Twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach

Riya Suchdev
Computer Engineering
VES Institute of Technology, University of Mumbai, Mumbai, India.

Pallavi Kotkar
Information Technology VES Institute of Technology, University of Mumbai, Mumbai, India.

Rahul Ravindran
Computer Engineering VES Institute of Technology, University of Mumbai, Mumbai, India.

Sridhar Swamy
Computer Engineering VES Institute of Technology, University of Mumbai, Mumbai, India.

ABSTRACT
Sentiment analysis is mainly concerned with identifying and classifying opinions or emotions that are expressed within a text. These days, sharing opinions and expressing emotions through social networking websites has become very common. Therefore, a large amount of data is generated each day, on which mining can be effectively performed to retrieve quality information. Sentiment analysis on such data can prove to be instrumental in generating an aggregated opinion on certain products. Twitter sentiment analysis often becomes a difficult task due to the presence of slangs and misspellings. Also, we constantly encounter new words, which makes it more difficult to analyze and compute the sentiment as compared to the usual sentiment analysis. Twitter restricts the length of a tweet to 140 characters. Thus, extracting valuable information from short texts is yet another challenge. Knowledge-based approach and machine learning can contribute considerably towards the analysis of sentiments from tweets. In this paper, we analyze people's sentiments in their tweets about certain companies. Computing a basic sentiment score and then classifying them as positive or negative would help to serve the company by providing them critical reviews about their products from people worldwide.

General Terms
Data Mining, Twitter API, Sentiment Score, Sanders Analytics, Hybrid Approach

Keywords
Sentiment analysis, Knowledge-based approach, Machine Learning, Twitter, Sentiment score

1. INTRODUCTION
The age of the internet has changed the way people express their ideas. With the ever-increasing popularity of social networking, microblogging and blogging websites, a large amount of data is generated every day. These social networking websites depend largely on the user-generated content. Typically, when people intend to purchase a product, they browse through a lot of websites to gain some information about the products before they make their purchase. They take into consideration the available reviews and ratings of these products on these websites before making purchases. The amount of information is inordinate for a normal person to analyse it using naive techniques. Thus, in order to make this process efficient and to automate it, several sentiment analysis techniques are used. Symbolic techniques or knowledge-based approach and machine learning techniques are usually used to develop such models. In knowledge-based approach, we require a comprehensive database that contains pre-defined emotion information and an efficient knowledge representation for identifying sentiments. Machine learning approach makes use of training data set to correctly identify the emotions of each word. Thus, this technique does not require the database of words like knowledge-based approach and therefore, is better. Using a mixture of both this techniques, predominantly machine learning technique, we can build a hybrid model which will be capable of analyzing sentiments of almost any text. A detailed overview of various techniques is discussed in this paper.

A sentiment summarization system is a system that takes the documents that have to be analyzed as input, and generates a detailed document summarizing the opinions in the input documents. This allows both public and the company to have access to summarized details pertaining to a certain product. There are two ways to summarize the data: (i) Extractive summarization where only the important sentences are used to depict the summarized portion of a document (ii) Abstractive summarization, which requires rigorous analysis of text and produces an abstract which may even include sentences that may not have been explicitly stated in the original data. Such summarized sentiment analysis reports provide a great deal of information not only to companies but also to the customers which will help them to judge a product wholly. There has already been a great amount of research on various methods to use web technologies to maximize the benefits of customers as well as companies in the market place [1].

Sentiment analysis is usually conducted at different levels varying from coarse-level to fine-level. Coarse-level analysis is mainly concerned with finding the sentiment score of the entire document whereas fine-level deals with attribute level. Sentence-level sentiment analysis comes in between these two [2]. There are many researches in the area of sentiment analysis of user reviews. The performance of sentiment analyzer is largely dependent on the topic. As a result, we cannot determine which classifier is the best.

Sentiment analysis in Twitter is quite difficult due to many reasons. Short texts owing to the character limit is an acute issue. Presence of slangs, emoticons and misspellings in the tweets makes an additional step of pre-processing obligatory. There are several methods for feature extraction that can be used to collect pertinent features from the text. These feature extraction methods can be effectively applied to tweets too. Feature extraction is done in two phases: First phase is the

extraction of data from Twitter. The extracted feature vector is converted into normal text. In the second phase, feature extraction is performed again on this text to get more feature vectors. By performing sentiment analysis on a specific domain, it is possible to identify the influence of domain information in choosing a feature vector.

2. RELATED WORK

There are two basic methodologies to detect sentiments from text. They are Symbolic techniques and Machine Learning techniques [3]. The next two sections deal with these techniques.

a) Symbolic Techniques

Much of the research in unsupervised sentiment classification using symbolic techniques makes use of available lexical resources. Turney [4] used bag-of-words approach for sentiment analysis. In this approach, relationships between the individual words are not considered and a document is represented as a mere collection of words. To determine the overall sentiment, sentiment of every word is determined and this value is combined with some aggregation functions. He determined the polarity of a review based on the average semantic orientation of tuples extracted from the review where tuples are phrases having adjectives or adverbs. He determined the semantic orientation of tuples using the search engine Altavista.

Kamps et al. [5] used the lexical database WordNet [6] to determine the emotional content of a word along different dimensions. They developed a distance metric on WordNet and determined the semantic orientation of adjectives. WordNet database consists of words connected by synonym relations. Baroni et al. [7] developed a system using word space model formalism that overcomes the difficulty in lexical substitution task. It represents the local context of a word along with its overall distribution. Balahur et al. [8] introduced EmotiNet, a conceptual representation of text that stores the structure and the semantics of real events for a specific domain. EmotiNet used the concept of Finite State Automata to identify the emotional responses triggered by actions. One of the participant of SemEval 2007 Task No. 14 [9] used coarse-grained and fine-grained approaches to identify sentiments in news headlines. In coarse-grained approach, they performed binary classification of emotions whereas in fine-grained approach, they classified emotions into different levels. Knowledge-based approach is found to be difficult due to the requirement of a huge lexical database. Social network generates huge amount of data every second, which is significantly larger than the size of available lexical databases. Therefore, sentiment analysis often becomes arduous and erroneous.

b) Machine Learning Techniques

Machine Learning techniques use a training set and a test set for classification. Training set contains input feature vectors and their corresponding class labels. Using this training set, a classification model is developed which tries to classify the input feature vectors into corresponding class labels. Subsequently, a test set is used to validate the model by predicting the class labels of unseen feature vectors.

A number of machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify reviews [10]. Some of the features that can be used for sentiment classification are term presence, term frequency, negation, n-grams and parts of speech [11]. These features can be used to find out the semantic orientation of words, phrases, sentences and documents. Semantic orientation is the polarity which may be either positive or negative.

Domingos et al. [12] found that Naive Bayes works well for certain problems with highly dependent features. This is surprising as the basic assumption of Naive Bayes is that the features are independent. Zhen Niu et al. [13] introduced a new model in which efficient approaches are used for feature selection, weight computation and classification. The new model is based on Bayesian algorithm. Here, weights of the classifier are adjusted by making use of representative feature and unique feature. 'Representative feature' is the information that represents a class and 'Unique feature' is the information that helps in distinguishing classes. Using those weights, they calculated the probability of each classification and thus improved the Bayesian algorithm.

Barbosa et al. [14] designed a 2-step automatic sentiment analysis method for classifying tweets. They used a noisy training set to reduce the labeling effort in developing classifiers. Firstly, they classified tweets into subjective and objective tweets. After that, subjective tweets are classified as positive and negative tweets. Celikyilmaz et al. [15] developed a pronunciation based word clustering method for normalizing noisy tweets. In pronunciation based word clustering, words having similar pronunciation are clustered and assigned common tokens. They also used text processing techniques like assigning similar tokens for numbers, HTML links, user identifiers and target organization names for normalization. After performing normalization, they used probabilistic models to identify polarity lexicons. They performed classification using the BoosTexter classifier with these polarity lexicons as features and obtained a reduced error rate.

Wu et al. [16] proposed an influence probability model for Twitter sentiment analysis. If @username is found in the body of a tweet, it is influencing action and it contributes to influencing probability. Any tweet that begins with @username is a retweet that represents an influenced action and it contributes to an influenced probability. They observed that there is a strong correlation between these probabilities.

Pak et al. [17] created a twitter corpus by automatically collecting tweets using Twitter API and automatically annotating those using emoticons. Using that corpus, they built a sentiment classifier based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features. In that method, there is a chance of error since emotions of tweets in training set are labeled solely based on the polarity of emoticons. The training set is also less efficient since it contains only tweets having emoticons.

Xia et al. [17] used an ensemble framework for sentiment classification. Ensemble framework is obtained by combining various feature sets and classification techniques. In their work, they used two types of feature sets and three base classifiers to form the ensemble framework. Two types of feature sets are created using Part of speech information and Word-relations. Naive Bayes, Maximum Entropy and Support Vector Machines are selected as base classifiers. They applied different ensemble methods like Fixed combination, Weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy. Certain attempts are made by some researches to identify the public opinion about movies, news etc. from Twitter posts. V.M. Kiran et al. [17] utilized the information from other publicly available databases like IMDB and Blippr after proper modifications to aid Twitter sentiment analysis in movie domain.

3. PROPOSED WORK

Initially, we acquire a dataset and partition the dataset into training and test data. The major approach used to classify the tweets is the machine learning technique. The tweets may be rife with slang words and misspellings. So, we need to perform sentence-level analysis for all these tweets. This is done in three phases. In the first phase, pre-processing is done. This is done mainly to eliminate the slang words, misspellings and other faults. In the second phase, a feature vector is created using relevant features. Finally, using different classifiers, we will be able to classify the tweets as positive, negative or neutral.

A) Acquiring the dataset

We use the Sanders analytics dataset to perform necessary actions. Sanders analytics dataset consists of a total of 5600 tweets containing tweets of companies like Apple, Google and Microsoft. The dataset is labeled and therefore, we know exactly which tweets are positive, negative, neutral and irrelevant.

B) Pre-processing tweets

Keyword extraction is difficult in Twitter due to misspellings and slang words. So to avoid this, a pre-processing step is performed before feature extraction. Pre-processing steps include removing URLs, avoiding misspellings and slang words. Misspellings are avoided by replacing repeated characters with 2 occurrences. Slang words contribute considerably to the emotion of a tweet. So they can't be simply removed. Therefore, a slang word dictionary is maintained to replace slang words occurring in tweets with their associated meanings. Domain information contributes much to the formation of slang word dictionary. Also, we use a technique in which if the overall sentiment of the tweet is obtained, we will be able to find out the sentiment score of the new term by just looking at its relative position in the sentence.

C) Creation of feature vector

Feature extraction is done in two steps. In the first step, Twitter-specific features are extracted. Hashtags and emoticons are the relevant Twitter-specific features. Emoticons can be positive or negative. Therefore, they are assigned different weights. Positive emoticons are assigned a weight of '1' and negative emoticons are assigned a weight of '-1'. There may be positive and negative hashtags. Therefore, the count of positive hashtags and negative hashtags are added as two separate features in the feature vector. Twitter-specific features may not be present in all tweets. So, a further feature extraction is to be done to obtain other features. After extracting Twitter-specific features, they are removed from the tweets. A tweet can be then treated as simple text. Thereafter, using unigram approach, tweets are represented as a collection of words. In unigrams, a tweet is represented by its keywords. We maintain a negative keyword list, positive keyword list and a list of different words that represent negation. Counts of positive and negative keywords in tweets are used as two different features in the feature vector. Presence of negation contribute much to the sentiment. So their presence is also added as a relevant feature.

All keywords cannot be treated equally in the presence of multiple positive and negative keywords. Therefore, a special keyword is selected from all the tweets. In the case of tweets having only positive keywords or only negative keywords, a search is done to identify a keyword having relevant part of speech. A relevant part of speech is an adjective, an adverb or a verb. Such a relevant part of speech is defined, based on their relevance in determining sentiment**.** A keyword that is an adjective, adverb or a verb shows more emotion than others. If a relevant part of speech can be determined for a keyword, then it is taken as special keyword. Else, a keyword is selected randomly from the available keywords as special keyword. If both positive and negative keywords are present in a tweet, we select any keyword having relevant part of speech. If relevant part of speech is present for both positive and negative keywords, none of them is chosen. Special keyword feature is given a weight of '1' if it is positive and '-1' if it is negative and '0' in its absence. Part of speech feature is given a value of '1' if it is relevant and '0' otherwise.

Thus, feature vector is composed of 8 relevant features. The 8 features used are part of speech (pos) tag, special keyword, presence of negation, emoticon, number of positive keywords, number of negative keywords, number of positive hashtags and number of negative hashtags.

d) Sentiment analysis

After creating the feature vector, sentiment classification is done using a combination of both knowledge-based and machine learning approach. Word-by-word sentiment analysis using knowledge-based approach is used along with different classifier techniques using feature vector.

3.1 Computing new word Sentiment

We derive day-to-day sentiment scores by counting positive and negative messages. Positive and negative words are defined by the subjectivity lexicon from OpinionFinder, a word list containing about 1,600 and 1,200 words marked as positive and negative respectively (Wilson, Wiebe, and Hoffmann 2005). We do not use the lexicon's distinctions between weak and strong words.

A message is defined as positive if it contains any positive word, and negative if it contains any negative word. (This allows for messages to be both positive and negative.) This gives similar results as simply counting positive and negative words on a given day, since Twitter messages are short.

A major issue when using such short messages is the presence of misspellings, emoticons, links and other unnecessary content. The pre-processing stage helps us remove these shortcomings considerably and makes the resultant text clean. In order to compute the sentiment score for a new word, we simply work the other way around. Once we have the score of the text, we simply associate the words with the general feel of the entire text. So, if soccer comes between great and happy we would associate soccer with a positive sentiment score.

3.2 Evaluation

Data contains 4685 rows. We create a subset of data by removing the irrelevant factor altogether. Resulting data set contains 3126 objects containing 5 variables. The recipe book contains the following variables:

1) Topic
2) Tweet id
3) Tweet Text
4) Tweet date
5) Sentiment

Topic: Apple, Microsoft, Google. Tweets concerning Apple, Microsoft and Google are used in the analysis of sentiment score of tweets.

Sentiment: This column actually describes the sentiment class of each tweet. Possible values are positive, negative, irrelevant and neutral.

Tweet ID: Contains the Twitter ID of the tweet
Tweet Date: Contains the date of the tweet
Tweet Text: Contains the actual text of the tweet

The original Sanders analytics data set histogram is shown below:

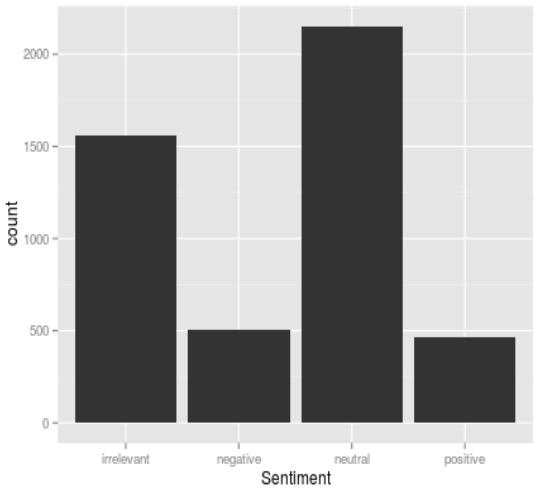

Fig 1. Histogram before eliminating irrelevant tweets

On performing a summary operation on the resultant data set, we get the following table.

Table 2. Tweet classification before elimination of irrelevant tweets

Irrelevant	Negative	Neutral	Positive
1559	509	2153	464

As we can see evident from the histogram, the dataset contains irrelevant tweets. Therefore, we eliminate them from the dataset. After eliminating the irrelevant tweets, the histogram obtained as follows:

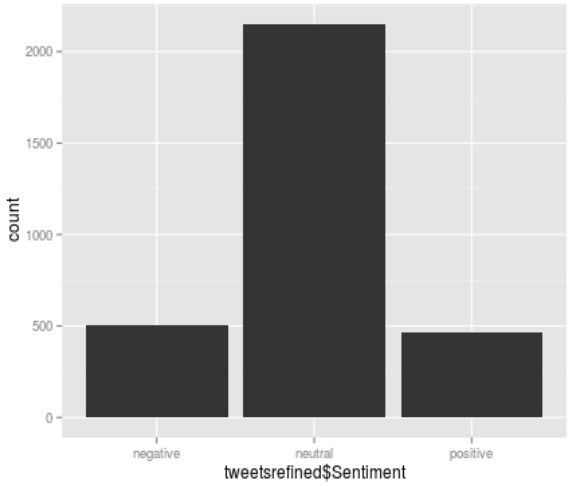

Fig 2. Histogram after eliminating irrelevant tweets

On performing a summary operation after the elimination of irrelevant tweets from the data set, we get the following table.

Table 2. Tweet classification after elimination of irrelevant tweets

Irrelevant	Negative	Neutral	Positive
0	509	2153	464

The following table demonstrates the scores that were assigned to 10 random tweets from the dataset.

Class: The label that was assigned to the tweet in the dataset
Tweet: The tweet text
Score: The sentiment score calculated by our model

Table 3. Sentiment score of 10 tweets using hybrid approach

Class	Tweet	Score
Positive	Now all @Apple has to do is get swype on the iphone and it will be crack. Iphone that is	2
Positive	@Apple will be adding more carrier support to the iPhone 4S (just announced)	3
Positive	Lmao I think @apple is onto something magical! I am DYING!!! haha. Siri suggested where to find food and where to hide a body lolol	3
Positive	Currently learning Mandarin for my upcoming trip to Hong Kong. I gotta hand it to @Apple iPhones & their uber useful flashcard apps	1
Negative	Why is #Siri always down @apple	-1
Negative	I just need to exchange a cord at the apple store why do I have to wait for a genius? @apple	-5
Negative	@apple AirDrop #fail - Immediate "declined your request." every time	-8
Negative	Dear @apple My new Air is now a notbook since your update killed #wifi #bug #destroying #productivity	-3
Neutral	Using @Apple's mobile @AirPort Utility http://t.co/TiDpaHYC	0
Neutral	#motoactiv? Methinks @apple and maybe @Nike are already prepping lawsuits	0

After training the classifier using hybrid method which combines knowledge-based approach and machine learning capabilities we get the following confusion matrix.

	Irrelevant	Negative	Neutral	Positive
Irrelevant	0	0	0	0
Negative	0	509	0	0
Neutral	0	0	2153	0
Positive	0	0	0	464

Thus, as we can see, we get an accuracy of 100%.

4. CONLCUSION

There are different Symbolic and machine learning techniques to identify sentiments from text. Machine Learning techniques are simpler and efficient than Symbolic techniques. A combination of these two techniques can be used to achieve an accuracy of 100%. In this paper we took the Sanders analytic dataset in order to analyze the tweets. After pre-processing the data we created the feature vector that is used for evaluating Twitter sentiments using Machine Learning techniques. Feature vector includes parameters like hashtags, emoticons etc. Knowledge-based approach is used to handle the other words. Slang word frequency count is measured using the backtracking approach wherein the new word is given a sentiment score corresponding to the overall sentiment score of the entire tweet. Using the hybrid approach we were able to achieve an accuracy of 100%.

5. REFERENCES

[1] Bahrainian, S.A., Bahrainian A.M, Salarinasab, M., and Dengel, A.,Implementation of an Intelligent Product Recommender System in an e-Store, Proceedings AMT-10, Int'l Conference on Active Media Technology, Toronto, Canada, 2010.

[2] Mejova, Y. (2009). Sentiment Analysis: An Overview. Comprehensive exam paper, available on http://www.cs.uiowa.edu/~ymejova/publications/Comps YelenaMejova.pdf [2010-02-03].

[3] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics ,pp. 417–424, Association for Computational Linguistics, 2002.

[4] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.

[5] C. Fellbaum, "Wordnet: An electronic lexical database (language,speech, and communication)," 1998.

[6] D. Pucci, M. Baroni, F. Cutugno, and A. Lenci, "Unsupervised lexical substitution with a word space model," in Proceedings of EVALITA workshop, 11th Congress of Italian Association for Artificial Intelligence,Citeseer, 2009.

[7] A. Balahur, J. M. Hermida, and A. Montoyo, "Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model," Affective Computing, IEEE Transactions on, vol. 3, no. 1,pp. 88–101, 2012.

[8] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14:

Affective text," in Proceedings of the 4th International Workshop on Semantic Evaluations, pp. 70–74, Association for Computational Linguistics, 2007.

[9] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey,"International Journal, vol. 2, no. 6, 2012.

[10] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1997.

[11] Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for microblog by machine learning," in Computational and Information Sciences (ICCIS),2012 Fourth International Conference on, pp. 286–289, IEEE, 2012.

[12] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36–44, Association for Computational Linguistics, 2010.

[13] A. Celikyilmaz, D. Hakkani-Tur, and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 79–84, IEEE, 2010.

[14] Y. Wu and F. Ren, "Learning sentimental influence in twitter," in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119–122, IEEE, 2011.

[15] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010, 2010.

[16] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[17] V. M. K. Peddinti and P. Chintalapoodi, "Domain adaptation in sentiment analysis of twitter," in Analyzing Microtext Workshop, AAAI, 2011.