Twitter Sentiment Analysis Approaches : A Survey

Amol S. Gaikwad[1]

[1]Lecturer, Department of Computer Engineering, Government Polytechnic, Gadchiroli, Maharashtra, India

---

Abstract - Due to internet revolution huge volume of opinionated information is generated online each day. Popular micro-blogging site twitter is one such valuable source of opinionated information. The sentiment analysis of twitter posts can help to gauge people's opinion on several topics ranging from products, politics to sports and culture. This paper explores various approaches to sentiment analysis of twitter posts. This paper describes several popular and recent trends in twitter sentiment analysis including machine learning, lexicon based, ontology based, and other unsupervised analysis methods. Work done by various authors on the described methods has also been introduced. This paper tries to present various approaches on one platform which saves time and efforts to study various ways of twitter sentiment analysis. It also proposes a system to classify twitter posts using combination of machine learning and ontology based approach.

Keywords: *twitter sentiment analysis, lexicon, label propagation, supervised, ontology*

1. INTRODUCTION

Sentiment analysis means identifying opinion or orientation of peoples towards particular product or entity from some text document. Various text sources are used for extracting sentiments like blogs posts, review sites, e-commerce sites, micro-bogging sites, comments, professional websites. Due to internet revolution huge amount of opinionated information is generated online each day and the quantity is increasing day by day with ever expanding web. The information generated can be effectively utilised to gain insight of people's opinion on various topics of interest. Sentiment analysis involves concepts from natural language processing. It is a subfield of natural language processing. Sentiment analysis aims at building a system which collects opinionated data, performs data cleaning, extracting opinion texts and finally determining the polarity of text.

Sentiment analysis finds applications in various domains. Sentiment analysis can help predict sales, in marketing it helps in judging success of ad campaign or new product launch, which product and services are currently popular among the masses can also be identified. Businesses can keep track of their rivals and adjust their strategy in time. Sentiment analysis of tweets find vital application in the field of economic and financial modelling. People can make their choices base on the current trends. Political parties can use sentiment analysis to track their popularity among masses. Thus sentiment analysis can influence various domains and can serve a cause for big gains.

2. MICRO-BLOGGING SITE TWITTER

Twitter has seen unprecedented growth in recent years. Twitter audiences varies from common man to celebrities, company representatives, politicians, and even countries presidents and prime ministers. Therefore, it is possible to collect text posts of users from different social and interests groups. Users from many countries are active on twitter. They create status messages called as tweets. These tweets contains users views on various topics. Twitter has about 500 million registered users and about 400 million messages per day. Millions of tweets are generated in short span of time hence twitter has become an epicentre of sentiment analysis activities. Twitter represents one of the largest and most dynamic datasets of user generated content. Twitter has several unique characteristic which differs from other data sources. Twitter's unique characteristics are mentioned as below.

Message length : Twitter messages called tweets are restricted to maximum length of 140 characters. As messages are short users directly express their views in more precise and compact manner.

Twitter API : Twitter data or tweets are publicly available through twitter API. Interested people can easily use twitter data in their applications after registering with the official website of twitter.

Domain *:* People can tweet on almost everything thus making twitter perfect for sentiment analysis of all domains.

Language : Language of tweets is significantly different from other sites. People can use very informal language while posting their tweets, misspellings, slangs, emoticons are quite normal and may appear frequently.

Real Time Data : Huge amount of data is generated in real time. This real time data can represent latest choice of people on various topics.

3. APPROACHES

Large amount of research has already been done in the field of sentiment analysis. Sentiment analysis can be done at blog level, document level, sentence level and phrase level. Blog and review level sentiment analysis involves large pieces of text whereas sentence level and phrase level comprises relatively short pieces of text. Twitter sentiment analysis involves analysing short pieces of text called as tweets. Following sections describe various approaches to twitter sentiment analysis.

3.1 Lexicon Based

Lexicon based approach makes use of dictionary called as opinion lexicon. This dictionary consists of list of positive, negative and neutral words. Various such opinion lexicons are available online like Bing liu opinion lexicons, OpinionFinder subjectivity lexicons. OpinionFinder is a system developed at University of Pittsburg, Cornell and Utah, that processes documents and automatically identifies subjective sentences as well as various aspects of subjectivity within sentences, including agents who are sources of opinion, direct subjective expressions and speech events, and sentiment expressions. Bing liu opinion lexicons consist of list of 2006 positive words and 4783 negative words. OpinionFinder lexicons contains collection of terms with sentiment labelled as either 'strong positive', 'weak positive', 'strong negative', 'weak negative'. Interested people can also create their own list of opinion lexicons, moreover domain specific opinion lexicons can also be created. We can also add more words to existing lexicons making it even more rich. Lexicon based sentiment analysis of twitter posts involves identifying number of positive and negative words in the given tweet. The difference between the count of positive and negative words is used as a sentiment score of the given tweet. If the difference is positive the tweet is considered as positive else if the difference is negative the tweet is assigned negative polarity. The tie is broken in favour of majority class or neutral class. Twittratr is a website that uses lexicon based approach to classify tweets, Twittratr's list of opinion words is publicly available.

Lexicon based method is relatively easy to implement. Lei Zhang, Ridhhiman Ghosh, Bing Liu [1] have combined lexicon based with learning based methods. In their work they have used lexicon based methods to produce automatic labelled training data for training machine learning classifiers. They have introduced lexicon based method as an automatic substitute for manually labelling of training data. Akshi Kumar and Teeja Mary Sebastian [9] proposed a hybrid approach using dictionary based and corpus based methods. In their work overall sentiment score of tweet is determined by individual sentiment score of adjectives, verbs, adverbs, emoticons, exclamation marks, capital words. They used POS tagger to identify adjectives, verbs and adverbs in the tweets. Semantic score of the adjectives was determined by the corpus based methods. Log linear Regression model with linear predictor was used. The semantic orientation of the adverbs was determined using dictionary based approach. WordNet was used to determined strengths of adverbs and verbs. WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. Opinion indicators like emoticons were also assigned sentiment score. The overall tweet sentiment is then calculated using a linear equation which incorporates all opinion indicators in the given tweet. Reynier, Yoan, Andres [15] proposed unsupervised system SSA-UO for sentiment analysis of twitter. This system uses a contextual sentiment classification method based on coarse-grained word sense disambiguation, using WordNet and a coarse-grained sense inventory (sentiment inventory) built up from SentiWordNet. SentiWordNet is a lexical resource for opinion mining which is publicly available. Each emotion in tweet is manually annotated with an emotion word and polarity value. Sentiments are divided into five classes highly positive, positive, highly negative, negative and objective. The final polarity of the tweet is determined using rule based classifier. The rule based classifier classifies tweets as positive, negative or neutral. The polarity of tweet is determined from the scores of positive and negative words it contains. Table 1 shows the evaluation of SSA-U0 system on two data sets of twitter and sms.

Table-1: SSA-UO Results in Polarity Classification

Runs	Dataset	F1	Constrained runs Rank
twitter-1	Twitter	50.17	25(35)
Sms-1	SMS	44.39	22(28)

3.2 Supervised Learning Methods

Supervised Learning or machine learning is one of the popular and traditional methods for solving classification problems. Machine learning methods work considerably well in case of twitter sentiment analysis. In machine learning based classification two set of documents are required : training and testing set. A training set is labelled set of documents which is used by the classifier to learn differentiating characteristics of the documents and the testing set is used to validate the performance of the trained classifiers. Machine learning classifiers Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM), have achieved great success in text categorization. K-nearest neighbour, ID3, C5, centroid classifier, winnow classifier are other well-known machine learning methods in natural language processing area.

Naive Bayes classifier is a simple but very power classification algorithm. It is widely used algorithm for document classification. The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The naive part of such a model is the assumption of word independence. The simplicity of this assumption makes the computation of Naive Bayes classifier far more efficient. A lot of work has been done on sentiment analysis of twitter using Naive Bayes algorithm. Suhaas Prasad [12] used several variations of Naive Bayes classifier along with feature selection methods. The different Naive Bayes approaches used include unigram multinomial, unigram Bernoulli, unigram Bernoulli with chi-square features, bigram multinomial, linear interpolated bigram and unigram, bigram back-off, and unigram weigh complemented multinomial. His work shows that classifiers perform better when dealing with just two classes. James Spencer and Gulden Uchyigit [7] used Sentimator tool for sentiment analysis of twitter data. Sentimator is a web based tool which uses Naive Bayes classifier to classify live twitter data based on positivity, negativity and objectivity. Table 2 shows the results of Sentimator using unigrams and Table 3 shows results for bigrams. Ravi Parikh and Matin Movassate [13] implemented two unigrams Naive Bayes models namely Multinomial unigram and Bernoulli unigram Naive Bayes models. Go, Bhayani, Huang [2] have also used Naive Bayes classifier to classify twitter tweets with unigram and unigram-bigram variations. They achieved 81.3 % and 82.7 % accuracy rate respectively. They noted accuracy increases for Naive Bayes classifier using unigram-bigram combinations.

Table-2: Results for Unigrams

Sentiment	Number of Samples	Correctly Identified	False Positives
Positive	108	37	9
Negative	75	45	45
Objective	33	19	61

Table-3: Results for Bigrams

Sentiment	Number of Samples	Correctly Identified	False Positives
Positive	108	47	16
Negative	75	47	44
Objective	33	19	43

A Maximum Entropy classifier is a probabilistic, feature-based model that favours the most uniform distribution of classes that adhere to a specific set of constraints that are determined based on the training data. The intuition of the Maximum Entropy model is to use a set of user-specified features and learn appropriate weights. Maximum Entropy models are feature-based models. In a two-class scenario, it is the same as using logistic regression to find a distribution over the classes. Maximum Entropy makes no independence assumptions for its features, unlike Naive Bayes. This means we can add features like bigrams and phrases to Maximum Entropy without worrying about features overlapping. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization so as to maximize the conditional probability. Go, Bhayani, Haung [2] implemented Maximum Entropy to classify twitter data. They have used Standford Classifier to perform Maximum Entropy classification. For training the weights they used conjugate gradient ascent and added smoothing (L2 regularization). They experimented with unigram and unigram-bigram approaches and achieved 80.5 % and 82.7 % accuracy rate respectively.

Support Vector Machine (SVM) are universal learners, important property of SVM is that their ability to learn can be independent of dimensionality of feature space. Go, Bhayani, Haung [2] have used SVM with linear kernel to classify tweets. Their input data consists of two sets of vectors. Each entry in the vector corresponds to presence of feature. If the feature is present, the value is 1, but if the feature is absent, then the value is 0. They have used feature presence as opposed to count which speeds up the overall process. Go, Bhayani, Huang [2] have used unigram, bigram and unigram-bigram variations of SVM to classify tweets. With the unigram approach they achieved 82.2 % accuracy for unigram-bigram approach accuracy of SVM declined to 81.6 %. Using only bigram features makes the feature space sparse. Table 4 summaries the accuracy of various classifiers using unigrams, bigrams, unigrams-bigrams.

Table-4: Classifiers Accuracy

Feature	Keyword	Naive Bayes	MaxEnt	SVM
Unigram	65.2	81.3	80.5	82.2
Bigram	N/A	81.6	79.1	78.8
Unigram+Bigram	N/A	82.7	83.0	81.6
Unigram+POS	N/A	79.9	79.9	81.9

3.3 Ontology Based

An ontology can be defined as an ‘‘explicit, machine-readable specification of a shared conceptualization’’ (Studer, Benjamins, & Fensel, 1998). An ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse. It is thus a practical application of philosophical ontology, with a taxonomy. Ontologies are used for modelling the terms in a domain of interest as well as the relations among these terms and are now applied in various fields, like agent and knowledge management systems and e-commerce platforms. Other applications include natural language generation, intelligent information integration, semantic-based access to the Internet and extracting information from texts.

The concept of ontology can be effectively applied to sentiment analysis of twitter data. Methods like machine learning and lexicon based gives you sentiment score for entire sentence. There may be different sentiments for different features of object in the same sentence, a single sentiment score for whole sentence cannot give insight of different sentiments. Ontology based techniques can be used for more fine grained analysis of twitter posts. Ontologies can be created for domains under considerations for sentiment analysis. Syed Zeeshan Haider [6] have refined the ontologies for mobile phones created by Yaakub into three categories smart phones, wet and dirty mobile phones and simple mobile phones. The Knowledge represented in such domain ontologies can be used to extract relevant tweets and then perform more fine grained analysis of tweets. Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, Nick Bassiliades [4] introduced an ontology based sentiment analysis of twitter posts. In their work they created domain ontology by extracting objects and attributes from retrieved tweets. They have used 'smartphone' as particular domain. Various semi-automatic learning techniques like OntoGen or manual methods can be used to create domain ontology. After creating domain ontology they used it to extract tweets that contain objects and attributes. The extracted tweets are then submitted to OpenDover for assigning sentiment score to the tweets. Figure 1 depicts the resulting sentiment values of each object-attribute pair of smartphone. OpenDover is a sophisticated webservice that allows you to extract the next generation semantic features within your blogs, content management systems, websites or other numerous applications. K. Vithiya Ruba and Venkatesan [3] build a custom sentiment analysis tool based on ontology for twitter posts. They used laptops as a domain. They extracted tweets relevant to the feature of a particular domain here laptops and then assigned score for each feature using sentiment package of R. Ontology is created using protege software and querying of object-attribute pair is done using ontoCAT package of R. Figure 2 shows the results for battery attribute of laptops of four different companies.

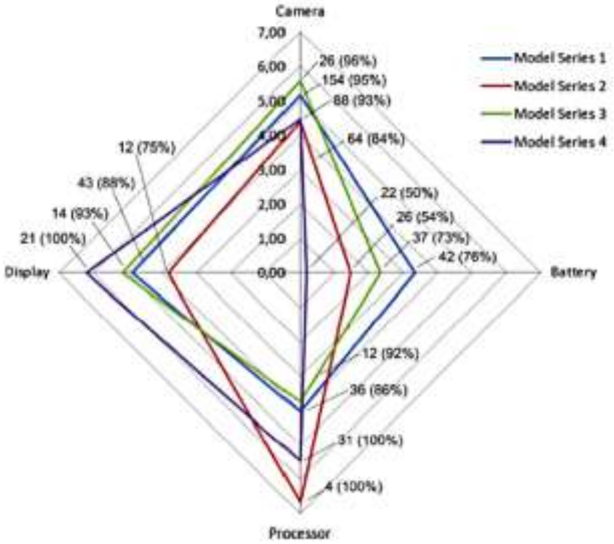

Fig -1: Sentiment value of smartphone attributes.

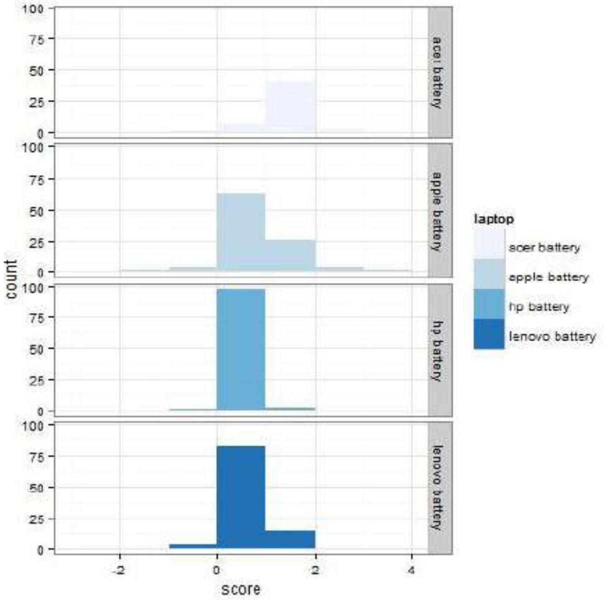

Fig -2: Comparable results of laptops.

3.4 Label Propagation (LP)

Label propagation is another way of performing twitter sentiment analysis. In label propagation method a weighted graph G = (V,E,W) is created where V is set of vertices comprised of users, tweets, and other features, E is a set of edges connecting vertices and W is set of weights associated with the edges. After creating such a graph structure, a label distribution is seeded at an initial subset of nodes and then spread across the graph until convergence.

Cristopher Johnson, Parul Shukla, Shilpa Shukla [14] proposed a semi-supervised approach using LP and Maximum Entropy. Their LP graph consisted of nodes representing users, unigram, bigrams, hashtags and tweets. Users are also connected to each other if they retweeted. The weights of user to tweet edges are fixed to a constant weight . The weights of tweet to unigram and bigram edges were related to the relative frequency ratio of the unigram or bigram between the training set and the Emoticon dataset. The tweet to hashtag edges also used a fixed edge weight. All unigrams contained in the OpinionFinder lexicon were seeded with a seed proportional to the term's lexical sentiment. All tweets were seeded with probabilistic predictions obtained from emoticon trained Maximum Entropy classifier. Modified Absorption based label propagation algorithm provided by the Junto software toolkit was used to converge the graph. Table 5 shows accuracy of different methods including LP methods. A

Table-5: Sentiment Prediction Accuracies

Method	Accuracy(%)
Lexical Ratio	43
Maximum Entropy Classifier	60
LP with no Retweet/Reply edges	75
LP with Retweet/Reply edges	67
LP withno Retweet/Reply edges and seeded hashtags	78

similar approach was also suggested by Michael Speriosi, Nikita Sudan [5], they have also used combination of LP and Maximum Entropy to classify twitter polarity.

4. CHALLENGES

Sentiment analysis in general do poses a lot of challenges. Sentiment analysis of micro-blogging site twitter does bring its own challenges in addition to traditional ones. Twitter in involves extensive use of jargons in authors tweets. The language used in twitter is extremely informal, thus large numbers of misspelled words, slang words, emoticons, appear in tweets which renders sentiment analysis difficult. Opinion words that appear positive in one context may appear negative in other contexts. People express their opinions in many different ways which is easy for humans to recognise but difficult for machines to parse. If one has to build a real world application then one must deal with opinion spam. Opinion spams are fake views that intentionally try to mislead leaders and their opinions.

Lexicon based approach can give good precision but low recall. Lexicon based methods totally rely on presence of opinion words in tweets to detect its polarity. The opinion words in the tweets should match with opinion lexicons. Twitter posts contains lots of slang words, misspelled words so using standard words dictionary may not be sufficient. Such words despite expressing opinions may not match any words in the dictionary. Emoticons, colloquial expressions, abbreviations, etc. are frequently used in tweets. These expressions contains valuable sentiments but they do not exist in general opinion lexicon. The writing style on twitter changes with time, this poses another challenge for lexicon based methods. Machine learning methods require huge amount of labelled training data. Manually labelling such huge data is time consuming task. Collecting positive and negative data can be achieved automatically to large extent using emoticons, but no efficient method exists for collecting neutral tweets. Machine learning methods are not suitable for real time application as training the algorithm consumes lot of time. Label Propagation approach has two main disadvantages, first due to users privacy it is not possible to construct their followers graph, secondly due to rapid growth of twitter user graph tends to grow quickly making it difficult to manage the graph. Lexicon based and machine learning methods give sentiment score for entire sentence, They fail to detect different sentiments about different features present in the same sentence. Ontology based approach gives more fine grained analysis of tweets but ontologies are domain specific, creating general ontology which suits for all domains is a difficult task. Moreover creating ontologies automatically without human intervention is still a distant dream. Unsupervised methods are time efficient as they do not require training but they are less accurate compared to machine learning methods.

5. CONCLUSION

Sentiment analysis of micro-blogging site twitter with its potential applications has become a hot research topic in recent years and is expected to continue in coming years. This paper introduced and summarised various approaches to twitter sentiment analysis. It also introduced the work of various authors on related topic. Twitter sentiment analysis approaches ranged from supervised to unsupervised and from more general to ontology based domain specific. Each approach has its own pros and cons. Various methods can also be integrated to form a more accurate analysis system. A similar approach is presented at the end of the paper which proposes combination of ontology based and machine learning methods to classify tweets.

REFERENCES

[1] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, Bing Liu. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. HPL Laboratories, HPL 2011.

[2] Go, Bhayani, Huang. Twitter Sentiment Classification using Distant Supervision. 2009.

[3] K. Vithiya Ruba and D. Venkatesan. Building a Custom Sentiment Analysis Tool based on an Ontology for Twitter Posts. In IJST 2015. SASTRA University Tamil Nadu.

[4] Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, Nick Bassiliades. Ontology-based sentiment analysis of twitter posts. In Elsevier 2013.

[5] Speriosu, Sudan, Upadhay, Baldridge. Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. University of Texas in Austin.

[6] Syed Zeeshan Haider. AN ONTOLOGY BASED SENTIMENT ANALYSIS a case study. University of Skovde.

[7] James Spencer and Gulden Uchyigit. Sentimentor: Sentiment Analysis of Twitter

Data. School of Computing, Engineering and Mathematics University of Brighton, Brighton, BN2 4GJ.

[8] G.Vinodhini, RM.Chandrasekaran. Sentiment Analysis and Opinion Mining: A Survey. In IJARCSSE 2012. Annamalai University, Annamalai Nagar, India.

[9] Akshi Kumar and Teeja Mary Sebastian. Sentiment Analysis on Twitter. In IJCSI 2012. Department of Computer Engineering, Delhi Technological University Delhi, India.

[10] Bing Liu. Sentiment Analysis a Multi-faceted Problem. IEEE Intelligent systems 2010. Department of Computer Science University of Illinois at Chicago.

[11] Aamera Z. H. Khan, Dr. Mohammad Atique, Dr. V. M. Thakare. In IJARCSSE 2015. Amravati University, Amravati, India.

[12] Suhaas Prasad. Micro-blogging Sentiment Analysis Using Bayesian Classification Methods.

[13] Ravi Parikh and Matin Movassate. Sentiment Analysis of User Generated Twitter Updates using Various Classification Techniques. 2009.

[14] Christopher Johnson, Parul Shukla,ShilpaShukla. On Classifying Political Sentiments of tweets. Department of Computer Science, University of Texas in Austin.

[15] Reynier Ortega, Adrian Fonseca, Yoan Gutierrez, Andres Montoyo. SSA-UO: Unsupervised Twitter Sentiment Analysis. University of Oriente, University of Matanzas, Cuba, University of Alicante,Spain.