A Review on Sentiment Analysis of Twitter Data

Preeti Mehrotra
Dept. of Computer Science
Truba Institute of Engineering & Information Technology

Devashri Anwekar
Dept. of Computer Science
Truba Institute of Engineering & Information Technology

Amit Saxena
Dept. of Computer Science
Truba Institute of Engineering & Information Technology

ABSTRACT
"Sentiment analysis is a sort of natural language processing" for measuring the sentiment of the public regarding a given product or issue. Sentiment analysis, that is often called "opinion mining", includes in constructing a system to gather and analyse opinions about product stated through comments, reviews, blog posts or tweets. "Sentiment analysis" may be valuable in numerous ways. In reality, it has moved from computer science to the management sciences as well as social sciences because to its relevance to society and business as a whole.

There has been much of effort in the topic of "sentiment analysis" of the twitter data. This study focuses largely on "sentiment analysis" of the twitter data that is beneficial to assess the information in the tweets where views are very unstructured, varied and are either "negative, positive or neutral" in certain situations. In this work, we give a survey and a current methodology for the opinion mining.

Keywords
Sentiment Analysis, Text Mining.

1. INTRODUCTION
Processing of real text elements is the focus of "Natural Language Processing " (NLP). NLP is used to convert the text into a machine-readable format. Artificial Intelligence (AI) combines NLP data plus a lot of mathematics to evaluate whether a situation is good or bad. " Natural language textual information" may be used to identify an author's perspective on a given subject. Machine learning is used in a variety of ways, each with various degrees of success. Opinion mining is a sort of "natural language processing" that focuses on gauging public sentiment about a certain product or issue. In addition to extracting ideas, emotions, and sentiments automatically from text, this programme keeps tabs on people's online attitudes and feelings. People use blogs, comments, reviews, and tweets to voice their opinions on a wide range of issues. Using the internet, you can keep tabs on items and companies to see how consumers feel about them. Opinon mining has a wide range of names and functions: it may be called anything from sentiment analysis to opinion extraction to sentiment mining. Artificial intelligence as well as computational linguistics all intersect in the topic of "natural language processing", which deals with how computers and human languages communicate with one another. As a result, NLP is a part of the human–computer interaction domain. There are a number of issues in NLP that revolve on natural language understanding as well as natural language generation, both of which need computers to be able to deduce meaning from human or the natural language input. Machine learning, particularly statistical machine learning, is at the heart of modern NLP algorithms. Machine learning has a different paradigm than most previous efforts to process language. In previous systems, enormous sets of rules had to be hand-coded in order to parse language. When it comes to the machine-learning paradigm, however, such principles may be learned via the study of enormous corpora of the typical real-world instances using generic learning algorithms—often, but not always, founded on statistical inference. It is possible to create a corpus by hand-annotating a collection of documents (or even individual phrases) with the values to be learnt. NLP problems have been applied to a wide range of machine learning techniques. Various "features" obtained from input data are fed into these algorithms, which then use them to do various tasks. Decision trees, one of the first commonly used algorithms, created systems of hard if-then rules akin to those previously popular in handwritten rule systems. A growing number of researchers have been focusing on statistical models, which utilise real-valued weights to make soft, probabilistic judgments. When used as part of a broader system, these models offer the benefit of being able to describe the relative confidence of many distinct potential solutions rather than just one.

Sentiment analysis is based on the notion that text documents or short phrases may be classified according to their polarity. "positive," "negative," or "impartial" are the three types of emotional polarity (neutral). "Sentiment mining" may be done on three levels, as shown in the following table [1]:

- "Document-level sentiment classification": A document may be categorised as "positive," "negative," or "neutral" at this level.
- "Sentence-level sentiment classification": At this level, each statement is categorised as either "positive," "negative," or "unbiased".
- "Aspect and feature level sentiment classification": "perspective-level assessment grouping" may be used to classify sentences/archives as "positive," "negative," or "non-partisan" based on specific elements of the sentences/archives.

2. SENTIMENT ANALYSIS
"Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP)". "Sentiment analysis" is the process of categorizing the views expressed in a piece of writing into "positive" or "negative" or "neutral" categories. A variety of terms are used to describe it, including opinion mining, subjectivity analysis, and assessment extraction. View, sentiment, opinion as well as belief are all synonyms, but there are some important distinctions to be aware of.

- *Opinion*: "A conclusion open to dispute (because different experts have different opinions)"
- *View*: "subjective opinion"
- *Belief*: "deliberate acceptance and intellectual assent"
- *Sentiment*: "opinion representing one's feelings"

Since the last several years, a number of scholars have been working in the topic of "Sentiment Analysis on Twitter." At first it was meant to be used for binary categorization, assigning comments or reviews to one of two categories, like "positive" or "negative".

An objective, positive, as well as negative classification paradigm was put out by [2]. They used Twitter API to gather tweets, and then used emoticons to automatically annotate those messages. A "sentiment classifier" based on the "multinomial Naive Bayes" approach was created using that data, and it makes use of characteristics like Ngram as well as POS tags. They utilised a training set that solely included tweets with emoticons, which was less effective.

For tweet classification, [3] used two models, a "Naive Bayes bigram model and a Maximum Entropy model". Classifiers based on Naive Bayes were shown to be more effective than those based on Maximum Entropy.

Distant supervision was used to train their system on tweets with emoticons that acted as noisy labels, and the result was a solution for the sentiment analysis for the Twitter data [4]. They use "MaxEnt, Naive Bayes, and Support Vector Machines (SVM)" to develop their models. Unigrams, bigrams, as well as POS were the only features they had. As a result, SVM was shown to be superior to other models, while unigrams were found to be more useful as feature attributes

"Two-step automated sentiment analysis" for tweets was devised by [5]. Objective tweets were separated from subjective ones, which were then further divided into positive & negative categories. Additionally, characteristics such as hashtags, exclamation as well as retweets were employed along with with features like preceding polarity of words as well as POS in the feature space.

Twitter streaming data from Firehouse API was utilized by [6] to get all publicly accessible tweets from each user in the real-time. "Multinomial naive Bayes, stochastic gradient descent and the Hoeffding tree" were also tried out during this time period as well. The "SGD-based model" when employed with an acceptable learning rate outperformed the rest of them, according to their findings.

3. CLASSIFICATION TECHNIQUES

It is possible to categorise unlabeled data using a variety of approaches in the machine learning discipline. Training data may be necessary for classifiers. "Maximum Entropy, Naive Bayes, and Support Vector Machine" are examples of the machine learning classifiers [7, 8]. Because they need the use of pre-existing data, these techniques are referred to as supervised machine learning. It's worth noting that improving the accuracy of future predictions is made possible by efficiently training a classifier.

A. Support Vector Machine

When it comes to sentiment analysis, the "support vector machine" (SVM) is recognised to perform well. Information is gathered using SVM, which then defines choice boundaries and employs components from the input space for the computation [11]. The most important information is supplied as two vectors of size m, one for each of the two layouts. At this step, each vector is sorted into a certain class. A boundary between the classes that is not present in any of the training data is then detected by the machine [12]. Expanding the categorization edge reduces the number of ambiguous possibilities. In a variety of text classification tasks, the SVM outperformed the "Nave Bayes classifier", as illustrated in [13].

B. Naïve Bayes

Bayes' Theorem is used in conjunction with strong (naive) independence assumptions for this classification approach. Classifiers using the Naive Bayesian approach assume that a feature (element)'s proximity to other features does not depend on how closely the other features are related to it. There are a number of characteristics that make an organic fruit an apple: its colour is red; its form is round; and its width is around three inches. This natural fruit's chance of being an apple means that a "Nave Bayes classifier" would treat these attributes as independent, regardless of whether they are linked or not. The Naive Bayes has been shown to outperform even the most advanced order techniques.

4. SENTENCE-LEVEL SENTIMENT ANALYSIS APPROACHES

Classifying sentences into negative, positive, or neutral categories is the goal of this study. "Sentence-level sentiment analysis" is an example of Twitter sentiment analysis, according to this definition. Twitter sentiment analysis methodologies are discussed in the next section. Text is classified using classification algorithms in machine learning systems. supervised learning as well as ensemble learning are the two most common methods of "machine learning". "Supervised machine learning", ensemble techniques, lexicon-based analysis, as well as hybrid analysis are the four main ways to analyzing Twitter sentiment. These four methods are outlined in detail as follows:

Twitter-Sentiment-Analysis using Supervised-ML Approaches

During the training phase, machine learning models are fed tagged datasets. Using these datasets, these models are trained to provide meaningful results. Both a training set and a test set are needed in machine learning systems. Classifiers, a machine learning method, may be used to determine the mood of Twitter users. As the volume of training data as well as feature sets are extractors, "Twitter sentiment classifiers" perform best. SVM and NB classifiers, two machine-learning-based techniques for analysing Twitter sentiment, are becoming more popular. In order to analyse sentiment on Twitter, supervised machine learning algorithms are used (see Fig. 1).

Three stages are involved in the "Twitter sentiment analysis" process. To begin, the classifier is trained on a variety of different types of tweets, including negative, positive, and neutral. The following are some examples of tweets that have been tweeted:

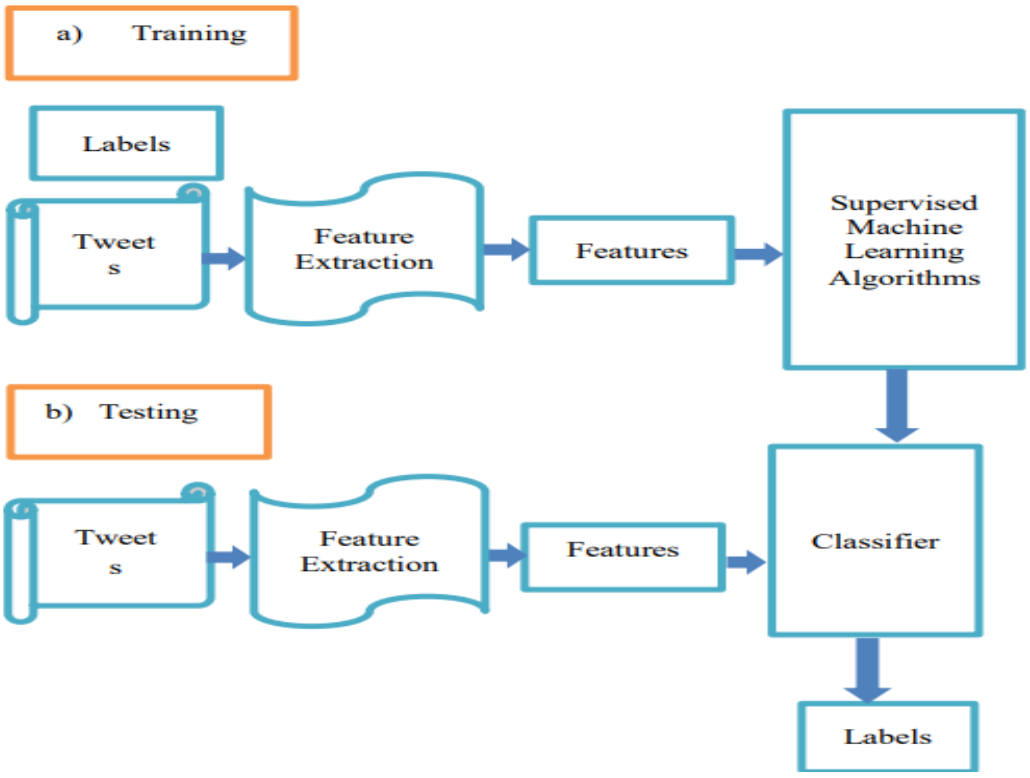

Figure:1. Sentiment Analysis using Supervised ML Algorithms

- The following tweets are examples of positive tweets:

1) "PM@narendramodi and the President of Ghana, Nana Akufo-Addo had a wonderful meeting. Their talks included discussions on energy, climate change and trade ties.
2) Billy D. Williams @Msdebramaye For the children, they mark, and the children, they know The place where the sidewalk ends.
3) @abdullah "Staying positive is all in your head". #PositiveTweets."

- Unbiased tweets

1) "(@Nisha38871234): "#WorldBloodDonorDay Blood Donation is the best donation in the world. Save a life!!"Good night #Twitter and #TheLegionoftheFallen. 5:45am cimes awfully early!
2) (@imunbiased). Be excellent to each other. Up a WV holler..or in NoVA
3) Today several crucial MoUs were signed that will boost India-France friendship."

- Negative tweets

1) "Any negative polls are fake news, just like the CNN, #DonaldTrump
2) Can Hillary please hire the genius/magician who dressed Palin in 2008 and stop dressing like my weird cat-lady aunt who works at JCPenney? — kara vallow (@teenagesleuth)
3) Sasha and Malia Obama, daughters have some selfie fun during the Inaugural Parade for their father President Obama on ... Follow @JessicaDurando"

Twitter Sentiment Analysis using Ensemble Approaches

Multiple classifiers are used in ensemble techniques in order to improve prediction precision and accuracy. "Text categorization using ensemble techniques is common, and it's possible that using these approaches to enhance the accuracy of "Twitter sentiment analysis" postings might be beneficial.

For emotion categorization, [14] tested the usefulness of constructing ensemble learners. The goal was to construct a more powerful classifier by combining several feature sets and classification methods. Sentiment categorization was done using an amalgamation of several skills and arranging processes. It is not possible to classify sentiments using traditional text classification methods since the bag of words (BOW) does not include all of the words in a sentence. 2 feature types – "POS & Word-relations" as well as 3 classifiers - "NB, MaxEnt, & SVM" were used in this research. "Fixed grouping, Weighted grouping, as well as meta-classifier grouping" were all tested and shown to be effective. The findings demonstrated that as compared to an individual classifier, the ensemble approaches produced much better results. Results also showed that combining two distinct classifiers with the different feature sets resulted in substantial gains.

Incorporating numerous classifiers into massive twitter data sets was one idea put up by [15]. They used hashed 4-grams as features to train "logistic regression (LR) classifiers". More than 100 million samples and ensembles of 3 to 41 classification algorithms were used as part of the training datasets. The experiment found that "sentiment analysis" of the Twitter data using several classifiers was more

accurate than using only one. Due to how many classifiers were used, the running time rose as the number of classifiers grew. With 21 classifiers as well as 100 million cases to classify, the best results were achieved with a classification accuracy of 0.81.

Twitter-Sentiment Analysis using Lexicon-based-Approaches (Unsupervised-Methods)

For sentiment analysis, lexicon-based methods rely on knowing that a text sample's polarity may be determined on the basis of its words' polarity. It's possible that this simplistic technique will fall short due to the fact that many nuances of real languages (like how close the negation is to a conjunction) aren't considered. Consequently, lexicon-based approaches have been proposed [16] to determine the mood of any given tweet T, that started by breaking down the tweet into a number of "small-scale phrases", like those indicated by the part signs appearing in the text [16]. It's a micro-phrase whenever there's a part signal in the text: adverbs as well as conjunctions are part of the part signal. After the splitting process, the polarity of every smaller micro-phrase was added to determine the emotion of a tweet. This was the moment at which the score was uniformed over the whole Tweet. Using micro-phrases to invert the polarity of the material was a simple and effective strategy in this case.

A large volume of unlabeled data from the social networks was cheap to obtain, but recognizing the sentiment labels of this data was exceedingly expensive, according to [17]. Thus, "unsupervised sentiment analysis" methodologies were required. Furthermore, as the amount of unlabeled data on social media grows, unsupervised learning approaches are becoming increasingly important.

Twitter-Sentiment-Analysis using the Hybrid-Methods

[18] devised a hybrid approach to analyze tweets for emotional content. Their system also used three categorization methods: "machine learning, rule-based, and lexicon-based". For machine learning, [18] employed the "SentiStrength dictionary and the SVM classifier". Hybrid systems beat all other classifiers tested, with a Fmeasure of 0.56 as opposed to the classifiers that used rules or lexicons or SVMs, which each had a Fmeasure of 0.14."

"Twitter opinion mining" (TOM) was introduced by [19] to classify tweets' sentiment. "Emoticon analysis, SentiWordNet analysis, and an improved polarity classifier" were all part of the suggested hybrid approach in [19]. Different pre-processing approaches and multiple sentiment algorithms were included in the suggested classifier to reduce the issue of sparse data. The suggested approach produced an "average harmonic mean" of 83.3 percent in testing involving six datasets.

5. APPLICATIONS OF SENTIMENT ANALYSIS

"Sentiment Analysis" has many applications in various Fields.

"Applications that use Reviews from Websites": Almost everything can be found online these days, thanks to the Internet's vast repository of user evaluations and opinions. This covers product evaluations, criticism on political topics, and comments on service. Because of this, it is necessary to have a system which can extract feelings from the reviews of a certain product or service. Let us automate the process of providing feedback or ratings for the product, item, etc. Both consumers and suppliers would benefit from this.

"Applications as a Sub-component Technology": The use of a sentiment prediction system in recommender systems is also possible. Data is connected to the nearest data scale tree by averaging. Since the decision tree doesn't "push" data into a category in an unnecessary manner, this is a benefit. Abuseful language and other bad aspects are common in internet communications. It's easy to see them if you're able to recognise a negative feeling and take action against it.

"Applications in Business Intelligence": People increasingly like to read evaluations of things accessible online before making a purchase, according to research. It's very uncommon for firms to rely on internet feedback to determine whether or not a product is a success or failure. As a result, corporations use Sentiment Analysis. To better their goods and, as a result, their reputation and consumer happiness, companies are interested in gleaning sentiment from internet evaluations.

"Applications across Domains": Sentiment Analysis, which identifies patterns in the expression of human emotions, has recently been used in research in subjects such as sociology, medicine, and sports.

"Applications In Smart Homes": In the future, smart houses are expected to be a common feature in homes. Humans will be able to manage any element of their house with a tablet device in future. The "Internet of Things" (IoT) has recently been the subject of a lot of study. The "Internet of Things" (IoT) will also include sentiment-analysis. As an example, a user's present mood or state of mind might be used to adjust the atmosphere of the house to create a serene and relaxing environment. Using Sentiment Analysis to anticipate future trends is also possible. Data on sales trends and consumer satisfaction may be gleaned by monitoring public opinion.

6. CONCLUSION

Diverse Twitter sentiment analysis methods were explored in this article including "machine learning, ensemble approaches and lexicon-based approaches". We also looked at "hybrid and ensemble methods" for analyzing Twitter sentiments. Some studies demonstrate that algorithms like SVM and naive Bayes, that are based on lexicons and take little effort in human-labeled documents, are the most accurate and should be considered the baseline learning methods. To further our research, we also explored how different characteristics affect a classifier. In order to improve "sentiment classification accuracy and adaptability" across a wide range of subjects and languages, we may investigate merging "machine learning" methods with the "opinion lexicon technique".

7. REFERENCES

[1] R. Sharma, S. Nigam, and R. Jain, "Polarity detection at sentence level," International Journal of Computer Applications, vol. 86, no. 11, 2014.

[2] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[3] R. Parikh and M. Movassate, "Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

[4] Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[5] L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010:

Poster Volume,pp. 36-44.

[6] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[7] K. P. Murphy, "Naive bayes classifiers," University of British Columbia, vol. 18, 2006.

[8] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," Comput. Linguist., vol. 22, no. 1, pp. 39-71, 1996.

[9] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support vector machine," Teori dan Aplikasinya dalam Bioinformatika, Ilmu Komputer. com, Indonesia, 2003.

[10] A. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," in Advanced Computing (ICoAC), 2016 Eighth International Conference on, 2017: IEEE, pp. 72-76.

[11] A. Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet, "Web Opinion Mining: How to extract opinions from blogs?," in Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, 2008: ACM, pp. 211-217.

[12] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002: Association for Computational Linguistics, pp. 79-86.

[13] J. Khairnar and M. Kinikar, "Machine learning algorithms for opinion mining and sentiment classification," International Journal of Scientific and Research Publications, vol. 3, no. 6, pp. 1-6, 2013.

[14] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences, vol. 181, no. 6, pp. 1138-1152, 2011/03/15/ 2011.

[15] J. Lin and A. Kolcz, "Large-scale machine learning at twitter," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012: ACM, pp. 793-804.

[16] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexiconbased approaches for sentiment analysis of microblog posts," Information Filtering and Retrieval, vol. 59, 2014.

[17] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in Proceedings of the 22nd international conference on World Wide Web, 2013: ACM, pp. 607-618.

[18] P. Balage Filho and T. Pardo, "NILC_USP: A hybrid system for sentiment analysis in twitter messages," in Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, vol. 2, pp. 568-572.

[19] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," Decision Support Systems, vol. 57, pp. 245-257, 2014