

ANALYSING THE SOCIAL DATA OPINION USING SENTIMENT SENSITIVE EMBEDDINGS

J.Sharmila,Saranya,Sri Nivethini,Swathi

Asst. Prof,UG Scholar

Department of Computer Science and Engineering
Sri Krishna College of Engineering and Technology, Coimbatore.
sharmilaj@skcet.ac.in ,13bc156@skcet.ac.in

Abstract

This project is entitled as “Analysing the Social Data Opinion using Sentiment Sensitive Embeddings”. The main objective of this project is to generate a real time suggestion from user’s comments in various social networks. Numerous comments are being posted in various social media daily, making it a valuable platform for tracking and analyzing public sentiment. This project is used to arrive at a conclusion from these comments in a quick and timely manner. By Using Artificial Neural Network (ANN) and Text categorization a Bag of words is generated. The resulting words will be used to analyze the public sentiment variations by micro analysis method which is called Pattern recognition. Two novel generative models like Clustering and Classification, which contains an inbuilt data dictionary are developed to solve the reason mining problem. Ranking is used for the estimation of relevance in posted comments to obtain a more specific outcome. All kind of user sentiments will be scrutinized from hitting comments whereas non-hitting sentences will be considered as moderate comment. An overall report in a pictorial format will be generated for each post from the given comments which can be used for supernumerary works.

Keywords— Sentimental analysis, data dictionary, opinion mining, pattern recognition.

I. INTRODUCTION

Analysis of user comments in social media is a challenging and a critical task. There are thousands of comments for a post which has to be analyzed to understand the human opinion on a

large scale. This helps in the generation of overall report concerning a product or an event. A timely

and an efficient output can be obtained through various techniques

used in this project, which helps many business organizations and government sectors to improve the success rate by easily understanding the changes to be incorporated in their products and schemes respectively.

Data mining for software engineering consists of collecting necessary data, extracting some knowledge from it and using this knowledge to improve the software engineering process, in other words “operationalize” the mined knowledge.

II. DATA MINING

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning statistics and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations complexity considerations, post processing of discovered structures, visualization, and online updating.

Data mining involves six common classes of tasks:

i. Anomaly detection – The identification of unusual data records, that might be interesting or data errors that require further investigation.

ii. Association rule learning (Dependency modeling) – Searches for relationships between

variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

iii. Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

iv. Classification is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

v. Regression attempts to find a function which models the data with the least error.

vi. Summarization providing a more compact representation of the data set, including visualization and report generation.

III. SENTIMENT ANALYSIS

It is a type of natural language processing (NLP) for tracking the mood of the public. It often uses machine learning, a type of Artificial Intelligence (AI), to mine text for sentiment. Opinion mining can be used to evaluate the success of an ad or a product launch which demographics the like or dislike of a product feature. This technique is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service and it aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. A different method for determining sentiment is the use of a scaling system whereby words commonly associated with having a negative, neutral or positive sentiment with them are given an associated number (most negative up to most positive). This makes it possible to adjust the sentiment of a given term relative to its environment.

Existing approaches to sentiment analysis can be grouped into three main categories: knowledge-based techniques, statistical methods, and hybrid approaches. Knowledge-based techniques classify text by affect categories based on the presence of unambiguous affect words such

as happy, sad, afraid, and bored. Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and Semantic Orientation (Point wise Mutual Information). Hybrid approaches leverage on both machine learning and elements from knowledge representation such as ontologies and semantic networks in order to detect semantics that are expressed in a subtle manner, e.g., through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts. For example the review on a website about a product may be broadly positive which can also be specifically negative. Being able to identify this kind of information in a systematic way gives the vendor a much clearer picture of public opinion than surveys or focus groups do, because the data is created by the customer.

IV. PROPOSED SYSTEM

The project is designed in such a way to generate real time suggestion for user's comments in social networks. To analyze the user's comments, Pattern recognition and data dictionary has been implemented.

The techniques used are,

i. Data dictionary contains group of all sentimental words and this contains a data set for pattern matching. The data dictionary contains definitions for fields as used in mining models. It specifies the types and value ranges. These definitions are assumed to be independent of specific data sets as used for training or scoring a specific model. A data dictionary can be shared by multiple models, statistics and other information related to the training set stored within a model.

ii. Pattern recognition classifies the data based on statistical information extracted from the patterns. It is a branch of machine learning that focuses on the recognition of patterns and regularities in data, although it is in some cases considered to be nearly synonymous with machine learning. Pattern recognition systems are in many cases trained from labeled "training" data (supervised learning), but when no labeled data are available other

algorithms can be used to discover previously unknown patterns (unsupervised learning). This generates pattern for all possible inputs and to perform the most likely matching of the inputs, taking their statistical variation. Through pattern recognition method the user's comments will be categorized and the given input will be separated into various sentiments.

iii. Clustering is a process of partitioning a set of data into a set of meaningful sub-classes called clusters and aggregate them based on their similarities. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

iv. Classification is the process of organizing data into categories for its effective and efficient use. The goal of classification is to accurately predict the target class for each case in the data. Classifications are discrete and do not imply order. A classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. Classification models are tested by comparing the predicted values to known target values in a set of test data. A product will include a positive and a negative opinion. The word which is positive in one situation may be considered as negative in another situation. Example the word "long" may be considered as positive in case of battery life and negative in case of start-up time. Thus the opinion systems which gather opinions will not be same for all the products.

v. Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: -100), impossible data combinations, missing values, etc. Analyzing data that has not been

carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. The product of data pre-processing is the final training set.

vi. Ranking can be calculated using global patterning report. Ranking is determined by using the most hit words in the comments. This technique helps to come up with more specific details related to human opinions and generate precise charts. Through the generated chart, the admin can get a clarity report for the user's comments.

V. FUNCTIONING OF THE SYSTEM

There are two inputs to the system-Raw dataset and trained data. A raw dataset has been used as the input which contains Posted by, Date, Time, Posted topic, User name, and User's comments. This is the backbone of the project as the entire project depends on this information. Using Data Dictionary, admin can create a data set for pattern matching. Data dictionary contains group of all sentimental words. This process will be done by the admin. Sentimental words, Positive and negative words can be updated by the admin.

In order to eliminate redundancy in comments like repetition of comments by the same person, posting the comments of other users, preprocessing technique is applied.

Text categorization is performed to split up the user's comment into various parts like noun, verb, object, adjective etc and put them in array. This is performed by comparison with trained data to obtain the required words for sentimental analysis. Through pattern recognition method the user's comments will be categorized. During this process, the given input will be separated into various sentiments like positive and negative. Moderate comments will be the non-hit words from all the sentimental classifications.

After this process, the individually recognized comments are categorized into various groups based on the pattern they follow by using

clustering and classification methods. This helps to generate chart. All the comment will be grouped using global patterning reports and various charts will be generated. Instead of reading all the comments commented by the users, chart will represent the graphical information of the effectiveness of the discussion. Then Ranking is performed to obtain more detailed outcome. Ranking is based on most hit words in the comments. Through the generated chart, a clear report about the post can be obtained.

VI. BENEFITS OF SENTIMENT ANALYSIS

Some of the benefits of sentiment analysis are as follows,

1) Improve Customer Service

Sentiment Analysis gives useful insights about the current and future customers' purchase preferences, brand affiliations, topics of interests, opinions, point of views on discussions, likes and dislikes in products/ services and much more. This useful information let organizations to drastically improve their customer service and engagement strategies by building on the positive sentiments and formulating methods to combat negative sentiments.

2) Revive Brand

One of the best uses of Sentiment Analysis is that it allows organizations to quantify perceptions about brands, products and services, marketing campaigns, social engagement initiatives, online content etc. Organizations can use this information for devising better and more effective branding and marketing strategies and thus improve brand reputation.

3) Beat Competition

Sentiment Analysis allows organizations to know sentiments surrounding the competitors too. This allows to benchmark the performance against that of their competitors. Using the reported sentiments, one can also predict trends and develop one's specific social strategies to leverage these trends.

4) Gain Business Intelligence

Sentiment Analysis empowers organizations by providing extensive, insightful information regarding their target audience's

sentiments. Made use of correctly, these sentiments are a gold mine of newer business possibilities and opportunities. Thus, Sentiment Analysis provides an insightful business intelligence using which one can take impactful decisions that would leverage your business.

VII. CONCLUSION AND FUTURE WORK

This project plays a vital role in decision making sector of business organization that would fetch more income or higher success rate using freely available information on the social network with the help of sensitive sentiment analysis process. In situations where the customers cannot be directly reached, this project would be the best option to rely on, in order to understand human opinion.

Inductive learning algorithms have been suggested as alternatives to knowledge acquisition for expert systems. However, the application of machine learning algorithms often involves a number of subsidiary tasks to be performed as well as algorithm execution itself. It is important to help the domain expert manipulate his or her data so they are suitable for a specific algorithm, and subsequently to assess the algorithm results. These activities are often called pre processing and post processing.

The future enhancement discusses issues related to the application of the text categorization algorithm, an important representative of the inductive learning family. A prototype workbench which has been developed to provide an integrated approach to the application of text categorization is presented.

- Can implement for web based application
- Handshakes with inductive learning algorithm
- Improvisations can be done in the performance evaluation
- In case of huge range of data set, data load balancing can be done

VIII. REFERENCES

- [1] "Cross-domain Sentiment Classification using Sentiment Sensitive Embeddings",

- DanushkaBollegala Member, IEEE, Tingting Mu Member, IEEE, JohnY. Goulermas Senior Member, IEEE,IEEE Transactions on Knowledge and Data Engineering (Vol:28 , Issue: 2) February 2016.
- [2] “A Survey on Sentiment Analysis Algorithms for Opinion Mining”,
Vidisha M. Pradhan, Jay Vala,PremBalani,International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016.
- [3] “Feature selection and classification approach for sentiment analysis”,GautamiTripathi , Naganna S(Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.2, June 2015).
- [4] “Sentiment Analysis of Twitter Data”,Apoorv Agarwal BoyiXie Iliia Vovsha Owen Rambow Rebecca Passonneau, Department of Computer Science, Columbia University, New York, NY 10027 USA,2014.
- [5] “Crowd sourcing Predictors of Behavioral Outcomes”
Josh C. Bongard, Member, IEEE, Paul D. H. Hines, Member, IEEE, Dylan Conger, Peter Hurd, and Zhenyu Lu. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2013.
- [6] “Automatic Sentiment Analysis for Unstructured Data”,
Jalaj S. Modha,Prof& Head Gayatri S. Pandi,Sandip J. Modha (Volume 3, Issue 12, December 2013).
- [7] “Modelling public mood and emotion: Twitter sentiment and socio-economic phenomena”,J. Bollen, H. Mao, and A. Pepe inProc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
- [8] “Sentiment-oriented contextual advertising”,T.-K.FanandC.-H.Chang, Knowledge and Information Systems, vol.23,no.3,pp. 321–344,2010.

IJSER