Twitter Sentimental Analysis Using Neural Network

Avudaiappan.T, Jenifer, Sisay tumsa[3], Subashrree, T.Jayasankar

Abstract: Sentimental Analysis can be referred to the process of analyzing and determining the thought process of the writer based on their messages on various social networking sites. Twitter is one of the most renowned social networking websites where user can read and post messages about a person, an event, a product and the current happenings all over the world. These are normally 140-280 characters in length. In this system, tweets are used as the raw data. The tweets are collected through Twitter API using a secret token. Then they are preprocessed using text mining package to reduce the noise in the words. The score is computed for each pre-processed tweet using Dictionary-Based Approach. For positive tweets, the score is 1, for negative tweet the score is -1 and 0 for neutral tweet. The pre-processed tweets along with the scores are stored in CSV format for further process. The train data and test data is provided in the ratio 60:40 to construct the classification model. After classification, it is observed that - Convolutional Neural Network is unformulated to compute the probability of the tweets. The system uses K-fold cross validation method to improve over the holdout method. Finally, as the result the opinion of the sentiment related to the given tweets is predicted using probability of the positive tweets by hybrid approach. This system produces a better performance measure when compared to other method

Index Terms: Data Mining, Sentimental Analysis, Deep Learning, Dictionary Approach, CNN

1. INTRODUCTION

1.1 Sentimental Analysis

Sentiment Analysis can be referred to as the computational investigation of individuals' suppositions, attitudes and feelings towards an element that utilize data mining procedures and methods to extricate information for investigation to perceive the subjective sentiment of a record, like blog entries, audits, news articles and online networking posts like tweets and notices. The element may be people, occasions or subjects. Sentiment Analysis is not only applied on item surveys but also can be connected on stock markets [12], news articles [20], or political debates [9]. It can be considered as a upgraded field of research in text mining. There are three fundamental order levels in Sentiment Analysis such as document level, sentence level, and aspect level. This paper focus on Document level Sentiment Analysis, it's easily distinguishing the document into positive or negative sense. It considers the entire document as a fundamental data unit. Sentence-level Sentiment Analysis expects to make conclusions in each sentence. The first and foremost step involved would be to find if the current post or sentence is internal fact or external fact. It recognizes the sentiment communicated through a content at that point and then examines it. Aspect-level Sentiment Analysis aims at classifying specific aspects of entities and the sentiments involved in it. In this manner, the sentiment analysis technique automatically detects the polarity of a sentence or a post.

1.2Dictionary-Based Approach

Dictionary based approach is commonly used one because dictionaries having the equivalent and opposite of each word. A simplest technique involved in dictionary based method is to use a many seed emotional words to build a dictionary structure using equivalent and opposite words. Specifically, this method works as follows: A small bunch of sensitive words with known equivalent and opposite words are collected manually, and then algorithms uses this group of sensitive words by searching in the online dictionary for their equivalent and opposite words. The lately identified words are added to the seed list. This processes repeated for multiple iterations. The iterative processes finally stop when further new words can be found. After the process is completed, a manual inspection step is used to clean up the list.

1.3 Deep Learning

Off late, Deep learning has become one of the most powerful learning techniques that learn quickly about features of the data and produces state-of-the-art prediction results [17].The main concept of deep leaning algorithms are automatically extract the information from the data. Deep learning algorithms are using a huge amount of unstructured data to automatically extract information from the complex unstructured data. These deep learning algorithms are largely inspired by the Nature, which has the common goal to mimic the human brain's activities. The human brain is an amazing processor. Its exact workings are still a mystery. The most basic element of the human brain is neuron. Deep learning algorithms constitute to abstract representations as many abstract representations are built accumulating less abstract ones. A significant merit of more abstract representations is that they can be invariant to the local changes in the input data. Learning such invariant features would be the major ongoing goal in pattern recognition. Deep learning algorithms are actually Deep architectures consisting of many consecutive layers. In this case, Non-linear transformation is applied on the input and the representation of the output is provided.

2. LITERATURE SURVEY

2.1 SENTIMENT ANALYSIS ON SOCIAL MEDIA

The web acts as a big platform where the users can share their opinions. Their reviews are very significant for giving credits over good products and efficient plans. This paper

- [1]*Assistant Professor, K. Ramakrishnan College of Technology, Trichy, Tamilnadu, India. E-mail: avudaiappanmecse@gmail.com*
- [2,3]*Lecturer, Faculty of Computing and Software Engineering, Arbaminch Institute of Technology, Arbaminch University, Arbaminch, Ethiopia*
- *. E-mail: [2]jenifer.mahilraj@amu.edu.et, [3]tumsa@amu.edu.et*
- [4]*UG Student K. Ramakrishnan College of Technology, Trichy, Tamilnadu, India. E-mail: subashrree.satya@gmail.com*
- (This information is optional; change it according to your need.)

proposes a system that uses Facebook as a social media site to gather the opinion of the public through Facebook posts [8]. 1000 Facebook posts of La7 and Rail news programs of Italy are collected. This used a knowledge mining system used by Italy security agencies for government. K-Means clustering algorithm is applied for the clustering of news posts of La7 and Rail. To enhance the results, semantic and linguistic approaches are included. Each sentence is analyzed to assign the context with the appropriate meaning to provide accurate result. Then the posts are preprocessed by removing different Parts of Speech i.e. noun, verb, adverb, adjective to decrease the complexity. Sentimental analysis is performed through polarity mining and syntactic tree. The Bayesian learning method is used for classification process. Recall and precision is adopted as the evaluation measures for computing the performance of the proposed system.

2.2 TWITTER DATA ANALYSIS USING SUPERVISED CLASSIFIERS& COMPARISON

Twitter achieves greatest measures of consideration on things related to product, movie reviews, stock exchange, government policies etc. There can be any kind of person who use Twitter like - a user can be a student, film actors, celebrities, sportsman and many leaders across the whole world containing data of different languages [10]. The proposed system performs sentiment prediction on movie reviews using machine learning algorithms. Supervised machine learning algorithms like SVM, Max Ent and NB is used here. The unigram, bigram and hybrid (which is a combination of both unigram + bigram features) are used for classifying data. In this system - 15000 movie reviews, 5000 tweets for training set and 2000 tweets as test data set. The stop words, repeated words and links are preprocessed to avoid unwanted processes. Finally, the tweets are categorized into positive, negative and neutral reviews. Accuracy is used as the performance measure in this approach.SVM using hybrid feature outperforms all other classifiers and selection feature with accuracy of 84% for movie reviews.

2.3 SENTIMENT ANALYSIS BASED ON DICTIONARY APPROACH

Sentiment analysis has become emerging field in recent years, and sentiment dictionaries have become very essential for research in this field [3]. For any online purchase, Consumers take decisions based on the reviews provided by other consumers. Any review – either good or bad, gives an impression about the product to the consumers and help them decide to make their purchase and learn respectively. So, these kind of review – tweets are pre-processed into a more structured information. Tokenization is used for identification of words in each text. Pre-processing of text is followed by analysis of the tweets. This analysis step is considered as the core of text mining, because this is where some type of useful, nontrivial knowledge is extracted from the text. The performance of sentiment classification is evaluated accurately in this SVM (hybrid method).

2.4 A HOLISTIC ANALYSIS OF TWITTER SENTIMENTS USING DEEP LEARNING METHOD

Deep learning is part of neural network that consists of many hidden layers and reflects the process of neurons in human brain to provide accurate result. This helps in increasing the accuracy of sentiments of the tweets [1].This proposed system uses tweets from Twitter for sentimental analysis. The data set is downloaded from Twitter API. To clean the words preprocessing steps are performed which includes stemming, removal of numbers, punctuations, stop words, white spaces and converting all the tweets into same case letters. Then the sentence is tokenized. The data set consists of 2000 training data set and 2000 testing data set in both Korean and English languages. Then FNN is applied on the data set. This system also uses Tensor Flow platform for creating neural network. Accuracy is used as the evaluation measure. This inputs 100 neurons for processing. This neural network uses three hidden layers for sentiment prediction. Multilayer perceptron produces accuracy of 52.60% as a result.

3. SYSTEM ANALYSIS

3.1 Existing System

Information about a data has become one of the most important metadata to extract useful and heterogeneous knowledge from the unstructured data. We can get the knowledge from text; it's provided information such as explanation, direction, opinion and also includes reviews, emotions, or feelings. Internal Fact information can be expressed through different textual genres, like blogs, forums, and reviews, through social networks and micro blogs. These sites having a large amount of information, due to millions of users sharing opinions on different aspects of their everyday life. Extracting this kind of subjective information has a great value for both general and expert users.

3.1.1 Dictionary Based Approach

Sentiment analysis involves categorizing the viewpoints involved in texts within certain categories like positive, negative or neutral [8]. The input which is fed into the sentiment analysis system is the corpus of documents. The chief component or part of the system is the document analysis module, which utilizes linguistic tools to cover the pre-processed records with sentimental annotations. Observations are attached to the whole document, to individual sentences or to specific features of entities. These observations are the outcome of the system and they maybe represented to the users utilizing different varieties of conception tools.

Disadvantages:

- Less accuracy rates

3.2 Proposed System

Sentiment analysis is a popular topic and is largely applied to data that comes with self-labeled information reviews from Twitter API. To access Twitter API, an app is developed to join as developer member of Twitter. After creating the app, Twitter provides4 types of tokens to provide authentication to retrieve tweets. The tweets collected are pre-processed. The pre-processing steps includes removal of links, numbers, punctuations, retweets, stop words, prepositions, references to other screen names, hash tags, tabs, new line, spaces at the beginning and end, emoticons and convert the tweet to lower case. The ability to identify the positive or negative sentiment which lies behind a piece of text is much more interesting when it comes to social media data. Each review is computed score using dictionary-based approach. A collection of positive and negative word list is read, and it is compared with the tweets to find score of the tweet. If the tweet is positive then the score is 1, for negative tweet it is -1 and 0 for neutral review. After that, the tweets are stored along with their score in a CSV file. Convolutional Neural Network is made up of neurons that have learnable weights and biases. The training set of data and test set of data are then provided. Document-term matrix and vocabulary is constructed for model creation. Then sentiment is analyzed using deep learning process. Finally, a probability graph depicting the probability of positive tweets is presented.

Advantages:

- Improved accuracy rate.
- Higher precision.
- Higher recall.

4. SYSTEM IMPLEMENTATION

4.1 Modules Description

- Data Collection
- Tweet Preprocessing
- Polarity Mining using Dictionary-Based Approach
- Sentiment Analysis using Hybrid Approach
- Performance Evaluation

4.1.1 Data Collection

There is a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc from Social media. For analyzing the tweets, an app is created to retrieve tweets from Twitter API. After authentication by the user the tweets are retrieved. To perform polarity mining using dictionary-based approach positive words and negative words are stored in text document format. Once the tweets are pre-processed the data is stored as tweetfile.CSV.

4.1.2 Tweet Preprocessing

Data cleaning or pre-processing is the process of finding and cleansing the data by replacing, modifying, or deleting the dirty or coarse data [18]. Once the tweets are retrieved, the cleaning process is done.

The basic preprocessing steps involved are:

- Removal of Digits
- Removal of Expressions
- Removal of links
- Removal of symbols
- Removal of Stop-words
- Removal of Prepositions
- Removal of Punctuations
- Removal of retweet
- Clean hash tags
- Remove tab
- Removal of reference to other screen names
- Convert tweets to lower case

4.1.3 Polarity Mining using Dictionary-Based Approach

After the pre-processing is done, the positive and negative words are read and saved for comparison [8]. The pre-processed tweets are compared with positive and negative words and then classified into either positive or negative or neutral. The score for positive or negative or neutral tweet are 1 or -1 or 0 respectively [11]. Then the tweets are stored with scores in CSV file. The graph is plotted to represent the sentimental analysis using Dictionary-Based Approach.

4.1.4 Sentiment Analysis using Hybrid Approach

Once polarity mining is done using Dictionary-Based Approach, Sentimental Analysis is performed using CNN. The tweets stored in the CSV file are retrieved [16]. Then the train and test data is provided in the ratio 60:40. The vocabulary and document-term matrix is formed. The CNN model is trained, using the training and test dataset. This uses K-fold cross validation to estimate how accurately a predictive model can perform practically. In a forecast problem, a model is generally give a dataset of well known data on which training is run, and a dataset of unfamiliar data against which the model is tested. The goal of cross validation is to define a dataset to "test" the model in the training phase, in order to limit problems like over fitting, give an insight on how the model will generalize to an independent dataset. Here k is assumed to be 5. The positive probability of each tweet is determined and stored as matrix. Then the total positive, negative and neutral tweets are known. Finally, the probability graph is plotted. This graph shows the sentiment probability from 0 to 1 in y-scale and created time in x-scale. The probability of each tweet is plotted on the graph. Construction of CNN modelIn recent times, deep learning is able to solve many real time troubles like computer vision, video recognition, and CNN also helped in the field of image analysis and image classification. The major reason to take CNN in image analysis and classification is that CNN can take out group of features from image and is able to connect the relationship among these features. In NLP, texts data features can be retrieved as matrix formation and CNN also make the relationship among these features. Currently, CNN has a convolutional layer to extract information from large volume of text, so we work for sentiment analysis with CNN, and we design a simple CNN model and test it on benchmark.

CNN has 3 layers namely,

- Convolutional Layer,
- Pooling Layer
- Fully-Connected Layer

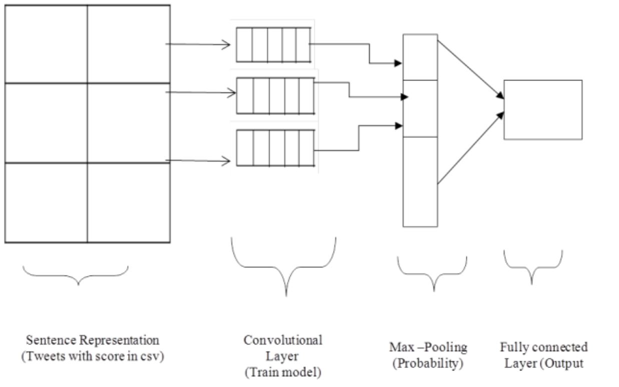

Fig: 4*.1.4 CNN Construction*

4.1.4 Performance Evaluation

Below metrics can be used to measure performance namely –

- Precision
- Recall
- Accuracy

4.1.4.4 Precision

Precision measures the accuracy of a classifier [19]. If there are very few false positivies, then the precision is considered to be higher, on the corollary, if there are more false positive then it is considered as lower precision
Precision=TP/ (TP + FP) --1

4.1.4.5 Recall

Recall measures the completeness or sensitivity of a classifier. It is also known as Sensitivity [14]. Higher recall refers to very few false negatives, while lower recall means more false negatives.
Recall=TP/ (TP + FN) ---2 2

4.1.4.6 Accuracy

A measure of how often a sentiment rating was correct. It tracks how many of the tweets were rated correctly.
Accuracy = (TP+TN) / (TP+TN+FP+FN) --------------------------------3

5. EXPERIMENTAL RESULT

We choose the tweets from Twitter API as dataset for sentimental analysis.

DATASET	Twitter API
NO. OF TWEETS	2500
LANGUAGE	English

***Table 5.1:** Detail of Dataset*

The tweets are pre-processed, and score is calculated using Dictionary-based approach.

NO.OF NEGATVIE WORDS	2195
NO. OF POSTIVE WORDS	4096
LANGUAGE	English

***Table 5.2**: Train data for Dictionary-based Approach*

The training data set along with their scores is split into 2 groups namely - train set which acts as the input as training samples and the development set which is used to verify the accuracy of the checkpoint of the CNN. For each of the dataset, we set the train set around 60% of the whole data amount and the development set is around 40%. CNN model is used for training purposes.

NO. OF DATASET	NO. OF TRAIN DATASET	NO. OF TEST DATASET
2500	1500	1000

***Table 5.3**: Detail of Train and Test dataset*

Confusion matrix for Dictionary-based approach and CNN is calculated and with the help of confusion matrix precision, recall and accuracy are calculated.

N=2500	Reference	
Prediction	0	1
0	TN	FP
1	FN	TP

***Table 5.4:** Confusion Matrix for Dictionary-Based Approach*

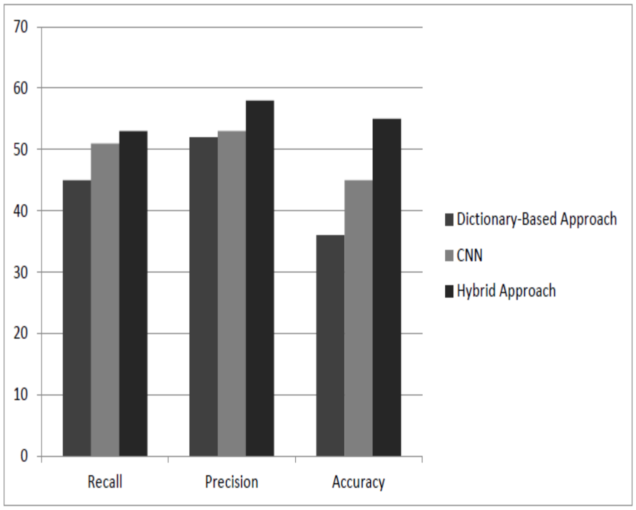

***Fig.: 5.1:** Comparison between Dictionary-Based Approach, CNN and Hybrid Approach*

Algorithm	Recall	Precision	Accuracy
Dictionary-based approach	45%	52%	36%
CNN	51%	53%	45%
Hybrid Approach	53%	58%	55%

***Table 5.5:** Comparison of accuracy*

Compared to the existing system i.e., Dictionary –Based Approach and CNN the hybrid approach produces high rates in terms of precision, recall and accuracy.

6. Conclusion and Future Enhancement

6.1 Conclusion

Twitter sentimental analysis could be a little tedious process because it is very hard to recognize emotional words from tweets. The sentiment could vary between people. These kind of problems can be handled by feature vector. Before extracting feature, text pre-processing is done on every tweet. Then CNN model is constructed for the CSV file which was created along with the scores for the tweets. Finally, the probability of the positive tweets is calculated, and the performance is evaluated. The above said method was classifying tweets in positive, negative and neutral sentiments. The proposed model has gone through preprocessing stage, features generation stage and classifiers learning stage. The statically evaluation of the above proposed model is done in terms of precision, recall and accuracy. The results of the CNN method are then observed by comparing against the Dictionary-based approach. The comparative results show that the proposed model has improved many factors like - precision, recall and accuracy of tweets' class prediction of sentiments.

6.2 Future Enhancements

A method to predict the location and user' s information of certain tweets though anonymous can be identified in the future. It is proposed to stream real time live tweets from twitter using Twitter API, and to perform sentimental analysis on multiple other languages. This will help to enhance any analysis on any social media especially Twitter in this case.

REFERENCES

[1] Adyan Marendra Ramadhani and Hong Soon Goo,"Twitter Sentiment Analysis using Deep Learning Methods." Engineering Seminar (InAES), 2017 7th International Annual.

[2] Ali Selamat and NurulhudaZainuddin ,"Sentiment Analysis Using Support Vector Machine" 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology, 2014 .

[3] ApekshaPande , Reshma Bhonde , Binita Bhagwat ,et.al, "Sentiment Analysis Based on Dictionary Approach", International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 51-55.

[4] Bhumika , Jadav and VimalkumarVaghela, "Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis." International Journal of Computer Applications Volume 146 – No.13, July 2016.

[5] Bowen Baker, Otkrist Gupta, Nikhil Naik, et.al,"Designing Neural Network Architectures using Reinforcement learning" Published as a conference paper at ICLR 2017.

[6] Carlo Aliprandi, Federico Neri, Federico Capeci ,et.al, "Sentiment Analysis on Social Media." IEEE Computer Society Washington, 2012

[7] Daniel Jurafsky and James H. Martin "Naive Bayes and Sentiment Classification" Published on August 7,2017.

[8] Fatehjeet Kaur Chopra and Rekha Bhatia, "Sentiment Analyzing by Dictionary based Approach" International Journal of Computer Applications(0975-8887).

[9] Isa Maks and Piek Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications." Elsevier Science Publishers B. V. Amsterdam,2012-11-01.

[10] Joshi, Rohit, and Rajkumar Tekchandani, "Comparative analysis of Twitter data using supervised classifiers." Inventive Computation Technologies International Conference on. Vol. 3. IEEE, 2016.

[11] Adyan Marendra Ramadhani and Hong Soon Goo,"Twitter Sentiment Analysis using Deep Learning Methods." Engineering Seminar (InAES), 2017 7th International Annual.

[12] Ali Selamat and NurulhudaZainuddin ,"Sentiment Analysis Using Support Vector Machine" 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology, 2014 .

[13] ApekshaPande , Reshma Bhonde , Binita Bhagwat ,et.al, "Sentiment Analysis Based on Dictionary Approach", International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 51-55.

[14] Bhumika , Jadav and VimalkumarVaghela, "Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis."International Journal of Computer Applications Volume 146 – No.13, July 2016.

[15] Bowen Baker, Otkrist Gupta, Nikhil Naik, et.al,"Designing Neural Network Architectures using Reinforcement learning" Published as a conference paper at ICLR 2017.

[16] Carlo Aliprandi, Federico Neri, Federico Capeci ,et.al, "Sentiment Analysis on Social Media." IEEE Computer Society Washington, 2012

[17] Daniel Jurafsky and James H. Martin "Naive Bayes and Sentiment Classification" Published on August 7,2017.

[18] Fatehjeet Kaur Chopra and Rekha Bhatia, "Sentiment Analyzing by Dictionary based Approach" International Journal of Computer Applications(0975-8887).

[19] Isa Maks and Piek Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications." Elsevier Science Publishers B. V. Amsterdam,2012-11-01.

[20] Joshi, Rohit, and Rajkumar Tekchandani, "Comparative analysis of Twitter data using supervised classifiers." Inventive Computation Technologies International Conference on. Vol. 3. IEEE, 2016.