Purva Mimaansa | Refereed | Peer Reviewed
ISSN: 0976-0237 | Volume 12, Issue September 2021 | Impact Factor 4.765

A DYNAMIC APPROACH TO IMPROVE THE SENTIMENT SCORE USING MACHINE LEARNING ALGORITHMS ON TWITTER

KL BANSAL
Professor, Department of Computer Science, Himachal Pradesh University Shimla (HP)

SHAURYA VIR SINGH PATHANIA
Student of M. Tech in Department of Computer Science, Himachal Pradesh University Shimla (HP)

ABSTRACT

Microblogs and social media outlets now a day are among the most widely used online networking platforms. Lots of information representing people's views and attitudes is published and posted on platforms like Twitter, Facebook, Instagram on a daily basis by the users. This has recently opened up a lot of doors for the companies that want to track and control the credibility of their brands, products and businesses, as well as policymakers and leaders by helping them in evaluating the public opinions on their policies or political matters. This paper presents a study on twitter data analysis, in which tweets are collected in real time and machine learning is used to comprehend the sentiments and emotions expressed in posts shared on social media. The findings effectively divide the attitudes of social media users into three categories: favourable, negative, and neutral, to have a better understanding of their viewpoints

KEYWORDS: Sentiment analysis, machine learning, twitter.

1. INTRODUCTION

The Internet is a vast virtual space where people can express and share their personal views, affecting every area of life and having consequences for marketing and communication. Consumer habits are influenced by social media by shaping their perceptions and behaviours. Monitoring customer loyalty and sentiment towards brands or products via social media is a good way to measure customer loyalty and keep track of their sentiments. We can comprehend a lot from the massive amount of data that is being shared across multiple social media platforms. This information comes in the form of online journals, comments, and reviews, among other things. People nowadays choose to share their opinions on various topics using social media sites, such as Twitter, Facebook, Pinterest & Quora etc. Few decades ago, people used to express their opinions by writing or speaking in public places. These reviews were also used as suggestions for improvement. This data would then be processed manually. With the advent and increase of the usage of internet, people began to share their opinions about various things via emails or social media platforms.

Sentiment Analysis (SA), also known as opinion mining, is the process of classifying the emotions, conveyed by a text, as negative, positive or neutral. The data made available by social media has contributed to a lot of research activities within SA in recent times. Information gained by applying SA to social media data has many potential usages, for instance, to help marketers evaluate the success of an ad campaign, to identify how different demographics have received a product

release, to predict user behaviour, or to forecast election results.

Sentiment analysers are increasingly being used by businesses to identify flaws in their products or services. An emotion analyser that works as rationally as humans is the best. So, the goal is to close the research gaps in effective sentiment processing.

Microblogging websites like Twitter allow users to write textual entries of up to 280 characters. Commonly referred to as tweets, these have seen a significant increase in their popularity in recent years. Companies and media organisations are increasingly looking for ways to mine Twitter for information about what people think and feel about their goods and services as a result of this development.

Though, a reasonable amount of research has been carried out on how sentiments are expressed in online reviews and news articles, very less research has been done on how sentiments are expressed in micro blogging due to the informal language and message-length constraints.

2. RELATED WORK

There is a plethora of related works for sentiment analysis but, we are only interested in contributions for Twitter Sentiment Analysis. The task of categorizing or recognizing the text as good, negative, or neutral can be termed as sentiment analysis. It's a multifaceted activity that uses Natural Language Processing and Machine Learning approaches to conduct numerous detection tasks at various text granularity levels. There are three approaches to sentiment analysis, lexicon based, machine learning based and a hybrid one. For the purpose of this research, the review of literature is confined to the machine learning technique.

Xuan et al., (2012) introduced syntax-based pattern that is used to extract rich linguistic features and enriched the features with syntactic information of the text. From the investigation, opinions evidenced in the text. Combining these novel features with traditional features obtained from previous studies, it is seen that high accuracy (about 92.1%) appears in detecting subjective sentences on the movie review data.

Samsudin et al., (2012) proposed an Artificial Immune System (AIS) approach that identified Malaysian online movie reviews. In this opinion mining process used three string similarity functions namely cosine similarity, Jaccard coefficient and Sorensen coefficient. Additionally, the performance of AIS was compared with other traditional machine learning techniques, such as Support Vector Machine (SVM), Naïve Bayes and k-Nearest Network. The final results were analysed and discussed.

Kim et al., (2013) introduced a scheme to mine public's opinion from a group of user comments that are easily accessible on social networks regarding the trailer of a new movie. The further step is to predict whether the movie will be a box office hit, according to the public and other factors like the leading actor casting, director, and their previous releases. Through several experiments, it is seen that the scheme can produce satisfactory result.

Liang & Dai (2013) proposed a novel system of architecture that could automatically analyze the sentiments of the given message. The authors merged this system with manual annotated data from Twitter which is one of the most popular micro-blogging platform analyzing sentiments. This system teaches the machine to learn the automatic extraction of the set of messages that contains opinion and filter out non-opinion message and finally determines their sentiment direction (i.e., positive, negative). Final results verified the system effectiveness on sentiment analysis in real micro-blogging application

Habernal et al., (2014) experimented with state-of-the-art supervised machine learning techniques for analysing the sentiments. The authors explored several pre-processing methods and employed different features and classifiers. They also evaluated five feature selection algorithms and examined the influence of named entity recognition as well as pre-processing on the performance of sentiment classification. In addition to newly created social media datasets, it also reported result for various popular domains like movie and product reviews. This not only extended the present sentiment analysis research to other family of languages, but also encourages competition that potentially led to the development of high-end commercial solution

Sharma et al., (2014) investigated document based opinion mining systems that are able to classify the given documents as positive, negative and neutral. In addition, negation was also taken care in the proposed method. Experimental result with movie reviews proved the effectiveness.

Ganeshbhai & Shah (2015) proposed an autonomous text analysis and summarization system for reviews available on Web. Opinion mining aims for differentiating the emotions that are expressed in the reviews, segregating them into positive or negative and summarizing such that it could be easily interpreted by the users. Feature based opinion mining analyses fine-grain by recognition of individual features of an object upon which user has expressed opinion. This method gives a view of several techniques proposed in the domain of feature oriented opinion mining and also discusses the limitation of available works and further scope of feature based opinion mining.

Agarwal & Mittal (2016) investigated various machine learning algorithms that were extensively used for analysing sentiments. The Bag-of-Words (BoW) representation is the commonly used technique for sentiment analysis. This method works on the concept that it assumes the independency of a word and not considers the significance of semantics and subjective information included in the texts. Each and every word in the text is given equal preference. It results in the feature space high dimensionality. Machine learning algorithm reduces this high-dimensional feature space by using feature selection technique that selects significant features alone by removing noisy and irrelevant feature.

3. METHODOLOGY

"Sentiment Analysis is the computational study of people's feelings and opinions expressed in the text". In other words, it is the task of identifying the mood of people about a particular subject. It is also referred as Opinion Mining, Subjectivity Analysis, Appraisal Extraction and Review Mining with some associations to Affective Computing. It automates the retrieval of text from appropriate sources, extraction of relevant sentences, understanding its contents, summarizing it and presenting the results in an appropriate format. In Computational Terms it is defined "as a data mining technique that uses Natural Language Processing, Computational Linguistic and Text Analytics to identify and extract content of interest from a body of textual data"

Sentiment analysis is a data mining task that scientifically extracts and analyses online content in a real time. Despite of the volume and data structures, sentiment analysis presents unique opportunities to marketers, to learn customer's feelings and emotions without incurring any time delays. Thus, it is an expanding field that has its foundation in Text Mining, Computational Linguistics and Natural Language Processing.

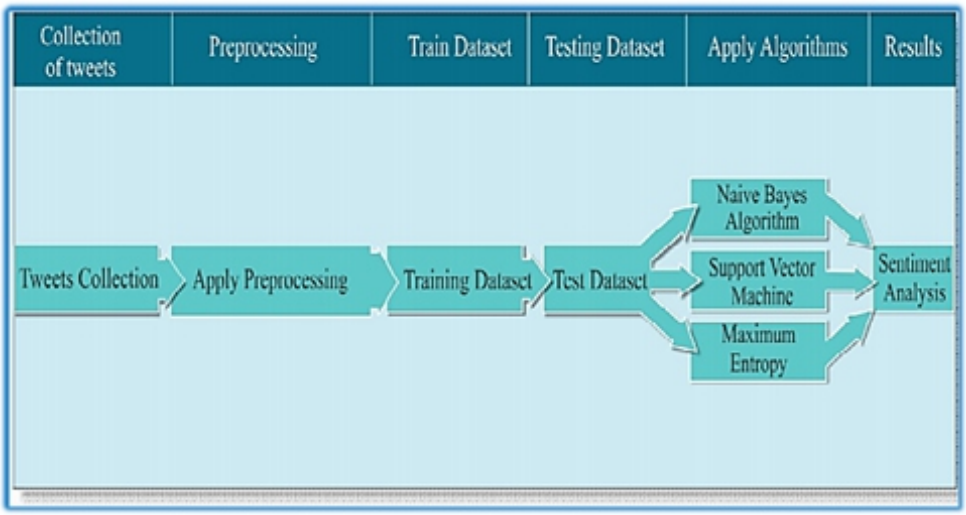
3.1 MACHINE LEARNING APPROACH

Machine learning approaches use a variety of training data to develop predictive models such as logistic regressions, neural networks, decision trees, and other models that construct forecasts on

documents that aren't in the training set. This approach has the benefit of being based on a learning model. It's helpful for creating quick and accurate forecasts.

Furthermore, algorithms are capable of discovering previously unthinkable and complex patterns that are beyond human comprehension. However it has significant drawbacks to develop the model. For this we need training data. Validation of the model is difficult due to its complexity. It's difficult and time-consuming. It is important to give each of these documents a rating. If there are document attributes, it is also necessary to assign a rating to them.

Another issue emerges when two separate reviewers give different ratings exactly to the same document. This can give rise to unexpected errors in the construction and measurement of the model performance. The approach starts with observations or data, such as examples, direct experience, or instruction, so that we can seek for patterns in data and make better decisions in the future based on the examples we provide.



3.2 COLLECTION OF TWEETS

Standard twitter dataset is not available for related data domain so in this research we collect twitter datasets from Twitter API and create twitter Management application. The proposed work twitter Management application account generates Access key, Secret Key, Access Token and Application Authentication ID this credential used for fetching data from Twitter Account.

For collection of tweets from twitter following are the basic R Packages required library (twitteR), library (ROAuth), library (plyr), library (tm) etc.

3.3 DATA PRE-PROCESSING

Data pre-processing is needed because in proposed work we get raw data with 16 attributes of twitter datasets. We apply pre-processing for removing favourited, favourite Count, replyToSN, created, truncated, replyToSID, id, replyToUID, status Source, screen Name, retweet Count, is Retweet , retweeted, longitude , latitude and text attributes for text attribute more pre-processing is required. We apply both Twitter-specific and standard pre-processing on the collected twitter datasets. The specific pre-processing is especially important for Twitter messages, since the Twitter community

has created its own unique phrases and forms to write messages. User-generated content in social media often contains slang and frequent grammatical and spelling mistakes. With Twitter-specific text pre-processing we try to handle these properties of the Twitter language and improve the quality of features. In this we consider the following options for Twitter-specific pre-processing to better define the classification

Data pre-processing is done to eliminate the incomplete, noisy and inconsistent data. Data must be pre-processed in order to perform any data mining functionality.

3.4 TRAINING DATASETS

This is the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict. If you are using supervised learning or some hybrid that includes that approach, your data will be enriched with data lab. This is used to measure the performance, such as accuracy or efficiency, of the algorithm we are using to train the machine. Test data will help us to see how well our model can predict new answers, based on its training. Both training and test data are important for improving and validating machine learning models.

3.5 APPLYING CLASSIFIERS

For this research work, applying the three machine Learning classifiers for performance evaluation

Support Vector Machine Classifiers

Classification refers to the task of predicting a class label for a given twitter data. The SVM algorithm has been often successfully applied. It has several advantages, which are important for learning a sentiment classifier from a twitter dataset it is fairly robust to over fitting and it can handle feature. SVM algorithm represents the positively and negatively labelled training dataset in the high dimensional space and separates them by a hyperplane. There are many possible hyperplanes which could separate the training dataset that belong to different classes, but the aim of the SVM algorithm is to choose the one which separates them with the largest possible gap, i.e., for which the margin between the training dataset of both classes and the SVM hyperplane is as large as possible. The larger margin indicates the lower classification error of the new unseen points.

Naïve Bayes Classifier

Naïve Bayes Classifier is supervised machine learning approach. This supervised classifier was given by Thomas Bayes and hence the name. Naïve Bayes is a classification algorithm that score how well each point belongs to each class based on feature. It has been used because of its simplicity in both during training and classifying stage. It is a probabilistic classifier and can learn the pattern of examining a set of sentences that has been categorized. It compares the contents with the list of words to classify the sentence to their right category.

Maximum Entropy Classifier

This is maximum entropy as a technique for estimating probability distributions using data. The most important rule in maximum entropy is that when nothing is known, the distribution should be kept uniform and it should have maximal entropy. In order to gather a set of constraints for the model, which describe class specific expectations for the distribution, labelled training data is utilized.

3.6 PERFORMANCE MEASURE

Following are the evaluative measures of classifiers. These evaluative measures find the

goodness of classifier and suitability of data.

True Positives (TP): These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes. E.g., if actual class value indicates that this passenger survived and predicted class tells you the same thing

True Negatives (TN): These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g., if actual class says this passenger did not survive and predicted class tells you the same thing. False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

False Positives (FP): When actual class is no and predicted class is yes. E.g., if actual class says this passenger did not survive but predicted class tells you that this passenger will survive

False Negatives (FN): When actual class is yes but predicted class in no. E.g., if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

The proposed work used three parameters to calculate Accuracy, Precision and Recall:

Accuracy: is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations

Accuracy = TP+TN/TP+FP+FN+TN

Precision: is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/TP+FP

Recall (Sensitivity): is the ratio of correctly predicted positive observations to the all observations in actual class

Recall = TP/TP+FN

3.7 SENTIMENT SCORE

The most important part of sentiment analysis to generate score each tweet, score. Sentiment ( ) function is used to iterate through the input text. It strips punctuation and control characters from each line using in R Programming platform regular expression powered substitution function, and matches against each word list to find matches. Sentiment ( ) function assigns score to the tweets using the formula as

Sentiment Score = sum (pos.matches) – sum(neg.matches)

Obtain the score file with

If Score > 0, means that the tweet has 'Positive Sentiment'

If Score < 0, means that the tweet has 'Negative Sentiment'

If Score = 0, means that the tweet has 'Neutral Sentiment'

4. EXPERIMENT & RESULTS

We can use the Twitter API to create microblogging posts, which implies we can obtain a large amount of Twitter messages about any subject for analysis. For frequent posting of tweets allows us to analyse not only a fixed amount of pre-collected Twitter messages, but also to apply the algorithms on Twitter.

In the machine learning approach, there are many possible algorithms that can be applied to

sentiment analysis Support Vector Machine (SVM), Naive Bayes (NB), and Maximum Entropy (MaxEnt). This approach and the algorithm, an important step is also to decide which text pre-processing techniques should be applied on the twitter data in order to improve its quality and prepare it for the algorithm. There exist several Machine Learning algorithm types which can be used in the context of sentiment analysis. Most of the approaches to determining the sentiment polarity of a sentence are based on supervised machine learning methods require a class labelled collection of Sentence, which have been pre-categorised. These methods analyse sentence, features, and organizes the sentence. In the context of a Sentiment Analysis and changes which can occur in it, the supervised machine learning approach is highly suitable since the classification model can be flexible

The machine learning common approach is designed. It performs the Tweet collection, Tweet Pre-processing, Train the Dataset, Test the Dataset and apply the classifiers and obtain the Results. The Twitter dataset selection method is well explained and then the pre-processing techniques for data enhancement used. After pre-processing feature extraction and selection method have been explained and then classified the feature selected by Machine learning sentiment analysis approaches in that perform training data, test data and apply the classifiers i.e. Support Vector Machine, Naïve Bayes, Maximum Entropy and at last obtain the results. The Performance evaluation using Accuracy, Precision and Recall

5. CONCLUSION

The overall goal of our research work was to establish a pure connection between different aspects of sentiment generation. As the social media has a vast source of information, there should be a reliable senti-score generation methodology. Sentimental analysis is the process of tracking public reviews about a particular topic, product or services. This process involves collecting the available information, extracting the features, selecting the needed features and finally making the classification to arrive at the opinions. With the rapid growth of e-commerce sites, public forums and social media on the Web, individuals and organizations are increasingly using public opinions available in these media for making their decision. In addition the information available for making the decision is also increased. Sentiment analysis process involves feature extraction, feature selection and finally sentiment classification. The feature selection step in sentiment analysis has high impact in determining the accuracy of sentiment classification.

WORKS CITED

Xuan, H.N.T., Le, A.C., & Nguyen, L.M. (2012, November). Linguistic features for subjectivity classification. In Asian Language Processing (IALP), 2012 International Conference, pp. 17-20. IEEE.

Samsudin, N., Puteh, M., Hamdan, A.R., & Nazri, M.Z.A. (2012, September). Is artificial immune system suitable for opinion mining?. In 2012 4th Conference on Data Mining and Optimization (DMO) pp. 131-136. IEEE.

Kim, D., Kim, D., Hwang, E., & Choi, H.G. (2013). A user opinion and metadata mining scheme for predicting box office performance of movies in the social network environment. New Review of Hypermedia and Multimedia, 19(3-4), 259-272.

Liu, Y., Yu, X., An, A., & Huang, X. (2013). Riding the tide of sentiment change: sentiment analysis with evolving online reviews. World Wide Web, 16(4), 477-496.

Habernal, I., Ptácek, T., & Steinberger, J. (2014). Supervised sentiment analysis in Czech social media. Information Processing & Management, 50(5), 693-707.

Sharma, N.R., & Chitre, V.D. (2014). 'Opinion mining, analysis and its challenges.' International Journal of Innovations & Advancement in Computer Science, 3(1), 59- 65.

Ganeshbhai, S.Y., & Shah, B.K. (2015, June). Feature based opinion mining: A survey. In Advance Computing Conference (IACC), 2015 IEEE International, pp. 919-923. IEEE.

Agarwal, B., & Mittal, N. (2016). Machine Learning Approach for Sentiment Analysis. In Prominent Feature Extraction for Sentiment Analysis, pp. 21-45. Springer International Publishing.