International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Prediction and Analysis of Unstructured Text Data for Efficient Decision Making

Saurabh Shashikant Kulkarni[1], Manthan Mukund Bhandare[2], Kartik Ramesh Hiwase[3] Sharayu Sudhir Shende[4]
[1, 2, 3, 4]*Computer Science and Engineering, Sant Gadge Baba University*

Abstract: ***An enormous growth of the internet has been instrumental in spreading social networks. The rapid increase in the internet use has led to users taking online reviews and opinions on the internet for a particular product or service, the growth of social networks had hiked sharing and communication. This activities generates unstructured data which accounts for 80% of data in the world.There are different approaches to the problem of sentiment analysis. In this seminar, we discuss a model about complete process of sentiment analysis. The feedback can be in various forms unstructured and structured which makes it hard for the organizations to identify the exact problem. In the proposed work, an unstructured text analysis system is created that analysis the text according to their domain and predicts the polarity of the domain giving an efficient summarization of the complete reviews for efficient decision making.***

Keywords: Sentiment analysis, Ontology, Pos tagger, Jaccard, Polarity.

I. INTRODUCTION

The enormous growth in the internet has led users all around the world to search and share their opinions in a wide manner. There are billion searches in a day and people go for online reviews to get more opinions on the internet. The information sharing and communication have gained significant importance and has led to increase in unstructured data which is 80% of the overall data in the world. The feedback can be in various forms unstructured and structured which makes it hard for the organizations to identify the exact problem. A review is an assessment of a service, company, publications, movie or for a product. The review is given by a particular individual or an organization.

Review refers to feedback written by the user about his experience with the reviewed product. Reviews can be freely given and can be accessed. There is a small difference between reviews, opinions and feedback. Opinions are usually subjective expressions that describe people's sentiments and feelings towards a subject or topic .whereas feedback on the other hand is suggestions given by the receiver to the sender about how effective the product was and what improvement should be made in the product. Majority of consumers are influenced by the reviews during buying a product. The reviews play a key role when a user thinks of buying a product. Business reviews can be found in many places online. Reviews usually take the form of a score out of an arbitrary number (usually 5 or 10) and a comment by the reviewer to justify or support their score. Quite often, a graphical representation of the business review score is depicted, usually in the form of stars.

In a world where we generate 2.5 quintillion bytes of data every day, sentiment analysis has become a key tool for making sense out of that data. This has allowed companies to get insights about the problems and automating the processes. According to corporate estimates, only 21% of the available data is present in structured format. Data is generated as we speak, or tweet, send messages on facebook and in various other activities. Majority of Unstructured data may have its own internal structure, and need not be always given in tabular format or in a database.

Most business interactions and meetings are unstructured in nature. Today more than 80% of the data generated is unstructured. The fundamental challenge of unstructured data is that they are difficult for data analyst to make a sense out of it because this data has to be used for analytics. The challenges faced with unstructured data are as follows-

Subjectivity and Tone, Context and Polarity, Irony and Sarcasm, Comparisons, Emoji's.

In our research, we are using unstructured data because that will help in extracting more accurate and relevant information from the data. The problems we face in today is that the business which are much concerned about critics and public reviews need a more broader view of what percent of people are in favour of their product, what is the flaw in the product which leads to declining of revenues. So for such broader view we need an application which can make sense out of this unstructured data and produce results which can be understood by a naive user as well.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

II. LITERATURE REVIEW

Many researchers have conducted various experiments and gained various results with different models and got accuracies in different ranges Eric Brill presented a rule based simple pos tagger which performs same as existing stochastic taggers, but has notable advantages over other existing taggers.

The tagger is intensely portable. Many advanced level procedures were used for improving the performance of stochastic taggers and this cannot transfer to different language and different tag set. Most of it is acquired by rule based tagger except for proper noun discovery making it much more portable than a stochastic tagger. If the tagger would have been trained on different corpus, a different set of patches which are suitable for corpus can be found automatically.

Bernard Merialdo in his paper demonstrated with some experiments how to use probabilistic model for text in English language, it assigned correct part of speech to words in the text according to their context. They looked for best way to estimate parameters in the model, depending on training data provided and used simple triclass marclov model. The unique thing is the use of untagged text in the training of the model. Particularly two approaches were compared-use the text which is hand tagged and calculate frequency count & train hidden markov model and use text without tags.

A dataset was created by Akshay Amolik with the help of twitter post related to the movies. Sentiment analysis was done on the sentences.it is done in three phases the very first thing is preprocessing then based on the important features we create feature vector and then various classifiers like SVM, Naive bayes, ANN were used. SVM showed 75% accuracy. He contradicted a paper which made an observation that if username is encountered that could affect the probability so Akshay Amolik replaced it with AT_USER and with the help of SVM the accuracy improved by 10%. POS Tagging process contains tokenization and then assigning tag to it and search for ambiguous words.

By looking at preceding and following word disambiguation is removed.

In Sentiment Analysis Using Product Review Data by Xing Fang and Justin Zhan discussed some important and general problems in sentiment analysis like sentiment categorization by considering a dataset which has over 5.1 million reviews of product from Amazon .com and they are categorized according to various subjects.

Subjective content is extracted in this paper for future analysis.

Christopher Manning studied the important aspects required to make the pos tagger accuracy from 97% as token accuracy which contains 56% sentence accuracy upto to total 100% accuracy. According to my research it is much possible to improve the taggers performance and study them for convenient improvements which is recently adopted in the Stanford POS tagger. Conversely, the amount of errors in analysis of specific errors shows that it is by far limited to be one or the other from good machine learning or more specifically having better features for differential sequence classifier.

The predictions of future gains from an semi supervised learning technique seems fairly limited. Relatively, my research provides insights to reveal the biggest chances of future progress when it comes for improving the taxonomic basis of dialectal resources on which the taggers will be trained.

That is, from the enhanced descriptive semantics. Though, I conclude by providing that there will some limitations in this process. The pattern of some words will not be able to be totally seized by assigning them into some of the categories. While agreements which will be used in this case for improving the consistency of tagger, there is an absence of powerful linguistic base.

Shravan Vishwanathan et al., proposed Reviews of rotten tomato is collected from the one of the database. Than the process of tokenization is done on the reviews and the tokens are then filtered based on their length, then the process of stemming is done which brings it to root words and remove words which are not used in sentiment analysis.

The operator is used which is used to compare whether the word is positive or negative in the dictionary if there is match it will put the word into that category. After that sum all the values at both positive database and negative database. Apply join operator which will subtract the sum of positive and negative and display the reviews to the user.

Ahmad Kamal in his paper designed a opinion mining framework that makes it easy for feature extraction and review summarization, objectivity or subjectivity analysis etc. objectivity and subjectivity reviews were classified based on supervised learning approach. The various techniques used by him were Decision Tree, Naive bayes and Bagging. By preventing extraction of irrelevant data and filtered noise the mining performance can be improved.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

III.SYSTEM DESIGN

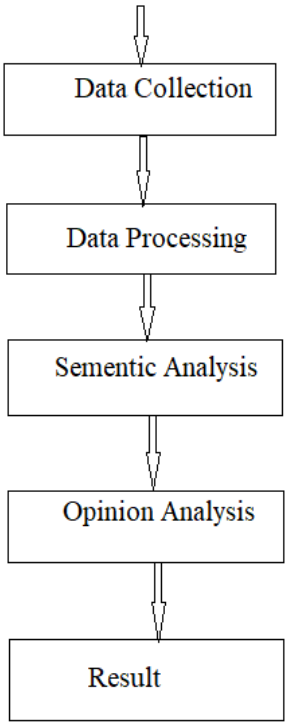

Figure 1 Step by Step model of Sentiment Analysis

A. Domain Ontology

Other font types may be used if needed for special purposes. 'Domain' name specifies the input is of relating to the some of the field or of the task. Domain is use to specifies the fields related to which user can search the data or user wants to get information related to it. And the domain ontology is specifying the various sub-fields i.e. the collection of the domain and sub-domain is known as domain ontology.

Similarity/Pattern matching algorithm: - Similarity algorithms are mainly used to determine the similar words by comparing the word. Similarity algorithms suggest or get output of the word which is similar to the input. In the similarity there where various types or various algorithms which gets maximum accuracy in processing.

Assigning sentences to domain or sub-domain after execution similarity algorithm: - After the similarity algorithms find out the domain and its sub-domain then that sentence get assigned to that particular sub-domain for getting the effective output. Firstly it select the domain with using the domain ontology and the similarity algorithms and the get assigned that sentences to that of the particular sub-domain. The ontology creates the organization structure like a binary tree diagram that helps in easily understanding the organization.

In our research, the proposed domain is about the movie. So the ontology is created on the basis of varies movie components like movie name, story, direction, acting, songs, etc. So we start the processing of the provided reviews after filtering out the significantly affecting reviews for the overall analysis. Tag clustering will provide the movie affecting reviews as such that the result of this further starts the post processing of the reviews on the ontology of the domain. The flow of ontology further creates a level wise classification on Domain, Sub-Domain, Entities and values

B. Part of Speech

Parts-of-speech (also known as POS, word classes), is important because of the large volume of information is classified on the basis of grammatical classes which will provide its sentiment of each word and its next vicinity for clarifying its existence which makes a difference. For finding out whether a word is a noun or a verb will help us a lot about occurring neighboring words (nouns are preceded by determiners and adjectives, verbs preceded by nouns) which makes tagger an important tool for parsing because it

makes checking for syntactic pattern on the words easier and will provide an efficient recognition of text. Parts of speech are useful in grammar as well as finding features for named entities like about a unique subject or people in text and other information for extraction of sentiment affecting part. Parts-of-speech deeply emphasize the possible morphological affixes which in turn can influence stemming of information retrieved, and will make the processing easier for choosing various types of nouns, verbs or other relevant words from a provided data. Our research is over Prediction and Analysis of Unstructured text data for efficient decision making. The input is the unstructured text which must be divided into meaningful group of words which in turn will used for predicting and then used for deriving actual meaning for various purposes depending upon the domain. The use of POS tagger is the pre-requisite for sentiment analysis which in turn required Machine Learning Algorithms which will used the properly tagged words by removing the ambiguity and hence making their meaning clear for the further analysis.
For Example- POS Tagger will provide Tagged words by removing ambiguity and can be used for polarity identification on the basis of particular domain specified like Social Media Review Analysis.

C. Jaccard Distance

The Jaccard distance is the measure of how dissimilar the sets are, it is complement of Jaccard coefficient and it is calculated by subtracting the Jaccard coefficient by 1. It can also be specified as difference of sizes of the two sets containing union and intersection divided by the size of the union. The two strings would be more similar if the value of the distance is less.
Jaccard Distance also depends on concept known as “Jaccard similarity index” which states that (the number in both sets)/(the number in either set)*100
$J(X,Y) = |X \cap Y| / |X \cup Y|$
The above calculated is Jaccard coefficient now we can calculate the Jaccard Distance as follows:
$D(X,Y) = 1 – J(X,Y)$

D. Polarity Calculation

Polarity normally means the intensity of the expressed emotions in the provided text. Emotions can be closely studied as the factors affecting the sentiment. The power of a sentiments or opinions has an direct impact towards the intensity of particular emotions, e.g. sad, joy, anger, surprised etc. Polarity can also be known as scale measure of the different levels of words in the data. Polarity is shown mathematically as +ve(positive), -ve(negative) & 0(neutral).

1) *Example:* Tweets on Sanju movie
2) Sanju movie is very good movie.
→(Polarity =+2) (It consist of positive word with stress (like: very) on which the sentence contain positive approach towards movie so that sentence make polarity of +2 for search)
Ranbir Kapoor act fabulous in Sanju movie.
→(Polarity =+2)
3) Songs of Sanju movie are good. →(Polarity = +1)

In our research, we are using the good and bad words dataset consist various polarity so that we can easily calculate the positive or negative approach of the sentence. We are defining three various polarity values regarding to their degrees.
For example: “Good” keyword consist +1 polarity while “Better” keyword consist +2 polarity and “Best” keyword consist +3 polarity. Similarly “Bad” keyword consist -1 polarity while “Worst” keyword consist -2 polarity. Also the polarity will be calculated with the help of adverb.

E. Experiment Results

Accuracy is not the only parameter for evaluating the classifier. The two metrics such as precision and recall can give greater insights about the classifier being used.
In our research, the precision is ratio of successfully implemented positive reviews to all the positive reviews present in the file. Precision means if we have less false positive then we will have high precision value. For example, the system searches similarity match between the words and the word list for the data set file and finds 131 sentences, 123 of which are really correct, then the system precision found is 93.8%.
Recall is the ratio of successfully implemented positive reviews to the sum of successful positive and false negative reviews present in the file. Higher value of recall means there are less false negatives. For example, the system searches similarity match between the words and the word list for the data set file and finds 166 sentences, 123 of which are really correct, then the system recall value is 74.09%.

F score is the average of precision and recall and takes both false positive and false negative into consideration. F score is also a very important metric and is useful as that of accuracy. The system f score value is 82.82%.
The values calculated on the basis of precision and recall are plotted in the graph given above.

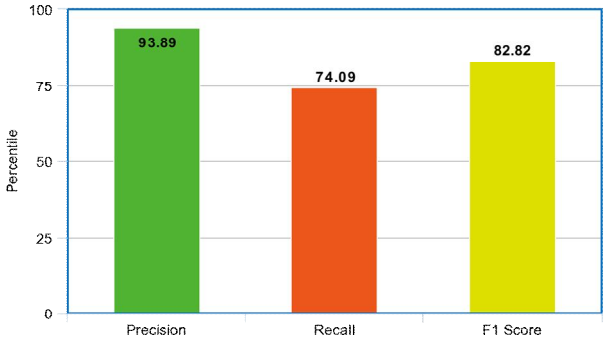

Figure 2 Metrics value

From the result of the sentiment analysis it is understood that the system will work more efficiently by using supervised machine learning models and using more trained dataset.

IV. CONCLUSIONS

The Sentimentizer tool gives best three parameters of the movie whose reviews are to be reviewed which is not present in the existing Movie review sites. This makes it easier for the user to get all the relevant information about the movie on single click. The use of tokenization and NLTK pos tagger tags the words in the sentences with the accuracy of 98%. The output of the pos tagger to which entity identification grammar enables the system to extract all the noun phrase from the sentences is then clustered according to noun tags with the help of jaccard distance. The keywords are matched with the domain ontology data set and gets forwarded to the classifier to calculate the polarity. The precision and recall of the system provide good results.

V. ACKNOWLEDGMENT

The authors would like to express their gratitude to Prof.Pratik Agrawal for his continuos encouragement and support.

REFERENCES

[1] Manning, C.D., 2011, February. Part-of-speech tagging from 97% to 100%: is it time for some linguistics.In International conference on intelligent text processing and computational linguistics (pp.171-189).Springer, Berlin.
[2] Amir Hossein Yazdavar, MonirehEbrahimi, Naomie Salim, 2016, Fuzzy Based Implicit Sentiment Analysis on Quantitative Sentences, Faculty of Computing, UniversitiTechnologi Malaysia, Johor, Malaysia, Journal of Soft Computing and Decision Support Systems vol 3:4, pp.7-18.
[3] Brill, E., 1992, March. A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing (pp. 152-155). Association for Computational Linguistics
[4] Merialdo, B., 1994. Tagging English text with a probabilistic model. Computational linguistics, 20(2), pp.155-171
[5] Fang, X. and Zhan, J., 2015. Sentiment analysis using product review data. Journal of Big Data, 2(1), p.5.
[6] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.
[7] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.
[8] Kamal A., 2015, Review Mining for Feature Based Opinion Summarization and Visualization
[9] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr .M. Venkatesan, Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, School of Computer Science and Engineering, VIT University, Vellore
[10] Manning, C.D., 2011, February. Part-of-speech tagging from 97% to 100%: is it time for some linguistics.In International conference on intelligent text processing and computational linguistics (pp.171-189).Springer, Berlin.