Sentiment analysis in social media Using machine learning techniques with R language

T. Praveen*, Rohit Kumar Sharma, Gurdeep Singh, Ram Shankar
[1]Department of Computer Science and Engineering,
Veltech Hightech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi Chennai 600062
***Corresponding author: E-Mail: praveen@velhightech.com**

ABSTRACT

Social networks have grown the way in which people communicate. Information Available from social media is beneficial for analysis of user opinion, for example measuring the review on a recently released product, looking at the response to policy change or the enjoyment of an ongoing event or about a product or a person. Manually observing and analyzing through this data is difficult and potentially expensive. Analyzing the Sentiment is a new area, involves getting user (twitter user) opinion automatically from the twitter or any social media. Analysis of Sentiments is a task to identify a text as comments, reviews or message. Here to implement an algorithm for automatic classification of text into positive, negative, neutral, or negation there are many ways in which social network data can be analyzed to give a better understanding of user opinion such problems are at the heart of natural Language processing (NLP) and data mining research. In this paper, we present a tool for sentiment analysis which is able to analysis Twitter data using Twitter stream API. Using this technique, we build a sentiment classifier that is able to determine positive, negative and objective sentiments for a document. Sentiment analysis of Twitter data. Sentiment or utilizes the naive Bayes Classifier to classify Tweets into positive, negative neutral, or negation We present experimental evaluation of our Live Review Twitter dataset and classification results, Sentiment Analysis is a task to identify a text as comments, reviews or message.

KEY WORDS: Sentiment analysis, Live Review, natural language processing, data mining, sentiment classifier.

1. INTRODUCTION

Data mining: Data mining is the extraction of information from large databases, it is emerging technology with great potential which helps the companies to work only on important information in their data warehouses. Data mining tools used to predict about new or existing trends and behaviors or point of view, allows the businesses to make proactive, knowledge-driven decisions. It can answer the business questions that traditionally time consuming to resolve the result. It can store large data into database and bring out effective tools to relate and save that data for future enhancement or retrieval only needed or relative information.

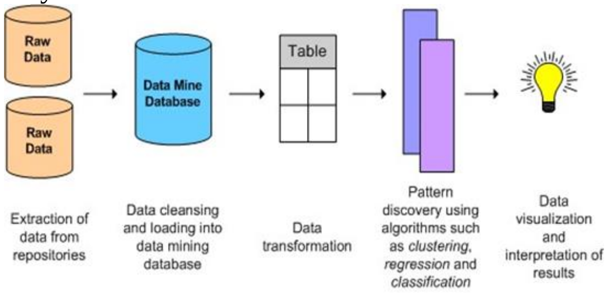

Raw Data
Raw Data
Data Mine Database
Table
Extraction of data from repositories
Data cleansing and loading into data mining database
Data transformation
Pattern discovery using algorithms such as *clustering, regression* and *classification*
Data visualization and interpretation of results

Figure.1. Data Mining

Data mining takes this process beyond retrospective data access and navigation to proactive information delivery. Data mining is application in business, because it supports the three technologies and they are:

- Massive data collection. Data mining algorithms.
- Powerful multiprocessor computers.

The data warehouses that integrate operational data with customers, market information, suppliers have resulted in explosion of information. In our project, it shows how we can analyses and bring out a result on a particular topic. First of all, we work on data that has been given by the user in social media, specially working with twitter, the user (twitter user) can post any data like about product, person, company, etc.…,

Hence, we work in our project to fetch this data and work on it to produce a report as result, either it is positive, negative, neutral negation. First, we fetch the data from the twitter given by the user and we do it by connecting it by ‘Twitter Stream API’. Stream R package allows users to fetch twitter Data in real time by connecting to Twitter Stream API.

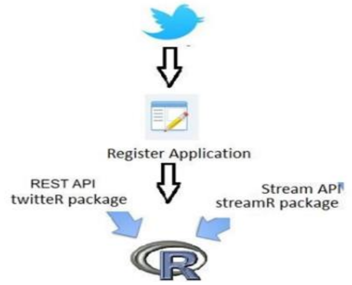


Figure.2. Streaming API

By this we can fetch the data and then we work on data as we broke it into words, phrases, symbols, or other meaningful elements called ‘tokens’. Hence, we broke the data in this form so we can compare this data with our data sets and bring on a result. We use the ‘twitter stream API’ as in [Figure 1] to connect to the Twitter so we can fetch the data and store it in ‘stop words’ and the compare the words with data sets to find its sentiment towards live review.

So here we work on fetching the data and comparing it with the data sets we create. We use ‘naïve Bayes classifier’ to parallel processing of fetching and comparing finally coming on a result about how the data is positive or negative or neutral in nature.

For example: If we run an organization and we have many products launched in our experienced period and finally new product has been launched by the company and eagerly waiting for the review of people. We can’t just post the details of product and read all the millions of comments and come across a solution either it was positive or negative. Hence our project works on that data that has been given in form of comment as a review. Hence it will collect all the data and broke that into works and analyses and finally produce a report as it is positive, negative, neutral or negation.

Problem Definition: Given a message, classify whether the message is of positive, negative, or neutral sentiment. Then for messages conveying both a negative and positive sentiment, whichever is the relative sentiment should be chosen. The problem in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level. Searching problem. Tokenization and classification. Reliable Content Identification. Here, as we mentioned we work with ‘naïve Bayes classifier’, it gives us parallel processing and by this we overcome the existing system.

$$\underset{\text{Posterior Probability}}{P(c\mid x)}=\frac{\overset{\text{Likelihood}}{P(x\mid c)}\,\overset{\text{Class Prior Probability}}{P(c)}}{\underset{\text{Predictor Prior Probability}}{P(x)}}$$

$$P(c\mid \mathrm{X}) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

Eqn.1. Bayes rule

The searching problem has been reduced as we are searching the data and fetching it by twitter stream API and then storing it in stop words to compare at a same time to analyze the sentiment in it. Tokenization is the process used to overcome this problem and makes this process reliable and efficient. It splits the sentence into words or phrases to compare it with stop words data.

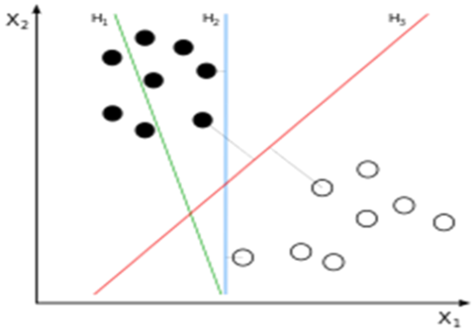
Advantages:

- Model is easy to interpret.
- Can be Domain-Specific.
- Can be more Robust.
- Efficient computation.

Related Work:

Support Vector Machine(SVM): In the machine learning, the support vector machine(SVM) shown in also supported by (support vector networks), are an supervised learning models with associated learning models with associated learning algorithms that will analyses the data used for the classification (by classifier) and analysis.

By doing the linear classifications’ can be efficiently perform a non-linear classification using the kernel trick, implicitly mapping to their inputs into high-dimensional features spaces.



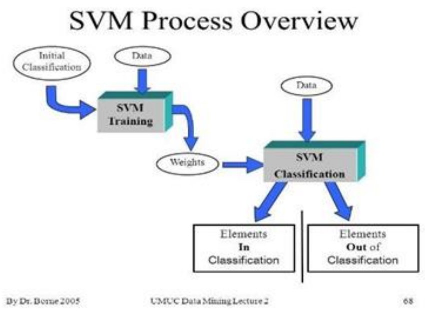


Figure.3. SVM **Figure.4. SVM Overview**

Linear and Non-linear data: Since the relationships are not linear (unstructured), we could not show the relationships using a linear data structure. The linear and non-linear data has been shown in [Figure 4].We cannot show relationships like list or stack. But, we needed something that looks more like tree. A tree is just an example of a non-linear data structure. some other examples of non-linear data structures are multidimensional arrays and graphs.

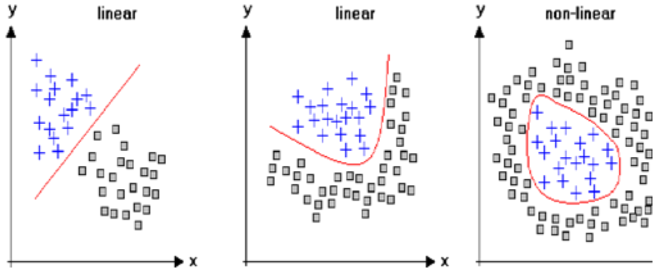


Figure.5. Linear and non-linear

As we mentioned we use Support Vector Machine in our existing system, where it is a serial process only do sequential process and it either work to do fetching the whole data and then comparing with the data sets at a time. It can do only single work at a time and cannot perform the parallel processing. Hence using the naïve Bayes classifier for parallel processing in our proposed system. In this system using the linear and nonlinear data is not efficient as it may not structure, hence it increases the work load to work to separate the data and make it structured for the further process while naïve Bayes classifier gets the data and process at a time to compare and produce result. Hence it is more efficient and reliable.

Advantages of existing system:

Easy to interpret: The model is easy to interpret, as it adapts to every situation. It gets the data sets from Twitter and interpret to change into structured format to classify its sentiment.

Domain-Specific: This is domain specific as it works to achieve full result and its all elements works efficiently to work to achieve goal.

Robust: It is strong in its algorithms and methods.

Efficient computation: We work to make the process parallel and efficiently handle the sentence data of large amount. Hence its efficient computation is our key concept.

Literature Survey:

Personalized Recommendation Combining User Interest and Social Circle: The personal interest denotes user's individual view on items, especially for our experienced customers and regular bloggers on reviews and this made this technique so unique. We conducted extensive experiments on three large real-world social rating datasets, and showed significant improvements over existing approaches that use mixed social network information. At present, the personal recommendation model only takes user rating records and interpersonal relationship of social network into consideration.

Personalized Recommendation Based on Ratings and Reviews Alleviating the Sparsity Problem of Collaborative Filtering: Personalized recommendation algorithm integrating users' reviews and ratings into a unified model TMCF. In TMCF, users' reviews are used to generate reviews' topic allocations and users' preferences. A new metric is proposed to calculate users' similarity based on users' most valued features. Finally, collaborative ratings are also utilized to make the final recommendations. Experiments on seven data sets of different domains show our model can achieve better recommendations than traditional CF when target users rate and review only few items and data sets are extremely sparse. Moreover, TMCF makes better recommendations than the state-of-the-art topic model based collaborative filtering algorithm on all data sets.

A Matrix Factorization Technique with Trust Propagation for Recommendation in Social Networks: Recommender systems are emerging as tools of choice to select the online information relevant to a given user. With the advent of online social networks, exploiting the information hidden in the social network to predict the behavior of users has become very important.

Social Contextual Recommendation: Conducted extensive experiment on large real world social network datasets, and showed that social contextual data can boost the performance of recommendation on these social data. In particular, we have gained increases of 24.2% and 20.7% in prediction accuracy and 21.7% and 12.3% in recommendation Precision upon previous approaches on these social networks, respectively. Also, the used algorithm is general and can be easy to adapt according to different real-world recommendation scenarios.

Circle-based Recommendation in Social Networks: The Recommendation accuracy by bringing up the concept of inferred circles of friends". The idea proposed is to find the best subset of a user's friends that is an inferred circle, for making recommendations in an item category of interest, they may die from their explicit circles of friends that have become popular in social networks. We proposed a way for inferring category-specific circle, and to assign weights to the friends within each circle. In our experiments on publicly available data, we showed significant improvements over existing approaches that use mixed social network information.

System Architecture: The system architecture shows how these steps being carried out from beginning to end. It gives the data flow. First it begin with the raw data where the data is fetched from the twitter, hence it collects the data in raw form. Then it goes for tokenizer were the data is splitted into small elements like words or phrases or other meaning full format.

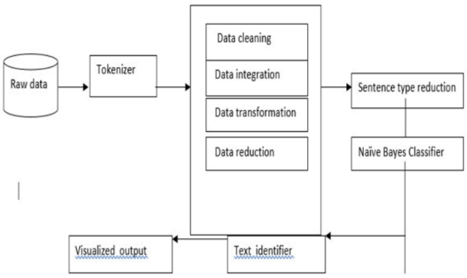


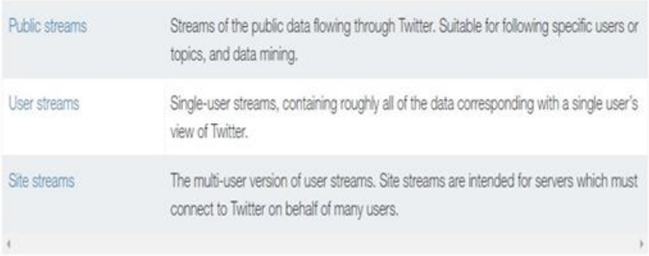
Figure.6. Architecture

Hence it goes for 3rd step into data preprocessing where the data goes through four parts of techniques namely data cleaning, data integration, data transformation, data reduction.by this the data gets very well structured according to the procedure which makes the further process very easy and efficient. Hence here the data is transformed to understandable format, now the sentence type reduction occur, the sentence is reduced to Declarative, Imperative, Interrogative Sentence. Now the naïve Bayes classifier is applied which makes this data broken into smaller parts into simple classification into words. Then the text is finally identified and come across the result that whether the result is positive or negative or neutral or negation. Hence the visualized output will be created and given.

Modules:

Fetching: In this project, we have to fetch the raw data from the twitter do to their analysis using R language. Stream R package allows users to fetch twitter Data in real time by connecting to Twitter Stream API. For implementing the project there is a need to interface all the comments that are commented in twitter for any particular feed. To interface the comments data or to get the data from twitter for sentimental analysis process we need to have full access to the commented data. For that we are requesting the twitter server to have access to twitter data. Twitter provides an API that helps other party application to have access to the twitter data using twitter stream API. The Streaming APIs provide developers low latency access to Twitter's global stream of Tweet data. The use of a streaming client will get pushed messages indicating Tweets and other events have occurred.

Twitter offers several streaming endpoints which are shown below



Public streams	Streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining.
User streams	Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter.
Site streams	The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users.

Figure.7. Twitter Streaming Endpoints

After the overall process of retrieving the data from twitter comments for sentimental analysis using twitter Streaming API. The next process is to analyze the collected data and further classification of data for further process is done in Tokenizing process explained in next module.

Tokenizing: In sentimental analysis, tokenization is defined as the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. All the tokens are taken for further processing such as text mining or parsing. Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of sentimental analysis. A *tokenizer* receives a set or stream of characters, breaks it up into small *tokens* (usually individual words), and outputs a stream of *tokens*. The tokenizer is responsible for recording the position or order of each term (used for phrase and word proximity queries) and the start and end *character offsets* of the original word which the term represents (used for highlighting search snippets). The tokenizing process is done with the help of some predefined data sets. The first one is step words data sets. Then there are two other data sets which defined the sentiments of the given words. The other two data sets are positive data sets which contains set of words which sentiments good or positive view. And the other data sets contains negative data sets which sentiments bad or negative view. The tokenizer scans the collected data classify based on the predefined data sets and convert the sentence into different tokens to ease the further process.

Data Pre-processing: Data preprocessing is a data-mining technique that transforms raw data into an understandable format. Real-world data is often inconsistent, incomplete and/or lacking in certain behaviors or trends, and it is likely to contain many errors. Data preprocessing is a proven best method of resolving such issues. Data preprocessing prepares raw data to process further. Data preprocessing describes every type of processing performed on raw data to prepare it for further processing procedure. Commonly used for preliminary data mining practice, data preprocessing transforms all the data into a format that will be more easily and effectively processed for the purpose of the user -- for example, in a neural network. There are large number of different tools and methods used for preprocessing, including: sampling, which is used to select a representative subset from a big amount of data; transformation, which transforms raw data to produce a single input; denoising, which remove all noises from data; normalization, which organizes all the data to have more efficient access, which pulls out data that is significant in some particular context. Data preprocessing is used for database-driven applications such as customer relationship management and rule-based applications.

Data goes through a series of steps during preprocessing:

- Data Cleaning: Data is cleansed with the processes such as filling in missing values, smoothing the data with noise, or resolving the inconsistencies in the data.
- Data Integration: Data with different type of representations are put together and conflicts within the data are resolved.
- Data Transformation: Data is simplified, aggregated and generalized.
- Data Reduction: This step mainly aims to present a reduced representation of the data in a data warehouse.

The tokens are processed by DATA PRE-PROCESSING i.e. Data cleaning, Data Integrity, Data Transformation and Reduction is carried out to an understandable format this is then implied to an identifier for identifying whether the given datas are positive, negative, neutral, or negation.

Sentence Detection: After completing the data preprocessing the data are transformed to the understandable format. Which gives the correct identification and accurate meaning of the data. And also it reduce the sentence as declarative, Imperative, Interrogative Sentence. In the sentence detection module the pre-processed data is evaluated and identify the exact meaning a sentence is referring to. It uses the data sets and with that it combines the step words and the words which are referring to sentiments to find out the exact meaning the particular phrase is referring to. For example the phrase "very good" in this it has two parts the first part represents the step word and the second part refers to the sentiment and when they both combine they give the exact meaning of phrase. So the sentence detection is a important module in sentimental analysis process because without it the sentence may not give out the exact sentiment the phrase is trying to express.

Classifying the Text: In this project the data are classifying using the naïve Bayes classifier. The naïve Bayes classifier is used for machine learning process and it classifies the data at particular manner such as positive or negative data etc. Naive Bayes is a very simplified classification process that makes some strong assumptions about the independence of each input variable. The Bayesian Classification represents a supervised learning method and also a statistical method for classification. Assumes an underlying probabilistic model and also allows us to capture the uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve predictive and diagnostic problems. This Classification is named after Thomas Bayes (1702-1761), who proposed this law Bayes Theorem. Bayesian classification gives practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides very useful perspective for understanding and finding many no. of learning algorithms. It evaluates explicit probabilities for hypothesis and it is robust to noise in input data.

Its easy to build and particularly used for very large datasets along with simplicity. Naïve Bayes is known to perform even highly sophisticated classification methods. Provides,

$p(^{c})=p(^{x})p(c)/p(x)$

xc

Simple classification of words based on Bayes Theorem. It is a Bag of words approach for Subjective Analysis of a content.

The classification of word is the important and main phrase as here only the sentence is actually categorized as whether it is a positive, negative, or neutral sentence.

Visualizing the Data: It produces the output of the classifying the data such as Positive, negative, neutral and negation are estimated from the twitter.

2. CONCLUSION

In this paper, a recommendation model is proposed by mining sentiment information from social users' reviews. We fuse user sentiment similarity, interpersonal sentiment influence, and item reputation similarity into a unified matrix factorization framework to achieve the rating prediction task. We conclude that using Naïve Baye's Classifier it is easier to classify the tweets and more we improve the training data set more we can get accurate results. Sentiment analysis of Twitter data. Sentiment or utilizes the naive Bayes Classifier to classify Tweets into positive, negative neutral, or negation We present experimental evaluation of our Live Review Twitter dataset and classification results.

Future Enhancement: In future we can further improve the classification process by adding some new features also. Like Now a day's, whenever retrieving Texts from the search Engines that retrieves Texts without analyzing their content, simply by matching user queries against the Text's filename and format, user-annotated tags, captions, and, generally, text surrounding the Text. Also the retrieved Text does not contain any textual data along with the Texts. We introduced the task of automatic caption generation for news Texts. The task fuses with computer vision and natural language processing and also holds promise for various multimedia applications, such as Text retrieval, development of tools supporting news media management, and for individuals with visual impairment. There is a way to learn a caption generation model from labeled data without costly our involvement. Instead of manually creating annotations, Text captions are treated as labels for the Text. Although the caption words are admittedly noisy comparison to traditional human created keywords, it show that it can be used to learn the correspondences between textual and visual modalities, and also serve to a gold standard for the task of caption generation. We have presented extractive and abstractive caption generation models. A key aspect of this approach is to allow both the visual and textual modalities to influence the generation task.

REFERENCES

Kim S.M, Hovy E, Determining the sentiment of opinions. In: Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, 2004.

Lin Y, Zhang J, Wang X, Zhou A, An theoretic approach to sentiment polarity classification. In Proceedings of the 2Nd Joint WICOW/AIR Web Workshop on Web Quality, Web Quality '12. ACM, New York, 2012, 35–40.

Liu B, Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, Second Edition, Taylor and Francis Group, Boca, 2010.

Sarvabhotla K, Pingali P, Varma V, Sentiment classification: a lexical similarity based approach to extract subjectivity in documents. Inf Retrieval 14 (3), 2011, 337–353

Xiaojiang Lei, Xueming Qian, Rating Prediction based on Social Sentiment, 2016.