Sentiment Based Recommendation System for Psychological Patterns of Social Media Users

[1]Dolly Tejwani, [2]Prachi Dabhade, [3]Vaidehi Satpute,
[4]Shital Gopatwad,
Pimpri Chinchwad College of Engineering, Pune

Prof. Shrikant Kokate
Asst. Prof. Department of Computer Engineering,
Pimpri Chinchwad College of Engineering, Pune

Abstract -Sentiment analysis is one of the fastest growing research areas in computer science. This paper deals with the recommendation based on the sentiment analysis of text, specially the posts that are posted by the users of the social media. The sentiment found within the posts and comments provide useful indicators for many different purposes as people nowadays tend to express their emotions on social media publicly. In existing systems sentiment analysis is done using lexicon based approach, rule based approach or by using machine learning algorithms In this proposed system, sentiment analysis will be done using text mining algorithms and depending on the intensity of the sentiment the user will be recommended with positive quotes and images to lift up their moods.

Keywords–Sentiment Analysis, Text Mining, Recommendation System, Social Media, K-Means, Apriori.

I. INTRODUCTION

Data mining is an analytical process designed to explore data in search of values and relationship between the variables and applying the new values to new subset of data. Data mining is the process of extracting value from a database. Data warehouse is a place where all this data is stored. The amount and the type of data largely depends on company to company. Data mining is a new term, but not the technology. First of all, what is the difference between Data and Knowledge? Data is collection of facts, numbers and figures that can be processed by the computer. There can be different sources of data like banks, insurance companies, hospitals, government etc. Information can be said as patterns or relation among the data. This information can be converted into knowledge of historical data and future trends.

Data mining mainly consists of five major elements, first is ETL that is extract, transform And load transaction data onto the data warehouse system. Second is to store and manage the data in multidimensional database system. Third is to analyse the data by application software and fourth is to analyse the data in useful format such as graph or table.

A. Data Mining Process

- *Data Integration:* Data from multiple sources are integrated into one.
- *Data Selection:* Data relevant to the problem or particular domain is selected. For example there is no need for insurance company to know about hobbies of a person. So only the relevant data should be selected.
- *Data Cleaning:* It is the process where noisy data, irrelevant data, missing values and erroneous data are handled.
- *Data transformation:* Data is transformed into the form which is suitable for mining.
- *Data Mining:* Intelligent Methods are applied to extract data patterns.
- *Pattern Evaluation:* To evaluate and select the truly interesting patterns among the discovered patterns
- *Knowledge Presentation:* Knowledge representation techniques and visualization are used to present the knowledge.

II. EXISTING CLASSIFICATION TECHNIQUES

The existing work on sentiment analysis can be classified from different points of views, technique used, view of the text, level of detailed of text analysis, rating level etc. From technical point of view, machine learning, lexicon based, statistical and rule based approaches.

- The machine learning method uses several learning algorithms to determine the sentiment by training on the known dataset.
- The lexicon based approach involves calculating sentiment polarity for review using the semantic orientation of words or sentences in the review. The semantic orientation is a measure of subjectivity and opinion to text.
- The rule based approach looks for opinion works in a text and then classifies it based on the number of positive and negative words. Different rules for classification such as dictionary polarity, negation words, booster words, idioms, emotions, mixed opinions etc.

- Statistical models represent each review as a mixture of latent aspects and ratings, it is assumed that aspects and their ratings can be represented by multinomial distributions and try to cluster head terms into aspects and sentiments into ratings.
- Traditional methods (e.g. self-report ratings, structured interview, and clinical judgment) cannot identify sentiments appropriately in real-time which might lead to delayed reporting. Usually self- reports regarding such issues is very meagre and clinical judgments starts after much more time than desired.

Opinion Digger: Opinion digger operates in two steps. First, it determines the set of aspects. After the pre-processing, each sentence is tagged with POS. It assumes that aspects are nouns so it first isolates the frequent nouns as potential aspects. With the sentences matching the known aspects, they determine opinion patterns, sequence of POS-tags that expressed opinion on an aspect. The frequent patterns used with known aspects are considered opinion patterns. If reviews with a "potential aspect" noun match at least two different opinion aspects, Opinion considers the noun as an aspect. The second phase is rating the aspects. For each sentence containing an aspect, Opinion Digger associates the closest adjective to the opinion. It searches two synonyms from the guideline in the Word Net synonymy graph. The estimated rating of the aspect is the weighted average of the corresponding rating in the guideline. Weight is calculated by the inverse of the minimum path distance between the opinion adjective and the guideline's adjective in the Word Net hierarchy The experiments show good performance in aspect determination and an excellent accuracy in ratings. The evaluation of aspect ratings was made using only the known set of aspects and compared to 3 other unsupervised methods. Opinion Digger performs an average ranking loss of only 0.49, meaning the difference between estimated and actual ratings. By incorporating new current information in the machine learning process, Opinion Digger increases the accuracy of unsupervised machine-learning methods.

III. PROPOSED TECHNIQUE

The main aim of our proposed system is to take certain actions based on the psychological patterns determined by the sentiment analysis of user's posts and also to provide solutions in extreme cases.

The proposed system consists of following stages:

- Pre-processing the data
- Classification into positive, negative and neutral
- Determining intensity of negative posts
- Providing recommendations according to the class and intensity of post

A. Pre-Processing the Data:

a) Transform characters to lower case.
b) Remove punctuation marks.
c) Remove unnecessary works like pronouns,

In this stage the unnecessary words like nouns, pronouns or helping verbs are omitted. Only key words like 'happy', 'sad' or 'lonely' etc. are found and considered for further stages of the system.

For example: Consider a sentence "I am happy."

Here, 'I' and 'am' will be omitted as these words do not provide any help in determining whether the post is positive, negative or neutral.

And only word 'happy' will be considered as a keyword and will be considered for determining the class of the post and for further sentiment analysis.

B. Classification into Positive, Negative and Neutral

1) After determining the context of extracted keywords this stage will come into picture. Two dictionaries will be maintained as basic data set. The keywords with their proper context will be compared to these dictionaries and the given post will be categorized into positive or negative or neutral category.

C. Determining Intensity of Negative Posts:

The further state for a positive post will be 'happy' category of sentiment. For negative posts the intensity may vary, which will be main concern. The intensity will be categorized into three levels 'Low', 'Medium' or 'High'. The intensity levels can be called 'Sad', 'Prone to depression', 'Depressed'. K-means can be used for this purpose.

For performing k means clustering on text data the words will be assigned with an intensity value according to the dictionary meaning of the word. Consider bad with intensity 10, so worse will have intensity 20 and worst will have intensity 30. Higher the impact of negativity in a word higher will be the intensity value.

Select the value of k as 3, randomly select 3 centroids for k = 3 such that these centroids are away from each other as small, medium and high intensity.

According to the intensity (numeric value) of the words, a particular word will fall into any one cluster which has minimum intensity difference.

All the words will be clustered and centroid will be rearranged by calculating the average at each step. Value for each cluster

will be calculated as multiplication of value of centroid and no of words into that centroid.
Ex. Value of cluster1 = Value of centroid of cluster 1 * No of words in that cluster.

We will now get value of 3 clusters. Calculate the average value and this value is the intensity of a particular post. According to the intensity of a particular post calculated above, the post will be classified into low, medium and high intensity.

If the post is of low intensity of negativity then we conclude the person is sad. If the post is of medium intensity we conclude the person is lonely and if the post is of high intensity the we conclude that the person is highly depressed.

D. Providing Recommendations According to the Class And Intensity of Post:

For 'happy' category the user will be guided towards even more positivity.

For 'Sad' category the user may be provided by positive and motivational thoughts and posts. This action will continue for other two categories 'Prone to depression' and 'Depressed'. In 'Prone to depression' category even closer observation of the user will be practiced. Certain threshold values will be set on the basis of time and frequency of negative posts of medium or high intensity. For example, if the series of negative posts continue for more than two weeks the person may be prone to depression. In extreme cases of depression, the actions like notifying the close people of the user may be taken.

The entire process takes place on client side.

IV. RESULTS

The purpose of this research is to design an algorithm that can efficiently compute the sentiment of data coming from social media and provide necessary recommendations to the users.

V. CONCLUSION AND FUTURE SCOPE

This system recognizes the psychological patterns of the social media users, using posts from social media and analysis the intensity of the emotion. Recommendation based on the intensity of sentiment analysis, is provided to the user in the form of motivational quotes and images. This system also recognizes the extreme level of depression and chances of a user to commit a suicide and takes the necessary action. The system works on client side.

VI. CHALLENGES

- Identification of bipolar and sarcastic sentences.
- Judgement of context.
- Meaning extraction of words with same spelling but different meanings. Achieving accuracy because of ambiguities.

REFERENCES

[1]. R.K.Bakshi, Navneet Kaur, Ravneet Kaur, Gurpeet Kaur, "Opinion mining sentiment analysis", New Delhi, INDIA, 2016,IEEE.

[2]. S.Huang, J.Zhang, L.wang, X.S.Hua, "Social friend recommendation based on multiple network correlation",2015,IEEE.

[3]. A.P.Jain, V.D.Katkar, "Sentiments analysis of twitter data using data mining", Pune, INDIA, 2015, IEEE.

[4]. M.kanakraj, R.M. Reddy, "NLP based sentiment analysis on twitter data using Ensemble classifiers", Manglore, INDIA, 2015, IEEE .

[5]. L.Povoda, R.Burget, M.K.Dutta, "Sentiment analysis based on support vector machine and big data", 2016, IEEE. [6]. S.Liu, X.Cheng, F.Li, F.Li, "TASC: topic adaptive sentiment classification on dynamic tweets", 2013, IEEE.

[6]. Malhar Anjaria, Ram Mahana Reddy Guddeti, "Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning", Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014 IEEE.

[7]. S. Gao and H. Li, "A cross-domain adaptation method for sentiment classification using probabilistic latent analysis," in Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011, pp. 1047–1052.

[8]. Y. Mejova and P. Srinivasan, "Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter." In *ICWSM*, 2012.

[9]. J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: Recommending people on social networking sites," in ACM CHI'09, New York, NY, USA, April 2009.

[10]. E. Zhong, W. Fan, and Q. Yang, "User behavior learning and transfer in composite social networks," ACM Transactions on Knowledge Discovery from Data, vol. 8, February 2014.