng A.Y. Improving word representations via global context and multiple word prototypes. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2012), 2012.
[17] Pennington J, Socher R, Manning C.D. GloVe: global vectors for word representation. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP 2014), 2014.