Sentiment Analysis for stack exchange data

Abstract— **With the development of Information of Technology, there is a vast incensement of data in digital form, for example in social media site, online communities and digital repository. People share their knowledge, opinions via online communities. Monitoring the activities related to social media is very important fact. Because it is very helpful to measure the quality of social media activities.**

Through this paper, describes sentiment analysis study for posts and comments of stack exchange. Sentiment analysis can be used to derive decision based on the knowledge mapping. Mainly three steps are involved for sentiment analysis. Data preprocessing, generating feature vector after feature extraction, generating a model for classification after training a model using machine learning algorithm.

I. INTRODUCTION

Online communities are the best places to share information and opinions with each other. With the increment of popularity of online communities, people adopt to use to share anything via social media without considering the qualitative of sharing information. There are plenty of social networks available at present but it is highly doubtful that these systems and their administration work according to the best interest of the public.

Sentiment analysis is a process of categorizing and identifying the meaning which is expressed through a text in a computational manner. According to the writer's view or desire it can be vary like positive, negative or neutral. This is a method of extracting subjective information from a source of a text. Sentiment analysis is called as Opinion mining, Opinion extraction, Subjective analysis and Polarity analysis also.

In this research paper, has been described clearly what are the related works related to sentiment analysis, data preprocessing steps, and when doing data preprocessing according to the feature extraction, features which are considered for creating feature vector for this analysis and about the classification.

II. RELATED WORK

As far, large number of researches have been done for sentiment analysis. They have been followed various types of technologies, methods and techniques.

Sentiment analysis are done by different levels. Such as document level, sentence level and entity and aspect level. Document level means whole document is considered as one entity in this level. Therefore analysis should be applied for whole document. As a result of that, the result of the analysis sometimes not correct. Sentence level can be introduced as sentence is considered as one entity. So that analysis should be applied for each sentence. Should be taken a summary as the overall result of a document. In entity and aspect level should consider about the positive sentiment or negative sentiment as quality of the features. And this method is built on summarization and opinion mining. [2]

Data preprocessing is an essential fact for sentiment analysis. Many of data that extracted from online platforms are noisy data. Therefore the data should be preprocessed before doing analysis. Data preprocessing can be done into three steps as tokenization, normalization and POS (Part-Of-Speech) tagging. [6] Can be followed different ways to do preprocessing. As the preprocessing techniques can be introduced removing URLs, removing tags, reducing lemmatization and stemming, removing synonyms, removing negation words and removing repeating words. [5]

Mainly sentiment analysis can be divided into two steps. Such as feature extraction, train the model using classification algorithm for testing purposes.

1) Feature extraction

Feature extraction is the main research part of sentiment analysis. Different types of methods have been used in various ways for different researches.

- n-gram features

As the first step should be removed stop words for identifying the n-grams witch are useful for sentiment analysis. Then identify the words which are followed or preceded the word 'not'. All of the unigrams and bigrams which are in training data set should be identified using Chi-squared related to the information gain. [6]

- Part-of-speech features

In documents or sentences that are needed to analysis, should be counted number of nouns, verbs, adjectives, adverbs and the other parts which are related to particular speech. [6]

- Lexicon feature

For this analysis use a dictionaries of words which are annotated as positive, negative and neutral. Should be annotated all of the words which are related to particular dictionary considering the meaning. [6]

- TF-IDF

Term Frequency – Inverse Document Frequency is used for classifying documents considering the meaning of those as positive and negative.

TF - A term is how many times has been occurred is called as term frequency.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

IDF – This is used for measuring how important a term for analyzing.

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

- Feature extraction when absence and presence the domain knowledge

Mainly two techniques have introduced when domain knowledge is available to feature extraction. One method is done when considering absence of domain knowledge. By considering absence of earlier information of the domain and, can create a new list by including only possible nouns as the features of the list of the review. And also should POS-tagged every words of the sentences to retrieve nouns. Then use an algorithm to prune features from the original list of words. In this algorithm check whether which are the words strongly match to keep as together. Those words are combined and others are pruned. The other method is done when information are available. All the features of domain can be extracted using Latent Dirichlet Allocation or HMM-LDA. When domain knowledge is presence, we can identify the specific features related to the domain easily and also can prune features directly. [4]

2) Sentiment Classification

Sentiment classification is mainly divided into three approaches, such as lexicon based approach, machine learning approach and hybrid approach. Hybrid method can be introduced as a combination of machine learning method and lexicon based method. Then lexicon based approach divided into two sections as dictionary based approach and corpus based approach. Corpus based approach can be done statistical way or a semantic way. Machine learning approach can be categorized as unsupervised learning and supervised learning. We mainly focus in sentiment analysis for supervised learning. It is categorized basically into four sections as decision tree classifiers, linear classifiers, rule based classifiers and probabilistic classifiers. When studying about linear classifiers, there are two sections as support vector machines (SVM) and neural network. And also probabilistic classifiers can be divided into Naïve Bayes, Bayesian Network and Maximum Entropy. [1]

When studying earlier researches can be concluded that most of researches have been used three machine learning algorithms for classifications. Such as Naïve Bayes, maximum entropy classification and support vector machines.

- Naïve Bayes classifier

This is simple than to other methods. And this is used for many classifications. Naïve Bayes based on Byes Theorem to decide probability based on label of the feature set that is given. This classification calculates the posterior probability related to a class, by considering the words spreading within the document. According to the feature extraction of Bag-of-words, this module is worked with ignoring the how word is positioned in the document.

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)}$$

The prior probability of a label or the likelihood that a random feature set the label can be introduced as $P(label)$. $P(features \mid label)$ is the prior probability that a given feature set(test data) is being classified as a label. $P(features)$ is the prior probability that a given feature set is occurred. As an assumption of Naïve can be said independent the all of the features that mentioned. According to this concept equation can rewrite as below,

Given the Naive assumption which states that all features are independent, the equation could be rewritten as follows:

$$P(label|features) = \frac{P(label) * P(f1|label) * \ldots.. * P(fn|label)}{P(features)}$$

In Naïve Bayes classification, consider every word as a feature. And also When applying this classification we supposed that there is no correlation among the words of the same text. [3]

- Maximum entropy classifier

This is introduced as a conditional exponential classifier. Using encoding this converts labeled feature sets to vectors. For calculating weights for each and every feature done by using encoded vector. As a result of that, for the set of features can define maximum likely label. The probability for all the labels is calculated using below one. [3]

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))forlinlabels)}$$

- Support vector machines

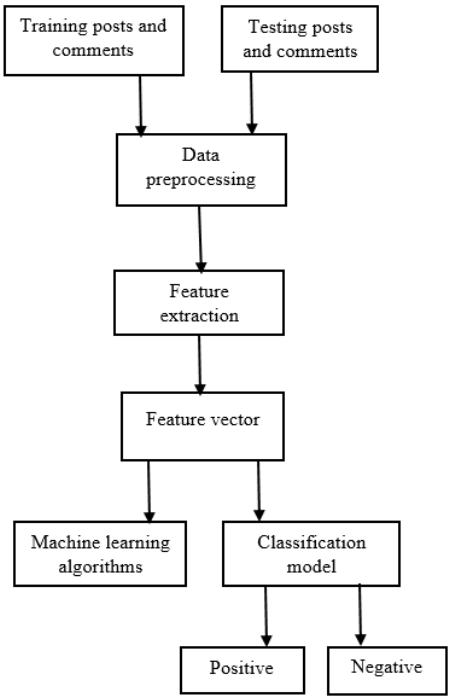
This is a supervised learning technique. This can be apply for classification and regression also. This performs the classification by finding hyperplane that maximize the margin between two classes. This is mainly use for binary classification and linearly separable data. Find the best line by maximizing margin width. [3]

III. APPROACH

As the process of this approach, sentiment analysis is done by using posts and comments that posted by users of the stack exchange. Sentiment analysis is done for analyzing which are the positive posts and comments and which are the negative posts comments.

For this should use data preprocessing techniques to construct data as needed. In this approach data preprocessing is done step by step. Because some data are essential for sentiment analysis. Therefore before preprocessing, should be extracted those data. For creating feature vector, nearly ten features are considered in this research. Then should be trained the model using machine learning algorithm. After that can check the positivity and negativity of the posts and comments using testing data set.

IV. METHODOLOGY AND IMPLEMENTATION



After extracting data from stack exchange basically, should be done preprocessing for both training and testing data. There are different types of preprocessing steps. Such as removing URLs and html tags, Sentence tokenization, POS tagging, case conversion, removing stop words and lemmatization.

Main research part of sentiment analysis is creating a feature vector to train the model. When creating this, should be considered various types of features. When gathering features, preprocessing should be done step vise. Otherwise data which are needed to sentiments can be lost.

As the first preprocessing step, should be removed URLs. Because there is no any contribution for sentiments from URLs. Then should do sentence tokenization. After tokenizing sentences should be done Part-Of-Speech (POS) tagging. Later can be identified each nouns, verbs, adverbs, adjectives, adverbs, particles, conjunctions and etc. This is very supportive for detecting keywords which are given sentiment value. While considering POS tags, adjectives are considered as keywords as the assumption. So can maintain separately positive keywords and negative keywords which are giving sentiment values for analyzing. Later can be done case conversion. In this step all the upper case letters are converted to lower case letters. As the next preprocessing part should be removed stop words. Because stop words are also not contributed as the sentiment values. Morphological features are another feature for sentiment analysis. On behalf of morphological features, prefixes and suffixes are considered. Some prefixes and suffixes are gained negative impact. And also some prefixes and suffixes are gained positive impact. According to the orthographic features, can be detected emoticons. Hence emoticons are given negative sentiments and positive sentiment also.

Number of negation words also can be identified as one feature. And also can be considered the bigrams as a feature for creating feature metric. When getting bigrams for all the words which are in the corpus, it is hard to handle them. Therefore should be considered only the bigrams which are added sentiment values. Such as bigrams with positive sentiment, bigrams with negative sentiments and bigrams with negations. Those bigrams can be kept as a dictionary and when creating feature vector, can be used those. Should be calculated negative bigrams and positive bigrams separately.

After considering all above features can be created feature vector. Using a machine learning algorithm (Naive Bayes Algorithm) can be trained the model. This model can be used for classifying the posts and comments as positive or negative.

REFERENCES

[1] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, pp. 1093 - 1113, 2014.

[2] A. D'Andrea, F. Ferri, P. Grifoni and T. Guzzo, "Approaches, Tools and Applications for Sentiment Analysis Implementation," International Journal of Computer Applications, pp. 1-8, 2015.

[3] S.-A. Bahrainian and A. Dengel, "Sentiment Analysis using Sentiment Features," in IEEE/WIC/ACM International Conference on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013.

[4] S. Mukherjee and P. Bhattacharyya, "Feature Specific Sentiment Analysis for Product Reviews".

[5] Y. Bao, C. Quan, L. Wang and F. Ren, "The Role of Pre-processing in Twitter Sentiment Analysis," pp. 1-2.

[6] E. Kouloumpis, T. Wilson and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.