A Survey paper on Sentiment Analysis Approaches On Tweets

[1]Kajal Rathore, [2]Deepika Punj
[1]Research Scholar, [2]Assistant Professor
[1]Information & Technology Department,
[1] J.C. Bose University of Science & Technology YMCA, Faridabad, India

Abstract : Data Analysis is a key to grow any business these days. It helps in determining more informed business decisions by scientists, researchers to verify models theories and hypothesis. As social media is coming into picture these days which concludes huge corpus of data with reviews from every user in public platform as a result sentiment analysis gained importance. Sentiment analysis (also known as opinion mining) plays a major role in data analytics useful for business and at various government aspects. Sentiment analysis is related to mining of sentiments (i.e. tweets) from online data present on www and classifying the opinion indicated by author into three parts- Positive, Negative, and neutral. Unstructured data can be classsified on twitter by sentiment analysis approach and then different techniques are carried out which are discussed in this paper.

IndexTerms **-Sentiment Analysis, Corpus.**

INTRODUCTION

Social computing is outgrowing for analysis of recent trends and modeling of social activities taking part. The internet has rapidly advanced from a static to an interactive medium. Now a days not only obtaining information is there but content is also extracted from it. Reports, forums, blogging, online discussion, comments etc. are main sources of information from social platforms. The text consists of opinions and facts which are used to extract views and opinions by natural language processing. Opinions are a people's sentiment, appreciations, demands, feeling towards entity. Whenever we see an opinion we do not talk about concerned topic instead we are much more interested in the opinions score if it is positive, negative or neutral. Internet consists of huge amount of information available to get an opinion about an entity. Opinion mining system analyzes: Part which expresses opinion, Written by whom, and what is the comment? Sentiment analysis is the finding out the popularity score, subjectivity and polarity strength of a piece of text. Sentiment analysis is carried out at different levels ranging from coarse level to fine level. Coarse level determines the whole documents whereas the fine level determines the attributes of the sentiment. The twitter data is carried out at sentence level which comes between coarse level and fine level. Sentiments in tweets (or text) are of two types- direct and comparative. Direct sentiments are those where the object is independent of other entities. Example- Dior has great perfumes. Comparative sentiments are dependent on other entities, comparison is done. Example – Tanishq has better gold than Kalyan jewelers.

In this paper we discuss various sentiment analysis classification techniques to be performed for determining id the opinions of the reader.

The existing sentiment analysis techniques are useful in various areas like disaster prone areas, business

development, marketing, advertising, finding job opportunities, various surveys, predictions, trade polls etc.

Sentiment classification is done in three stages-

Document level

Sentence level

Feature level

In document and sentence level the classification makes use of single entity and extracting of single opinion is done. Fig 1 shows the overall sentiment process from preprocessing of a dataset to resulting in positive negative or neutral review.

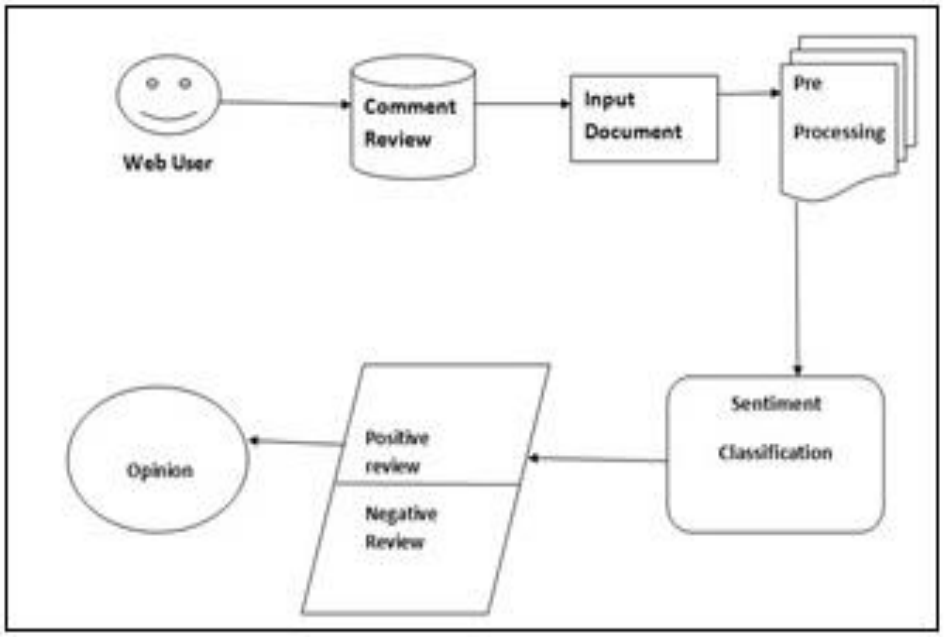

Fig 1 Sentiment Analysis Concepts

SENTIMENT CLASSIFICATION TECHNIQUES

Sentiment classification techniques are numerous. We will discuss some of the most commonly used approaches with the data set provided. Sentiment classification techniques are- Dictionary based, Ontology based, Machine learning.

Dictionary Based Approach

This approach is used to perform sentiment analysis by operating on sentiments. There are predefined dictionaries (WordNet, SentiwordNet etc) which consist of synonyms and antonyms for every sentiment. An easy approach is to use senti words to bootstrap based on synonyms and antonyms classification in dictionaries. The paradigm is as follows: few seeds of sentiment words with their values of positive(+1), Negative(-1) and neutral (0) is gathered manually. Further set is grown by finding in any dictionary for the respective synonym and antonym. Then the found words are combined with seed list. And the iterations proceed..

The procedure is ended when no new word is collected and at last when process is completed, it is checked manually to clean the list.

Ontology Based Approach

Ontology is a philosophical word used in various aspects meaning "theory of existence". This approach provides a mechanism for sentiment analysis to improve results of natural language processing. Ontology along with machine learning approach is most appropriate for sentiment analysis. This approach is used to determine structure of some specific domain. It understands the techniques and concepts in domain and also the relationship between these. There are three types of ontologies: (1) Domain Ontology. (2)Hybrid Ontology (3)Upper Ontology. Most preferred ontology techniques among all of them is domain ontology for opinion mining The important aspect of ontology is taxonomy which is developed in hierarchal form. It is developed using various ontology languages like OWL(Ontology Web Language) and RDF(Resource Description Framework). The tool used is Protégé.

It includes various constituents like individuals, classes & properties.

Machine learning approach

This approach is further divided into two categories: Structured and Unstructured Learning.

Machine learning is amalgamation of methods to determine accessible framework in given dataset [3]. It is also used to predict patterns to forecast future data. Supervised learning uses target value or label. Unsupervised learning does not use any label or target value.

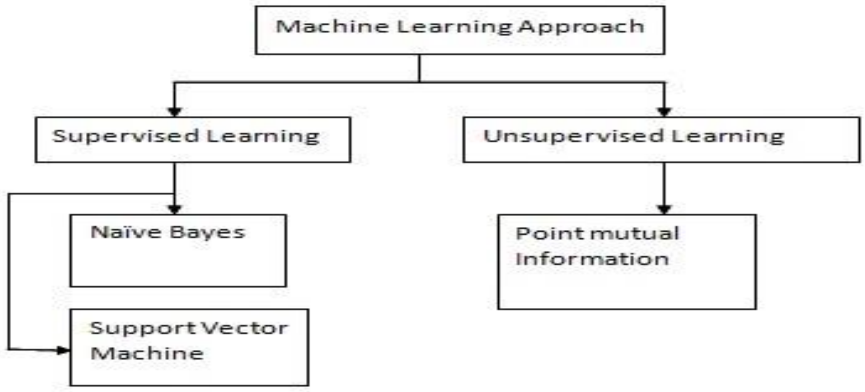

Fig2. Machine learning approaches.

A.1 Supervised Learning

- ***Naïve Bayes Approach-***

It is simple probabilistic classifier that uses concept of mixture models to perform classification P(C|F)

states conditional probability of C given F,

C is Class Label

F is Feature

$$P(C|F) = \frac{P(C)\,P(F|C)}{P(F)}$$

It is the process of calculating conditional probability from prior probabilities assuming present feature is unrelated to presence of any other feature.

Advantage of naïve bayes is small training set helps to train the model in efficient way. Also ensures and check quality of model which can be used. It is widely used in research area for sentiment analysis.

- ***Support Vector Machine***

SVM is used to solve text categorization problem efficiently. It performs better than naïve bayes. It supports concept of maximum margins also. It has two forms- linear and nonlinear with supervised classifier. A line is present which divides two classes separately in the search space. SVM classifier is widely used these days in research areas because of the improvement in overall good empirical formula. SVM classifier is used to find boundaries that separate group of dataset. It is constructed using set of points and separate them by mathematics formula Fig.3 shows the data flow of SVM.

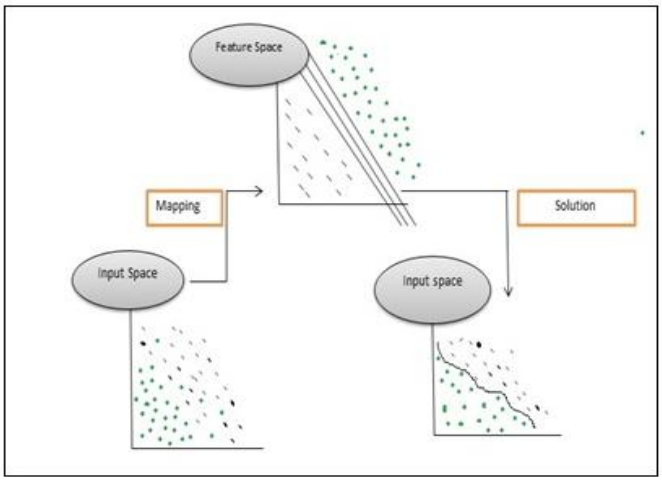

Fig3. Flow of SVM

A.2 Unsupervised Learning

- ***Lexicon based method***

There exist two main approaches to the problem of extracting sentiment automatically. The lexicon-based approach involves calculating orientation for a document from the semantic orientation of words or phrases in the document. The text classification approach involves building classifiers from labeled instances of texts or sentences. Dictionaries for lexicon-based approaches can be created manually. Much of the lexicon-based research has focused on using adjectives as indicators of the semantic orientation of text. Semantic orientation refers to the polarity and strength of words, phrases, or texts. Our concern is primarily with the

semantic orientation of texts, but we extract the sentiment of words and phrases first, a list of adjectives and corresponding Sentiment Orientation values is compiled into a dictionary. Then, for any given text, all adjectives are extracted and annotated with their SO value, using the dictionary scores. The SO scores are in turn aggregated into a single score for the text.

III.COMPARISION

Algorithm	Naïve Bayes	Support Vector Machine	Lexicon Based Method
Understanding Complexity	Very less	High	High
Theoretical accuracy	Low	High	Moderate
Theoretical Training Speed	High	High	Moderate
Performance with less no. of observations	High	Low	Low
Classifier used	Probabilistic	Linear	Lexicon
Accuracy	79.67%	87.15%	73.33%

S.No.	Title	Issues Addressed	Proposed technique	Dataset used	Limitations
1.	Fangzhao Wu, Yongfeng Huang, Yangqiu Song, Shixia Liu," Towards building a high quality micro blog -specific Chinese sentiment lexicon", Decision Support	Not Accurate	Micro blog specific sentiment lexicon	Dataset from Tecent Webio with 2013 labeled messages on 20 topics	Only for Chinese blog
2.	V.K. Singh, R. Piryani, A. Uddin, P. Waila," Sentiment Analysis of Movie Reviews", conference on	Polarity shift and accuracy	Lexicon based Approach	Movie Review Dataset	Difficult to update the dictionary
3.	RuiXia, FengXu, Jianfei Yu(2016)	Polarity shift Problem	PSDEE approach based on Rule based and statistical	Movie Review	Limited Accuracy

			method		
4.	Farman Alia, Kyung - Sup Kwaa,Yong -Gi Kimb," Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification".	Binary Classification Problem and Accuray	Fuzzy Ontology with Machine learning Technique	Hotel Review	Increased Complexity
5.	Afef Walha, Faiza Ghozzi and Faïez Gargouri "A Lexicon Approach to Multidimensional Analysis of Tweets Opinion"	opinion extraction from informal tex	Opinion lexicon building method, POLSentiment algorithm, based on lexical resources	SentiwordNet, Sanders twitter corpus.	Increased Complexity

IV.APPLICATIONS OF OPINION MINING AND SENTIMENT ANALYSIS

1. Automatic content analysis can help in preprocessing huge corpus of qualitative data.[1] this is emerging area in research.
2. It helps in determining new products and services of a product with customer's opinion attached to it.
3. It also helps in functional and non-functional requirements of the product using semantic analysis by gathering opinion.
4. Can also assist government in judging their strengths and weaknesses by processing sentiments and resulting in an opinion.[2]

V. CHALLENGES AND GAPS

1. Presence of fake and duplicate reviews can result in different opinion of a particular sentiment.
2. Continuous need for better usage and user interaction of tools, which are mainly used by data analytics.[1]
3. Limiting collaborative filtering.[1]
4. Pragmatics is also difficult challenge where there can be a sentence having no sentiment bearing word but consist of indirect sentiment.[2]

VI.CONCLUSION

Twitter sentiment analysis is a very important and challenging task. Twitter being a micro blog suffers from various linguistic and grammatical errors. In this paper, we studied various popularity methods which

incorporates the popularity effect of words on tweet sentiment classification and also emphasis on how to pre-process the Twitter data for maximum information extraction out of the small content. In future, we can work on analyzing the tweets based on hash tag segmentation and context dependency.

REFERENCES

[1] David Osimo and Francesco Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis"
[2] Sujata Rani and parteek kumar 2011,'Rule based sentiment analysis system for analyzing tweets',International conference on infocom technologies and unmanned systems(ICTUS),Dubai,UAE,18-20 Dec. 2017.
[3] A.M.Abirami and V.Gayathri 2016,"Survey on sentiment analysis methods and approaches',2016 IEEE eighth international conference on advanced computing(ICoAC), Chennai,India,19-21 Jan 2017.
[4] Go, A., Bhayani, R. and Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. In CS224N Project Report, Stanford University.
[5] Godbole, N., Srinivasaiah, M. and Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. In Proceedings of the International Conference onWeblogs and Social Media (ICWSM).
[6] He, B., Macdonald, C., He, J. and Ounis, I. (2008). An effective statistical approach to blog post opinion retrieval. In Proceedings of the 17th ACMconference on Information and knowledge management CIKM '08.
[7] Hu, M. and Liu, B. (2004). Mining Opinion Features in Customer Reviews. In AAAI. Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM *38*, 39–41.
[8] Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques.
[9] Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In ACL.