ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

Aspect Based Evaluation of Social Anxiety using Natural Language Processing Techniques

K.Jayanth [1], Dr.M.Sampath Kumar [2]

P.G. Scholar, Department of CS & SE, College of Engineering (A), Andhra University, Visakhapatnam, India[1]

Professor Department of CS & SE, College of Engineering (A), Andhra University, Visakhapatnam, India[2]

ABSTRACT: This paper proposes aspect based evaluation of social anxiety using Natural Language Processing (NLP) Techniques. With significant advancement in social media and web, the content in the social media websites and in the Internet has increased dramatically. Knowing these opinions play a vital role in decision making but because of its huge amount of data it is difficult to evaluate manually. Thus, Natural Language Processing Techniques helps to summarize the data and know the anxiety or opinion of the people. Aspect Based evaluation technique is used to know the opinion of people. Most opinion classification uses the Bag of words approach while leaving the main features of the sentence. This may mislead the classification algorithm especially when used for problems like opinion prediction. In Aspect Based evaluation the meaning of sentence is preserved by taking the aspects of the sentence and classifying them by applying machine learning algorithms on it, to know the opinion.

KEYWORDS: Aspect Based evaluation, Natural Language Processing, Sentiment.

I. INTRODUCTION

Now a day's Online services play a vital role in every individual day to day life. These services include daily News, weather, banking transactions, blogging, social media and much more. With growing web technologies, buying online goods and selling through online has increased drastically. With the ability to share their feelings in the form of opinion and reviews, knowing these opinions and its sentiment is important since it plays an important role in the decision making process of individual and effects the organization.

Looking at each product review or the movie's review is a laborious task and involves great time and effort to compile every review. Thus Natural Language Processing helps to solve this problem, it has ample approaches and techniques to extract the information from text. Typical approaches include counting number of words, identifying topics, summarizing, parsing the words to it's base form (Lemmatizing or stemming) etc. Using this structured data extracted from web, machine learning algorithms can be applied to generate classification and prediction models.

Mining opinions and analyzing sentiment from the web helps in various fields like analyzing mood of people on a particular aspect or event, knowing opinion of people on government schemes, knowing feedback, and event prediction. This also helps in discovering business demands, and potential areas to concentrate, this plays an important role in decision making.

There are a lot of Natural Language Processing(NLP) techniques available at present to aid the sentiment classification problem. Stemming, Stop Word removing, Parts of Speech Tagging (POS), Lemmatizing, Named Entity Recognition (NER), URL's removal, punctuations removal, case conversion followed by bag-of-words are some of the preprocessing techniques used for feature vector formation.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

In general there are two ways for opinion generation: traditional opinion mining and aspect based opinion mining. In traditional opinion mining the opinion is generated by two ways, one is by using Lexicon's or Dictionary like wordnet, vader sentiment, Textblob etc, other way is by using bag of words then applying machine learning algorithms on preprocessed data and predicting the sentiment like positive, negative, neutral. Whereas in aspect based opinion generation the aspect's of the sentence are tagged using parts of speech tagging(POS) and then the important features of the sentence are preserved and are used for generating the opinion. Using the Aspects in the opinion generation increases the prediction accuracy of the sentiment.

II. RELATED WORK

In literature survey there are two main approaches followed in mining opinion: Dictionary Based(DB) and Machine Learning Based(MLB). The Dictionary Based opinion mining refers to using the corpus or pre-built dictionary and finding the sentiment, the limitation in this is that the classification depends on the dictionary used and most of the DB uses the bag of words and their count as their feature vectors and may lead to misclassification. Machine Learning Based(MLB) uses the domain knowledge and trains on it by using different algorithms like Navie Basis, Maximum Entropy, SVM etc.

In 2017 Emad Shihab and Everton da Silva Maldonado [9], proposed a system to automatically analyse the technical debt in source code of different languages using NLP Techinques. Akshaya R. Garje and Karbhari V. Kale [3], classified the sentiment with sentiwordnet corpus and analysed them with machine learning algorithms. Prashant Raina [6], analysed the sentiment with corpus like ConceptNet and achieved 71% accuracy in classification. WARIH MAHARANI, DWI H. WIDYANTORO, MASAYU L. KHODRA [2], proposed system for aspect based summarization and compared text based summary and visual based summary and achieved better results with machine learning methods. Monisha Kanakaraj and Ram Mohana Reddy Guddeti [5], proposed sentiment classification with machine learning classifiers using ensemble classifiers, they eliminated class imbalance problem by training models with different set of data and achieved 90% precision using ensemble classifiers. Mark Hoogendoorn, Thomas Berger, Ava Schulz, Timo Stolz, and Peter Szolovits [1], assessed the patients mental behavior based on the email conversations with the doctor and patients over a time period and gives accurate results.

Our work is different from above mentioned works in terms of feature sets and sentiment classification. All the above methods use keyword based feature vectors while deciding polarity. Also the use of traditional machine learning algorithm in these systems leads to increased biasness towards a particular class. Our method preserves all important aspects by using the Aspect based classification.

III. PROPOSED SYSTEM

The proposed system gathers data from the twitter social networking site and processes data using NLP techniques to get feature vectors. Then ensemble methods are applied on the training data to form the training model followed by classification of the data. The proposed methodology has the high level process flow as shown in Figure 1.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

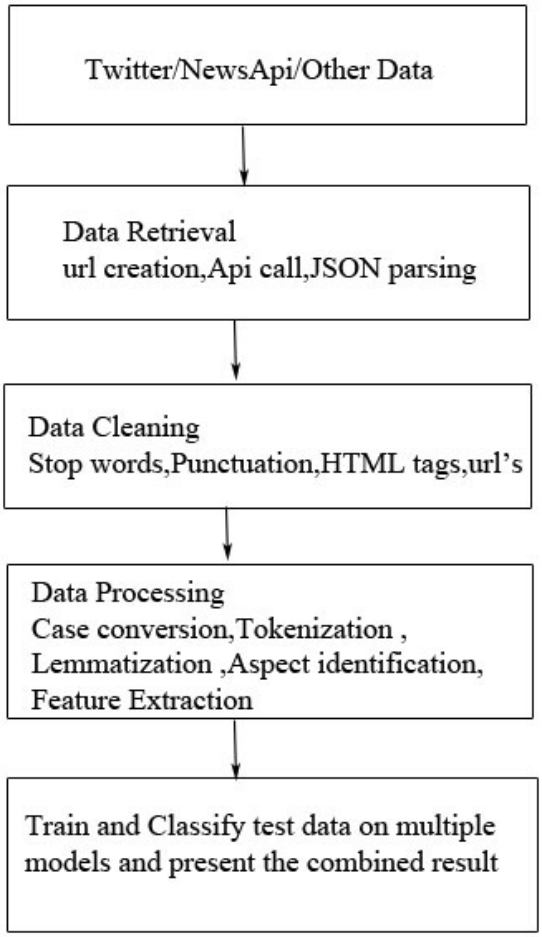

Figure 1

The above figure shows the proposed Methodology for aspect based classification
The system consists of the following modules

1. Data Retrieval Module
2. Data Cleaning Module
3. Data Processing Module
4. Classification Module

- Data Retrieval Module

The tweets or the News articles are fetched using the Twitter API and NewApi API. The API provides a more sophisticated programming interface through which search results can be obtained in JSON format. The Twitter JSON object helps in extracting specific tweet attributes say user name, geographic location, re-tweet count etc. The NewsApi API provides with attributes like News article source, description, url, article date published etc. Once the data is fetched preprocessing of the gathered data is done to extract features.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

- Data Cleaning Module

This module cleans the unwanted text or the data which does not have any valuable information or data which is not wanted for classification. In this module we clean the data like url's, stop word's, punctuation, HTML tags etc.

- Data Processing Module

In this module we process the data extracted from the previous module. The data undergo following processes to make classification process simpler

1. Case Conversion
2. Tokenization
3. Lemmatization
4. Aspect identification
5. Feature Extraction

1) Case Conversion: In order to simplify the process of analysis, we convert the whole data into Upper/Lower. So that words with same spelling but with different case are not differentiated by the classifier.

2) Tokenization: it is an act of breaking the sentences or text into smaller pieces such as words, keywords, phases, symbols etc. It is useful for word level classification and finding polarity for every word.

3) Lemmatization: it is the process of determining the root or lemma of a word. It is used to correctly identify the part of speech and meaning of a sentence.

4) Aspect identification: In this module the text is tagged with parts of speech and the aspect of it is determined then we preserve the aspect's of text by which we gain most information.

5) Feature Extraction: Once the text is tokenized and aspect's have been identified from the above module the next step is to form feature's. The Feature vector consists of key terms of text and their sense value.

- Classification Module

In Classification Module ensemble method is used which aims to combine the predictions of several classifications and machine learning algorithms in order to build a machine learning algorithm that minimizes the biasness of a single machine learning algorithm.

IV. EXPERIMENTAL SETUP

Python 3.6 version with NLTK library for Natural Language Processing (NLP) Techniques are used. It also uses pandas, numpy for data processing, matplotlib library for plotting the graphs, and sklearn library for the machine learning algorithms. The system gets the search keyword as input from the user. The user input keyword is used as search item in the Twitter API or News API. The results are analyzed by the NLTK and sklearn machine learning algorithm modules are applied on processed data.

V. RESULTS

Three parameters are taken for testing the system they are: Precision, Recall, F1 score
Precision: Ability of classification model to identify only relevant instances
Precision = TruePositive / (TruePositive + FalsePositive)
Recall: Ability of classification model to identify all relevant instances
Recall = TruePositive / (TruePositive + FalseNegative)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

F1 score: it combines recall and precision using harmonic mean
F1 Score = 2 × Precision × Recall / (Precision + Recall)

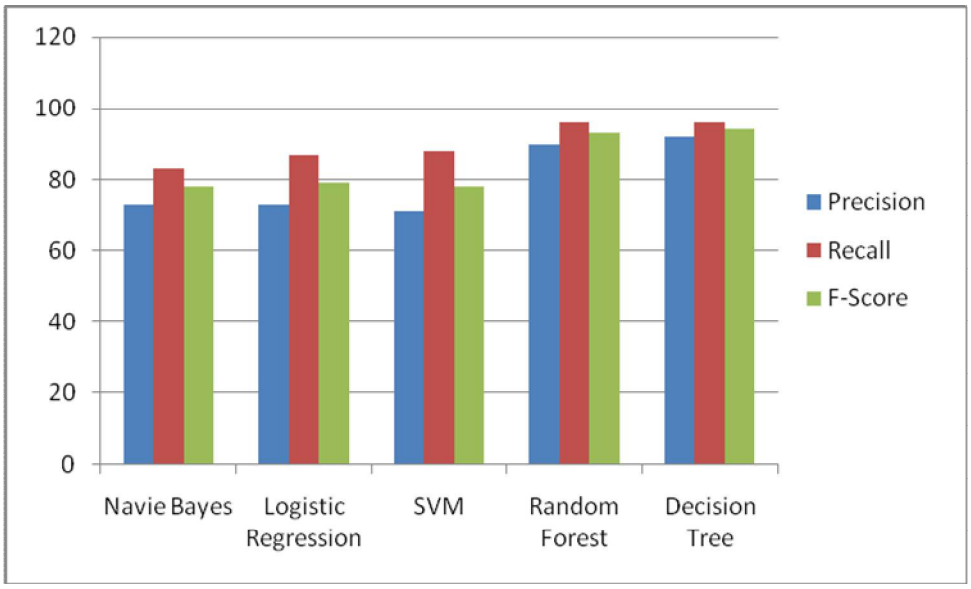

Figure 2

General Twitter classified tweets dataset with 200000 rows of tweets, 1000 of amazon tweets dataset, 748 classification tweets of IMDB and 1000 classified tweets of yelp datasets for training and testing. Figure 2 shows the comparative view of the machine learning algorithms. It can be observed that Random Forest and Decision Tree algorithms outperforms the traditional algorithms like Navie Bayes, Logistic Regression etc. It is also observed that accuracy increases with the increase in datasets.

VI. CONCLUSION

The proposed system collects data from social networking site like Twitter and NewApi and does the NLP techniques to extract the features from the dataset and preserves the aspects in the dataset. Then the machine learning algorithms are used on the training and testing sets to classify them as positive or negative. It is observed that Random Forest and Decision Tree algorithms outperforms the traditional algorithms like Navie Bayes, Logistic Regression.

REFERENCES

[1] Mark Hoogendoorn, Thomas Berger, Ava Schulz, Timo Stolz, and Peter Szolovits, "Predicting Social Anxiety Treatment Outcome Based on Therapeutic Email Conversations", IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 21, NO. 5, SEPTEMBER 2017

[2] WARIH MAHARANI, DWI H. WIDYANTORO, MASAYU L. KHODRA, "ASPECT-BASED OPINION SUMMARIZATION : A SURVEY", Journal of Theoretical and Applied Information Technology 31st January 2017. Vol.95. No.2

[3] Akshaya R. Garje, Karbhari V. Kale, "SENTIMENT POLARITY WITH SENTIWORDNET AND MACHINE LEARNING CLASSIFIERS", International Journal of Advanced Research in Computer Science Volume 8, No. 9, November-December 2017 ISSN No. 0976-5697

[4] Albert Weichselbraun, Stefan Gindl, Fabian Fischer and Svitlana Vakulenko, Arno Scharl, "Aspect-Based Extraction and Analysis of Affective Knowledge from Social Media Streams", IEEE INTELLIGEN T SYSTEMS

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

[5] Monisha Kanakaraj and Ram Mohana Reddy Guddeti, " NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers", 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)

[6] Prashant Raina, "Sentiment Analysis in News Articles Using Sentic Computing", 2013 IEEE 13th International Conference on Data Mining Workshops

[7] Ashwini Save and Narendra Shekokar, "Analysis of Cross Domain Sentiment Techniques", 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)

[8] Bruno Ohana and Bruno Ohana, "Sentiment Classification of Reviews Using SentiWordNet", 9th. IT & T Conference 2009-10-01

[9] Everton da Silva Maldonado and Emad Shihab, "Using Natural Language Processing to Automatically Detect Self-Admitted Technical Debt", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 43, NO. 11, NOVEMBER 2017

[10] N. Peter Whitehead, William T. Scherer and Michael C. Smith, "Use of Natural Language Processing to Discover Evidence of Systems Thinking", IEEE SYSTEMS JOURNAL, VOL. 11, NO. 4, DECEMBER 2017

[11] Kim Schouten and Flavius Frasincar, "Survey on Aspect-Level Sentiment Analysis", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 3, MARCH 2016

[12] Peiman Barnaghi, John G. Breslin1 and Parsa Ghaffari, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment", 2016 IEEE Second International Conference on Big Data Computing Service and Applications