Sentiment Analysis of Twitter Data – Survey

Priyanka S. Dongre[1] Dr. S. D. Sawarkar[2]
[2]Principal
[1,2]Dattameghe College of Engineering, Airoli, Navi Mumbai, India

Abstract— Twitter is a web service and social communication platform which allow users to address their tweets in different domains on internet. Public can easily write their perspectives and ideas on a wide variety of topics via social networking websites. As online data is openly available through different platforms like social networks, twitter, Facebook, etc... Analyzing the data is of paramount importance in drawing inference from the data. Hence, in our survey, we try to identify sentiment analysis on twitter data by using a learning algorithms. By using our research, we can identify the measures customers' opinions and perceptions and can be enhanced to any desired level depending on the data gathered from online resources.

Keywords: Twitter, Support Vector Machine (SVM), Lexicon

I. INTRODUCTION

Sentiment analysis (SA) tells user whether the information about the product is satisfactory or not before they buy it. The sentences that represent observations or attitude or opinions that is expressed as positive or negative or neutral emotions are called as sentiments. From twitter tweets are extracted in the form of unstructured data. The unstructured dataset is converted into structured form with pre-processing techniques then extracts features from structured review. The features of the words are selected and then classification technique is applied on extracted features are to classify them into its sentiment polarity that is namely either positive or negative or neutral. Feature words representation based on classifier used is the main algorithm used in system by information retrieval researchers to represent text corpus. It is an easy approach to convert unstructured text into structured data for only English language based on word by word and the grammar is neglected.

II. SENTIMENT ANALYSIS

Sentiment analysis is process is process of identifying emotions or opinions though the online medium. In this paper we have interpreted twitter data sentiment analysis using various algorithms. Twitter is a micro-blogging site that is rapidly growing in terms of number of users of different languages. Moreover, Tweets are mostly publically visible and limited to 140 characters that simplify the identification of emotions or sentiments in text. Though, the abundance of data, use of short forms, timing of different posts, smileys, images, videos, animations and diversity of language make the sentiment analysis process difficult for Twitter data. Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral" based on their reviews or satisfaction after the use of that specific product. This survey focuses mainly on sentiment analysis of twitter data which will be helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in given cases.

III. RELATED WORK

Sentiment analysis consists of two parts: Unsupervised learning and supervised learning of machine learning, each of them have their classification techniques. Before checking for the supervised/unsupervised learning data should be understandable format to algorithms. For which pre-processing is must and then classify tweets as per sentiments assigned.

A tweet contains a lot of sentiments or views or opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labelled into two categories i.e negative and positive and neutral polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Pre-processing of tweet include following points:

- Remove all URLs (e.g. www.domain.com), hash tags (e.g. #topic), targets (@username)
- Correct the spellings; sequence of repeated characters is to be handled mostly removed.
- Replace all the emoticons with their sentiment.
- Remove all punctuations, symbols, numbers, images and videos.
- Remove the Stop Words.
- Expand Acronyms(we can use an acronym dictionary)
- Remove Non-English Tweets.

A. Unsupervised Learning:

It does not consist of a specific category and they do not provide with the correct targets at all and therefore rely mainly on clustering.

1) Lexicon-Based Approaches:

Lexicon based method uses sentiment dictionary with opinion words and match them with the data to determine polarity. They assigns sentiment scores to the opinion words describing how Positive, Negative and Objective the words contained in the dictionary are[02].

Lexicon-based approaches mainly rely on a sentiment lexicon which are stored in database, i.e., a collection of known and precompiled sentiment terms, phrases and even idioms, developed for traditional genres of communication, such as the Opinion Finder lexicon[04].

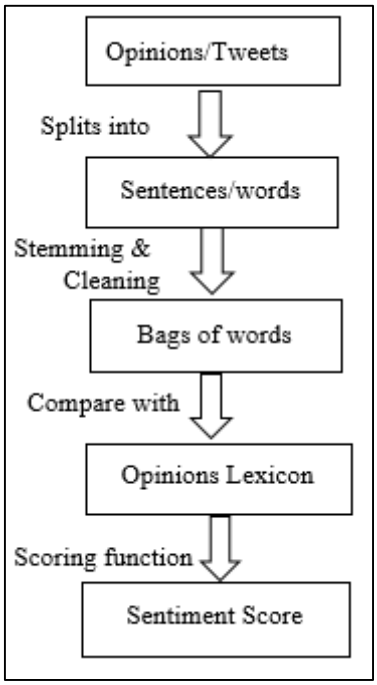


Fig. 1: Lexicon based Sentiment analysis approach

a) Dictionary-based:

It is based on the usage of sentences that are usually collected and annotated manually. This set grows by searching the synonyms and antonyms of a dictionary. An example of that dictionary is WordNet, which is used to develop a thesaurus called SentiWordNet dictionary.[10]

Drawback: Can't deal with domain and context specific orientations of data.[05]

b) Corpus-Based:

The corpus-based approach specifies objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of these opinion terms that grows through search of the related words by means of the use of either statistical techniques or semantic techniques [10].

- Methods based on statistics: Latent Semantic Analysis (LSA) is used how many times specific dictionary is being used.
- Methods based on the semantic such as the use of synonyms and antonyms or relationships are from thesaurus like WordNet dictionary may also represent an interesting solution. According to the performance measures like precision and accuracy we provide a comprehensive study of existing techniques for opinion mining, including machine learning, lexicon-based approaches and other supervised approaches.

2) Supervised Learning:

It is based on labelled dataset and thus the labels are provided to the model during the process. These labelled dataset are trained to get meaningful results when encountered during decision- making of process. The success of both these mentioned machine learning methods is mainly depends on the selection of datasets and extraction of the specific set of features used to detect sentiment. The machine learning approach applicable to sentiment analysis and opinion mining which mainly belongs to supervised classification. In a machine learning approch, two sets of data are needed:

1) Training Set

2) Test Set.

A number of machine learning techniques have been formulated to classify the tweets into classes depends on the results the formulas gives. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in sentiment analysis. Machine learning starts with collecting training dataset. Next we train a classifier on the training data. Once a supervised classification technique is selected, an important decision to make feature selection. They can tell us how documents are represented. The most commonly used features in learning techniques for sentiment classification are:

- Term presence and their frequency
- Part of speech(POS) information
- Negations(negative words)
- Opinion words and phrases

a) Maximum Entropy:

The maximum entropy relies on probability distribution of estimation technique to perform classification for sentiment analysis. In this technique, firstly the categorized feature sets are converted into definite vectors using any of the encoding schemes. Secondly, this encoded vector is used to compute weights for each of the extracted features that can collectively support in determining the most prospective label for a feature set. It is used for various natural language processing (NLP) tasks such as text classification of datasets. It depends on the probabilistic approach like Naïve Bayes [06-07]. The concept of maximum entropy is that if much information regarding the data is not known, the distribution should be extremely uniform. This constraint eliminates the probability of non-uniform distribution.[11] The probability is derived from the categorized training data and denoted as expected values of extracted features as follows:

$$P(c \mid d) = \frac{1}{Z(d)\{\exp(\sum \lambda i \, fi(c,d))\}}$$

Where $f_i(c, d)$ is a feature, λ_i is a parameter to be predicted and Z(d) is a normalization function. Unlike NB, maximum entropy doesn't make any assumption regarding the feature independency. The motivating idea behind maximum entropy learning is building a uniform model that satisfies all the given required constrains. For example, consider a four-way text classification problem with a constraint given in percentage as: on average 30% of documents having a word "professor" is labeled as faculty class. And co-incidentally when given a document with "professor" word within, we assume that it has a 30% chance to be labeled as a faculty class, and a 15% chance to be labeled as each of the other three classes. If a document does not have "professor" word within then as per the law of uniform class distribution, we assume the probability of that document to be in each class is 25%. This model is precisely the maximum entropy model that conforms to each given or known constraint [11].

b) Support Vector Machine (SVM):

Support vector machine (SVM) solves the traditional text classification problem effectively, mostly outperforming Naïve Bayes as it supports the concept of maximum margin.

The main principle of SVMs is to determine a linear separator that separates various different classes in the search space with maximum distance covered i.e. with maximum margin [13-17]. If we represent the tweet using t, the hyper plane using h, and classes using a set Cj € {l, -1} into which the tweet has to be classified. Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space.[02] The main principle of SVMs is to determine a linear separator that separates different classes in the search space with maximum distance i.e. with maximum margin, the solution is written as follows equivalent to the sentiment of the tweet[11].

The idea of SVM is to determine a boundary or many boundaries that separate distinct clusters or groups or array of data. SVM performs this task for constructing a set of points and separating those points using mathematical formulas. Fig. 1 illustrates the data flow of SVM.[01-02]

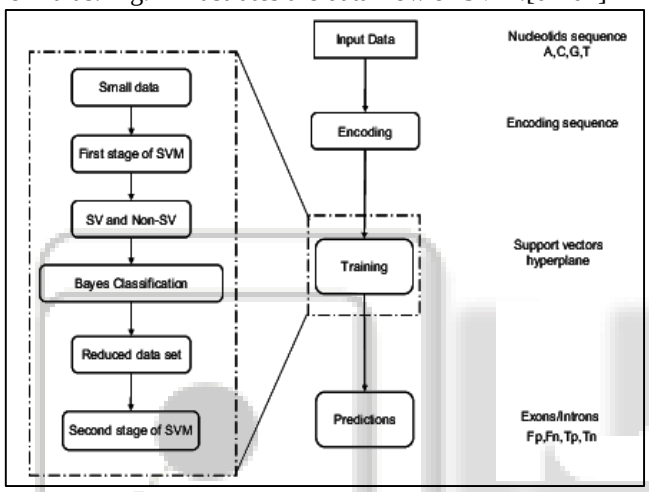


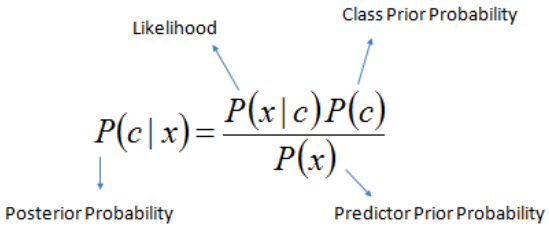
Fig. 1: SVM Algorithm

c) Random Forest

Random Forest classifier is a tree-based classifier. It consists of numerous classification trees that can be used to predict the class label for a given data point based on the categorical dependent variable [19]. For a given data point, each tree votes for a particular class label for each turn and the class label gaining the maximum votes will be assigned to that data point which gives maximum accuracy. The error rate of this classifier depends on the correlation or association among any two trees in the forest in addition to the strength of definite or individual tree in the forest. In order to minimize the error rate, the trees should be strong and the degree of associativity should be as less as possible. In the classifier tree, the internal nodes are represented as their features, the edges leaving a node are represented as tests on the feature's weight, and the leaves are represented as class categories.[11] It performs classification preliminary from the root node and moves incrementally downward until a leaf node is detected. The document is then classified in the category that labels the leaf node. This algorithm is used in many applications of speech and language processing.

d) Naïve Bayes Classifier:

It is used to predict the probability for a given words to belong to a particular class. It is used because of its easiness in both during training and classifying steps to generate sentiment analysis. Pre-processed data is given as input to train input set using Naïve Bayes classifier and that trained model is applied on test to generate either positive or negative or neutral sentiment. The Bayes theorem is as follows:



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

where X- Tuples, H-Hypothesis, P(C|X) represents Posterior probability of H conditioned on X i.e. the Probability that Hypothesis holds true given the value of X, P(C) represents Prior probability of H i.e. The Probability that C holds true irrespective of the tuple values, P(X|H) are represents posterior probability of X conditioned on C i.e. all the Probability that X will have certain values for a given Hypothesis, P(X) represents Prior probability of X i.e the Probability that X will have certain values. The proposed system understands whether the tweet is positive or system understands whether the tweet is positive or negative based on the dictionary methods of score.

IV. Conclusion

Twitter is an excellent source for social media analysis. People directly share their opinions through Twitter to the general public. One of the very common analyses which can perform on a large number of tweets is sentiment analysis. In the paper, we provided a survey and comprehensive study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with cross domain. Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy and hence machine learning algorithms are mostly used in analysis kind of implementations in real time world, while lexicon-based methods are very effective in some cases, which require few effort in human-labeled document .We also studied the effects of various features on classifier. We can conclude that more the cleaner data, more accurate results can be obtained.

References

[1] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He, Member, "Tweet Segmentation and Its Application" to Named Entity Recognition", IEEE VOL. 27, NO. 2, FEBRUARY 2015

[2] Vishal A. Kharde and S.S. Sonawane "Sentiment Analysis of Twitter Data: A Survey of Techniques" (0975 – 8887) Volume 139 – No.11, April 2016

[3] Dr.S.Kannan "Preprocessing Techniques for Text Mining", 2731-7322, VOL 31, NO. 25, 05 March 2015

[4] Kavya Suppala, Narasinga Rao "Sentiment Analysis Using Naïve Bayes Classifier". ISSN: 2278-3075, Volume-8 Issue-8 June, 2019

[5] Varsha Sahayak Vijaya Shete Apashabi Pathan "Sentiment Analysis on Twitter Data" Issue 1, Volume 2 (January 2015)
[6] M. Vadivukarassi, N. Puviarasan and P. Aruna Annamalai University, Tamil Nadu, India. "Sentimental Analysis of Tweets Using Naive Bayes Algorithm" World Applied Sciences Journal 35 (1): 54-59, 2017
[7] Bhagyashri Wagh1, J. V. Shinde2, N. R. Wankhade3, "Sentimental Analysis on Twitter Data using Naive Bayes" IJARCCE Vol. 5, Issue 12, December 2016
[8] Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan, "Mining Social Media Data for Understanding Students' Learning Experiences" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 7, NO. 3, JULYSEPTEMBER 2014
[9] Lukas Povoda and Radim Burget and Malay Kishore Dutta, "Sentiment Analysis Based on Support Vector Machine and Big Data" IEEE 978-1-5090-12886/16 ©2016.
[10] Govin Gaikwad and Prof. Deepali J. Joshi. "Multiclass Mood Classification on Twitter Using Lexicon Dictionary and Machine Learning Algorithms" IRJET Vol. 15, 14 December, 2015
[11] Survey Mitali Desai & Mayuri A. Mehta. "Techniques for Sentiment Analysis of Twitter Data: A Comprehensive" ICCCA2016 ISBN: 978-1-5090-1666-2/16/ ©2016 IEEE

