ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com
Journal DOI: 10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

A REVIEW ON SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA USING TEXT MINING AND MACHINE LEARNING.

GURPREET KAUR[1] and MANOJ KUMAR[2].

1. M.Tech. Student (Computer Engg.), Yadavindra College of Engineering, Punjabi Univ. Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India.
2. Assistant Professor (Computer Engg.), Yadavindra College of Engineering, Punjabi Univ. Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India.

Manuscript Info

Manuscript History:

Received: 15 March 2016
Final Accepted: 22 April 2016
Published Online: May 2016

Key words:
KNN, Naive Bayes, SVM, twitter, facebook.

Corresponding Author

GURPREET KAUR.

Abstract

This paper proposes to utilize this source of information and predict the sentiments of public towards a particular topic. Twitter data is utilized for the same and live tweets of Indian origin are extracted using twitter API called 'tweepy' Twitter API was used for streaming of tweets. Entire data can be obtained through access token and secret key. The score for every tweet was evaluated by using one of the score based approach. Moreover, KNN algorithm will be applied on the tweets.

Introduction:-

In recent years, due to the popularity of social networking has dramatically increased and the vast amount of data being produced by social networks such as Twitter, Facebook, Google+, etc.,. Social networks become popular among millions of people who share's their thoughts in everyday life. Social media web sites are rich source of data for sentiment analysis. Sentiment Analysis has been used to understand the people's opinion on particular product or service. Twitter, one of the biggest and most popular social website which contains unstructured data. Figure 1shows how the sentiment analysis works.

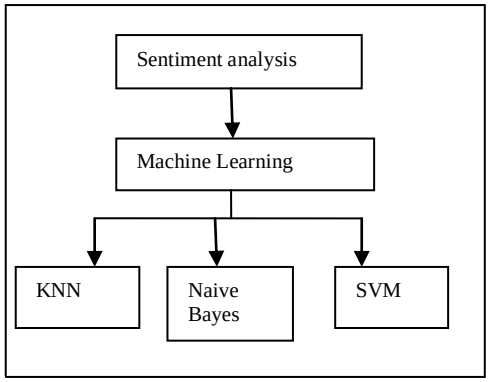

Figure: 1 Sentiment Analysis Processing

Social media is a great medium for exploring developments which matter most to a broad audience and it is the means of interactions among people in which they create, share, and exchange information and ideas in virtual communities and networks. Social media technologies take on many different forms including magazines, Internet

forums, weblogs, social blogs, micro blogging, wiki, social network, podcasts, photographs or pictures, video, rating and social bookmarking. Micro blogging websites have evolved to become source of varied kind of information. The use of social media is increasing day by day. Increasing growth of social media users over internet has also increased their participation in various discussions and activities simultaneously. Manually reading such a bulk amount reviews is a very difficult task. So there is a need of a automatic system which will lead to automatically extract the positive and negative features of the product and make the decision making process easier. There are many sites and companies which perform these activities.

There are various approaches to design machine learning algorithms. The purpose of ML algorithms is to use observations as input and this observation can be a data, pattern and past experience. Thus Ml algorithms use to improve the performance of instances, which can be done by any classifier by trying to classify the input pattern into set of categories or to cluster unknown instances. As the nature of ML algorithms it enhances its performance from past experience or by receiving feedback. It can be divided into two categories supervised and unsupervised approach. Supervised: In supervised learning, the instances are labeled with known or target classes labels. Here before classification the dataset knows the target class. Thus it is very helpful for the problems which have known inputs. Unsupervised: In unsupervised learning, the algorithm groups the instances by their similarities in values of features and makes different clusters. In it no prior class or clusters are given, the algorithm itself defines their clusters automatically and statistically. Sentiment analysis is a natural language processing and information extraction task.

This technique aims to extract writer's feelings expressed in comments or reviews. Sentiment analysis does not only deal with extracting polarity but also deals with extracting features from the text. Sentiment analysis is also about finding subjectivity or objectivity of the opinion. There are many applications of sentiment analysis. The one of the main applications of sentiment analysis is in business and government intelligence. Business intelligence seems to be one of the main factors behind corporate interest in the field. Sentiment analysis is also known as opinion mining. Sentiment analysis is a natural language processing and information extraction task that aims to obtain writers feelings expressed in positive or negative comments, questions by analyzing a large number of documents. Sentiment analysis is considered as a classification process. There are main three classification levels in sentiment analysis: Document Level, Sentence Level, and Aspect Level.

Related work:-

Marketa et al. [1] proposed a model that collects tweets from social networking sites and then gives their own opinion of business intelligence. In this framework, there are two layers in the sentiment analysis tool, the data processing layer and sentiment analysis layer. Data processing layer deals with data collection and data mining, while sentiment analysis layer use a application to present the result of data mining. Mahalakshmi R and Suheelapropose a method of sentiment analysis in this paper which is based on twitter by using Hadoop and its ecosystems. The proposed method will process the large volume of data on a Hadoop and the mapReduce function will perform the sentiment analysis.

Ortiga et al. [2] presented a new approach for sentiment analysis in social site (facebook). This analysis is starting from the message written by its user. The results obtained through this proposed scheme shows that it is possible to execute sentiment analysis in Facebook easily with high accuracy of 83.2 percent.

According to Alexandra et al.[3] identified three subtasks that need to be addressed: defining the target; separating the bad and good news content from the bad and good sentiment expressed; and finally analysis of clearly mentioned opinion that is expressed unambiguously, not needing understanding or the utilization of world knowledge. Furthermore, distinguish three dissimilar views on newspaper articles (text, author and reader), which have to be handle differently while analyzing sentiment.

According to Gupta et al. [4] proposed Sentiment analysis which aims to investigate the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (the emotional state of the author when writing), or the intended emotional communication (the emotional effect the author wishes to have on the reader.

Aisopos et al. [5] presented a method which is based on two orthogonal and complementary sources of evidence. One of the evidence is context-based method captured by polarity ratio and other evidence is content-based features

obtained with n-gram graphs. These two methods are language-neutral and tolerant to noise. The performance shows that these proposed method assured high robustness and effectiveness. Furthermore, some other group of features having low extraction cost develop dimensionality reduction techniques and discretization techniques.

In this paper, Agarwal and Apoorv [7] proposed micro blog named as Twitter which is very popular and build models to classifying the "tweets" into positive and negative sentiment or they can be neutral. author worked with micro blog data named as Twitter and build models for classifying the "tweets" into positive negative and neutral sentiments. Author builds models for classification tasks: first one is binary task of classifying sentiments of users into positive and negative classes and second is a 3-way task of classifying sentiments of users into positive negative and neutral. Author experiments with three types of models: (1) unigram model (2) feature based model (3) tree kernel based model. Results from paper indicate that unigram model is a hard baseline. Our feature based model that uses only 100 features achieves similar accuracy as unigram model that uses over 10,000 features. The tree kernel based model outperforms both these models by a significant margin.

In this paper, Pak and Alexander et al. [8] analysis of sentiments has been proposed and main focus is given to the users who use twitter as it is one of the popular micro blogging website. In this work, linguistic analysis technique is being performed and it explains the obtained phenomena. The performance results demonstrate that the proposed approach is more efficient and works well in comparison with existing techniques. In addition, it is also indicates that the proposed approach may also used with other language. Author used a text classification scheme based on Multinomial Naïve Bayes to deals with Twitter messages. The effectiveness of this scheme is evaluated using TASS-SEPLN twitter data sets and it achieves maximum macro averaged F1 measure rate of 36.28%. The effectiveness results provided by TASS-SEPLN organizers indicate that the proposal based on MNB is rather effective

Existing techniques:-

KNN Method

It is type of instance based learning or lazy learning. In this learning the function is approximately locally and all computation is deferred until classification. It is simplest of all machine learning algorithms. In KNN classification, the output is class membership. An object is classified by majority votes of its neighbors by the object being assigned to class most common among its k nearest neighbor (k is positive small integer). The nearest neighbor is determined using similarity measure usually distance functions are user.

Naive Bayes Method:-

Algorithm is named after famous statistician Thomas Bayes who proposed Bayesian theorem. The Naïve bayes algorithm is also based on Bayesian theorem. This theorem assumes that all the attributes are conditionally independent to each other. In this algorithm, conditional probability for each attribute with respect to certain class level is calculated The starting point is the Bayes theorem for conditional probability, stating that, for a given data point x and class
C:

$$P(C/x) = P(x/C)* P(C)/ P(x)$$

Furthermore, by making the assumption that for a data point x = {x1,x2,…xj}, the probability of each of its attributes occurring in a given class is independent. Training a Naive Bayes classifier therefore requires calculating the conditional probabilities of each attributes occurring on the predicted classes, which can be estimated from the training data set.

Support Vector Machine (SVM):-

Support vector machine is a method used for classifying the linear data. The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. The SVM method uses a non linear mapping to transform the training data set into high dimensions. The SVM finds the hyper plan using the support vector. SVM has been employed successfully in text classification and in a variety of sequence processing application.

Comparison of existing methods:-

Table: 1 Comparison of Techniques

S. No.	Technique	Learning Methodology	Advantage	Disadvantage
1	SVM	Supervised	Very high accuracy Lesser over fitting Robust to noise	Computatio-nally expensive Slow
2	Naive Bayes	Supervised	Faster training and classification Not sensitive to irrelevant features Handles streaming data well	Less accurate than SVM
3	KNN	Supervised	Simple and easy to understand	Biased by value of K High computation complexity

Concluding remarks:-

Opinion Mining is an important concept in today's world and due to the advent of social media it has become a huge source of database. Since almost everybody in the modern era is involved with some social media platform, the public mood is hugely reflected in the social media today. In this paper, Food price crisis is being studied and also public opinion is predicted for the topic. This paper discussed the classification of sentiments in Indian market for food prices. This paper proposes to utilize this source of information and predict the sentiments of public towards a particular topic.

References:-

1. Horakova, Marketa (2015), "Sentiment Analysis tool using Machine Learning", Global Journal on Technology.
2. Ortigosa, Alvaro, J. M. Martín, R.M. Carro (2014), "Sentiment analysis in Facebook and its application to e-learning", Computers in Human Behavior (31), pp. 527-541.
3. Balahur, Alexandra (2013), "Sentiment analysis in the news", arXiv preprint arXiv, pp. 1309.6202.
4. Gupta, Aditi (2013), "Sentiment analysis for social media", International Journal of Advanced Research in Computer Science and Software Engineering, pp. 216-221.
5. Aisopos, Fotis (2012), "Content vs. context for sentiment analysis: a comparative analysis over microblogs", Proceedings of the 23rd ACM conference on Hypertext and social media.
6. Jebaseeli, A. Nisha, E. Kirubakaran (2012), "A Survey on Sentiment Analysis of (Product) Reviews", International Journal of Computer Applications, pp. 47.11.
7. Agarwal, Apoorv (2011), "Sentiment analysis of twitter data", Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics.
8. Pak, Alexander, Patrick Paroubek (2010),"Twitter as a Corpus for Sentiment Analysis and Opinion Mining", LREC. Vol.10.