Sentiment Analysis of Twitter Data containing Emoticons: A Survey

Jayanthi T
Assistant Professor, Department of Computer Applications
Mohandas College of Engineering and Technology, Trivandrum, Kerala

Sreeja K
Assistant Professor, Department of Computer Applications
Mohandas College of Engineering and Technology, Trivandrum, Kerala

Abstract - Sentiment Analysis (SA), also called opinion mining, as the name suggests is the field of study which analyzes people's opinions, sentiments, evaluations, appraisals, attributes and emotions towards elements such as products services, organizations, individuals, issues, events, topics, and their attributes through twitter. Microblogging (twitter) is used by people to talk about their daily chores and to share information, it is an online social networking. Micro-blogging service enables users to send and read "tweets", which are text messages limited to 140 characters. In this survey paper we propose a model that can identify the public opinion with their emotions. The main target of this survey is to give SA techniques used for emoticons and the related fields with brief details. The main contributions of this paper include the illustration of the recent trend of research in the sentiment analysis and its related areas.

Keywords: Sentiment Analysis, Microblogging, Emoticon, Lexicon based classifier, Naïve based classifier, Support Vector Machine

I. INTRODUCTION

Sentiment analysis or opinion mining refers to the type of natural language processing used to understand the moods, opinions and sentiments of the public regarding a particular product or a movie or an event. The availability of large amounts of data and the human tendency to always factor what other people think has been influential in a decision making process. This unique feature plays a vital role in deciding on matters that have financial, medical, social or other implications. Seeking second or third or many more opinions have fuelled the interest of researchers in the field of sentiment mining. With multiple reviews available for a single product and the enormous growth in the number of internet users it has become indispensable to develop a system that collects, builds, analyzes, and classifies the comments or a review posted online. There are instances where people are biased in their opinions and automatically that has an impact on the content they contribute to the forum as review or blog posts or tweets. As the number of such people contributing content surges it has become a huge challenge to classify and organize the real problems and prospects of the product which makes the user to doubt the reliability of the content.

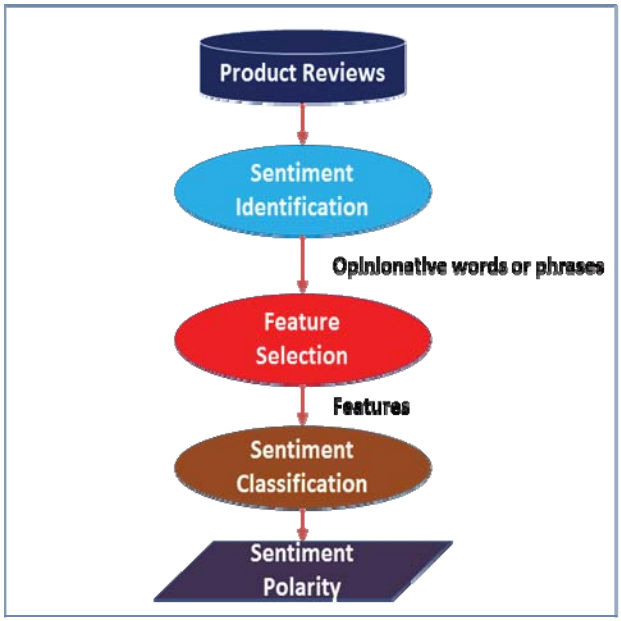
Micro blogging is a web service that allows the subscriber to broadcast short messages to other subscribers of the service. The strict 140 character limit, imposed by Twitter, is a curse. We are forced to abbreviate words, concepts and sentences. Prolific use is made of acronyms, creative contractions and the stylistic use of punctuation is common. Prominent in Tweets is the copious use of emoticons.

Emoticons are ASCII art. They are sometimes called "Smileys". They are formed through the creative use of letters, numbers and punctuation symbols. Most often (though not always) they attempt to represent facial features. As their name implies, emoticons are glyphs designed to add emotional flavor to plain text messages. Just as simple punctuation can convey surprise ! or pose a question ?, emoticons can convey happiness and joy :-), sadness :-(, laughter :-D , or cheekiness ;-)

There are many emoticons present in tweets. These emoticons can be analyzed to form the sentiment of a given tweet. A given emoticons can be classified in a three way as positive negative and neutral. This along with the words will determine the overall sentiments of a sentence/tweet.

Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. The target of SA is to find opinions, identify the sentiments they express, and then classify their polarity as shown in Figure 1. [2]

Figure 1: Sentimental Analysis Process



II. DATA DESCRIPTION

Since the opinions contributed by people and companies are to be evaluated, we consider the data from blogs, review sites, web discourse and news articles.

Review: There are many user generated reviews available on the internet that aids a customer in buying a product. E-commerce sites such as www.amazon.in, www.flipkart.com and www.reviewcentre has millions of customer reviews for products, whereas www.rediff.com/movies/reviews, www.indiaglitz. com and www.rottentomatoes.com has reviews for movies and www.yelp.com,www.burrp.com has restaurant reviews [13].

Web Discourse: A blog is a personal website or web page on which an individual records opinions, links to other sites, etc. on a regular basis. They are the fastest growing sections of the emerging communication systems. The simple and no-nonsense style of writing a post and uploading it on the web has made the blogging world an indispensable source of data in the case of sentiment mining [10].The micro blogging site Twitter is also flooded with opinions that are decisive in determining the election results even [9].These opinions can also be used for classifying sentiments [11]. People record the daily events in their lives and express their opinions, feelings, and emotions in an on-line journal, or blog or on Twitter [12].

News Articles: The websites like www.thesun.co.uk, www.cnn.com and www.thehindu.com has news articles that allow users or readers to comment. This helps in recording the opinions of the people in issues that are of current relevance and importance.

Out of the **96,269,892** Tweets that contained emoticons, the top 20 smileys accounted for 90% of all occurrences. They are listed below:

Table 1: The top emoticons in twitter [3]

S.No	Emoticon	Usage	Percent (%)	Meaning
1	:)	32,115,789	33.36	*Happy face*
2	:D	10,595,385	11.00	*Laugh*
3	:(	7,613,014	7.90	*Sad face*
4	;)	7,238,295	7.51	*Wink*
5	:-)	4,254,708	4.42	*Happy face (with nose)*
6	:P	3,588,863	3.72	*Tongue out*
7	=)	3,564,080	3.70	*Happy face*
8	(:	2,720,383	2.82	*Happy face (mirror)*
9	;-)	2,085,015	2.16	*Wink (with nose)*
10	:/	1,840,827	1.91	*Uneasy, undecided, skeptical, annoyed?*
11	XD	1,795,792	1.86	*Big grin*
12	=D	1,434,004	1.49	*Laugh*
13	:o	1,077,124	1.11	*Shock, Yawn*
14	=]	1,055,517	1.09	*Happy face*
15	D:	1,048,320	1.08	*Grin (mirror)*
16	;D	1,004,509	1.04	*Wink and grin*
17	:]	954,740	0.99	*Happy face*
18	:-(	816,170	0.84	*Unhappy*
19	=/	809,760	0.84	*Uneasy, undecided, skeptical, annoyed?*
20	=(	760,600	0.79	*Unhappy*

III. CLASSIFICATION METHODS

In this section we review fundamental aspects of three classification methods. They are lexicon based method which comes under unsupervised learning, Naïve Bayes method and Support Vector Machines which come under supervised learning methods.

A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. It is domain specific whereas unsupervised learning is the machine learning task of inferring a function to describe hidden structure from unlabeled data.

A. LEXICON-BASED CLASSIFICATION

The lexicon based approach is based on the assumption that the contextual sentiment orientation is the sum of the sentiment orientation of each word or phrase.

The classifier is based on estimating the intensity of negative and positive emotion in text, that is, the output of the classifier is one of $\{0, +1, -1\}$.

Sentiment Analysis on Twitter data is not confined to raw text. Analyzing Emoticons have been an interesting study. Go et al. (2009) used emoticons to classify the tweets as positive or negative and train standard classifiers such as Naive Bayes and Support Vector Machines. Hashtag may have some sentiment in it. Davidov et al. [21] used 50 hashtags and 15 emoticons as sentiment labels for classification to allow diverse sentiment types for the tweet. Negation and intensifier play an important role in Sentiment Analysis. Negation word can reverse the polarity, where as intensifier increases sentiment strength. Taboada et al. [20] studied role of the intensifier and negation in the lexicon based Sentiment Analysis. Wiegand et al. [19] survey the role of negation in Sentiment Analysis.

Preprocessing

The objective of the preprocessing step is to normalize the text into an appropriate form to extract the sentiments. The preprocessing steps used are:

POS Tagging: POS Tagger gives part of speech tag associated with words. POS tagging is done using Natural Language Tool Kit (NLTK).

Stemming: Stemmer gives the stem word. Non- stem words are stemmed and replaced with stem words. For example, words like 'loved', 'loves', 'loving', 'love' are replaced with 'lov'. This would aid the engine to do the word match from the text to the lexicon. Stemming is done using Natural Language Tool Kit (NLTK).

Exaggerated word shortening: Words which have same letter more than two times and not present in the lexicon are reduced to the word with the repeating letter occurring just once. For example, the exaggerated word "NOOOOOO" is reduced to "NO".

Emoticon detection: Emoticon has some sentiment associated with it. Twitter NLP is used to extract emoticons along with the sentiments in the Twitter data.[14]

Hashtag detection: The hashtag is a topic or a keyword that is marked with a tweet. Hashtag is a phrase starting with # with no space between them. Hashtags are identified and sentiments are extracted from them.

Stop Words: All the stop words (like a, an, the, is etc.)and discourse connectives are discarded.

The work flow model is shown in Figure 2.

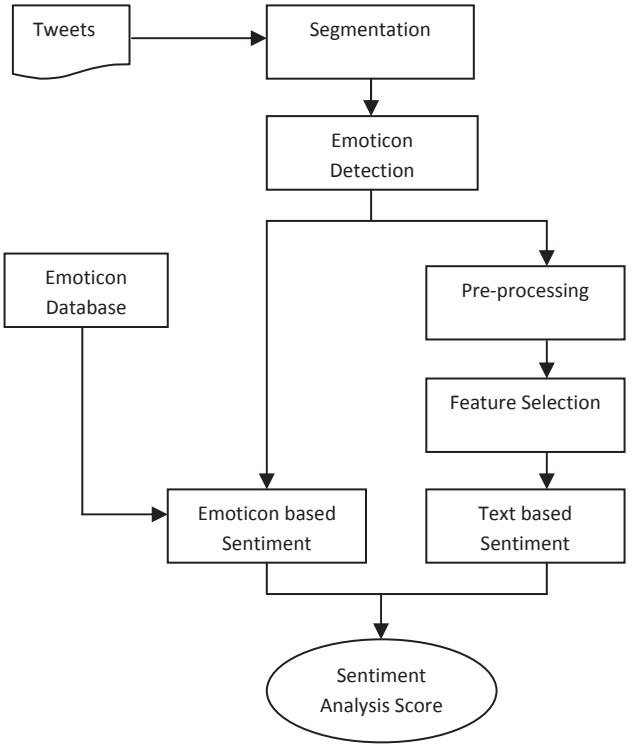


Figure 2: Workflow of model

EMOTICONS AND SENTIMENT [4]

In order to exploit emoticons in automated sentiment analysis, we first need to analyze how emoticons are typically related to the sentiment of the text they occur in. Insights into what parts of a text are affected by emoticons in which way are crucial for advancing the state-of-the-art of sentiment analysis by harvesting information from emoticons.

Emoticons as Cues for Sentiment

Emoticons can generally be used in three ways. First, emoticons can be used to express sentiment when sentiment is not conveyed by any clear positive or negative words in a text segment, thus rendering the emoticons to be carrying the only sentiment in the sentence in such cases. Second, emoticons can stress sentiment by intensifying the sentiment already conveyed by sentiment-carrying words. Third, emoticons can be used to disambiguate sentiment, for instance in cases where the sentiment associated with sentiment-carrying words needs to be negated. Some examples can be found in Table 2. Table 2 clearly shows that the sentiment associated with a sentence can differ when using different emoticons, i.e., the happy emoticon ":-D" and the "-_-"emoticon indicating extreme boredom or disagreement, irrespective of the position of the emoticons. The sentiment carried by an emoticon is independent from its embedding text, rendering word sense disambiguation techniques [11] not useful for emoticons. As such, the sentiment of emoticons appears to be dominating the sentiment carried by verbal cues in sentences, if any.

Table 2: Typical examples of how emoticons can be used to convey sentiments

Sentence	**How**	**Sentiment**
I love my job :-D	Intensification	Positive
The song was bad :-D	Negation	Positive
:-D I got a promotion	Only sentiment	Positive
-_ - I love my job	Negation	Negative
The song was bad -_ -	Intensification	Negative
I got a promotion -_ -	Only sentiment	Negative

Prabu Palanisamy et al [16] presented a lexicon based method for Sentiment Analysis with real time Twitter data. They have provided practical approaches to identifying and extracting sentiments from emoticons and hashtags. In their method, the non-grammatical words converted to grammatical words and normalize non-root to root words to extract sentiments. They have conducted experiments using 9451 subjective expressions are marked from the tweets. They got an F-score of 0.8004 on the test data set.

B. NAIVE BAYES CLASSIFICATION

Naive Bayes is a probabilistic learning method that assumes terms occur independently. In order to incorporate unlabeled data, the foundation Naïve Bayes was build. The task of learning of a generative model is to estimate the parameters using labeled training data only. The estimated parameters are used by the algorithm to classify new documents by calculating which class the generated the given document belongs to [4]. The naive Bayesian classifier works as follows: 1. Consider a training set of samples, each with the class labels T. There are k classes, C1, C2, . . . ,Ck. Every sample consists of an n-dimensional vector, X = { x1, x2, . . . ,xn}, representing n measured values of the n attributes, A1,A2, . ,An, respectively. 2. The classifier will classify the given sample X such that it belongs to the class having the highest posterior probability. That is X is predicted to belong to the class Ci if and only P(Ci |X) > P(Cj |X) for $1 \leq j \leq m$, $j \neq i$. Thus we find the class that maximizes P(Ci |X). The maximized value of P(Ci |X) for class Ci is called the maximum posterior hypothesis.

By Bayes Theorem P (A|B) = (1) The simplicity of the naïve bayes theorem is very useful when it comes to document classification. The main idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The simplicity of the Naïve Bayes algorithm makes this process efficient. HanhoonKhang Et.al [22] has proposed an improved version of the Naïve Bayes algorithm and a unigrams + bigrams was used as the feature, the gap between the positive accuracy and the negative accuracy was narrowed to 3.6% compared to when the original Naïve Bayes was used, and that the 28.5% gap was able to be narrowed compared to when SVM was used.

Pablo Gamallo, Marcos Garcia [17] presented Naïve Bayes strategy when the classifiers are implemented with the Binary strategy, when they use a polarity lexicon, and when multiwords are considered as features. The two systems submitted to Semeval competition were those obtained the best scores: CONSTR-BIN-LEX-MW and UNCONSTR-BIN-LEX-MW. In the Tweets2014 test corpus, the constrained system reached 0.62 F-score while the unconstrained version achieved 0.63.

C. SUPPORT VECTOR MACHINES (SVM)

Support Vector machine is Vector space based machine-learning method aiming to find a decision boundary between two classes that is maximally far from any point in the training data (possibly discounting some points as outliers or noise). Apart from performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. This discriminative classifier is considered the best text classification method [23]. M. Rushdi Saleh Et.al [24] has applied the new research area by using Support Vector Machines (SVM) for testing different domains of data sets and using several weighting schemes. They have accomplished experiments with different features on three corpora. Two of them have already been used in several works. The SINAI Corpus has been built from Amazon.com specifically in order to prove the feasibility of the SVM for different domains [14].

Prerna Chikersal [18] trained a Support Vector Machine (SVM) on 9418 tweets allowed to be used for training purposes. They got an F-score of 0.662 for all features.

IV. CONCLUSION

Data in social media particularly in micro blogging are short, real time and very informal. This poses a challenge to the sentiment analysis of these data. In this paper we tried to review the presence of emoticons in the micro blogging

site twitter and the different techniques for classifying these emoticons in order to predict the sentiment of the given tweet. Different classification techniques are analyzed in this paper viz, lexicon based, Naïve Bayes and SVM.

There are many potential future extensions of this work. It is interesting to investigate the contributions of other emotion indication information available in other social media. It is also interesting to study other methods to measure the sentiments orientation of the social media posts.

REFERENCES

[1] Apoorv Agarwal, BoyiXie Ilia, Vovsha, Owen Rambow, Rebecca Passonneau, Sentiment Analysis of Twitter Data, Department of Computer Science, Columbia University, New York, NY 10027 USA

[2] WalaaMedhat, Ahmed Hassan, Hoda Korashy, Sentiment Analysis Algorithms and Applications: A Survey, Ain Shams Engineering Journal,Volume 5, Issue 4, December 2014, Pages 1093-1113.

[3] *www.**datagenetics**.com/**blog**/october52012/index.html*

[4] Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Exploiting Emoticons in Sentiment Analysis.

[5] Waghode Poonam B, Prof. Mayura Kinikar, MIT Academy of Engineering, Twitter Sentiment Analysis with Emoticons, International Journal of Engineering and Computer Science, April 2015

[6] Royden Lewis, Steven Ware, Kayhan Moharreri, Sentiment analysis using emoticons

[7] M.Govindarajan, Romina M. , A Survey of Classification Methods and Applications for Sentiment Analysis

[8] Bo Pang and Lillian Lee, Opinion Mining and Sentiment Analysis .

[9] Ahmed Abbasi, Stephen France, Zhu Zhang and Hsinchun Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 3, pp. 447-462, 2011.

[10] Robert P. Schumaker ,Yulei Zhang , Chun-Neng Huang , Hsinchun Chen," Evaluating sentiment in financial news articles". Decision Support Systems 53 (2012) 458–464.

[11] M. Rushdi Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez, L.A. Ureña-López,"Experiments with SVM to classify opinions in different domains" Expert Systems with Applications 38 (2011) 14799–14804.

[12] Introduction to Information Retrieval by HinrichSchutze (course 2010) and chapter 15 of the book [MRS08], all available at www.informationretrieval.org

[13] Singh and Vivek Kumar, "A clustering and opinion mining approach to socio-political analysis of the blogosphere". Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference.

[14] Melville, WojciechGryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", KDD˝09, June 28–July 1, 2009, Paris, France.Copyright 2009 ACM 978-1-60558-495-9/09/06.

[15] Xia Hu, Jiliang Tang, Huji Gao and Huan Liu, Unsupervised sentiment analysis with emotional signals

[16] Prabu Palanisamy, Vineet Yadav and Harsha Elchuri, Serendio: Simple and Practical lexicon based approach to Sentiment Analysis

[17] Pablo Gamallo, Marcos Garcia, Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets

[18] Prerna Chikersal, Soujanya Poria, and Erik Cambria, SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning

[19] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, Andr´ es Montoyo. 2010. A survey on the role of negation in sentiment analysis. Proceedings of the workshop on negation and speculation in natural language processing 60–68, Association for Computational Linguistics.

[20] Maite Taboada, Julian Brooke,Milan Tofiloski, Kimberly Voll and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. Computational linguistics, volume 37, number2, 267–307, MIT Press.

[21] Dmitry Davidov, Oren Tsur and Ari Rappoport. 2010, Enhanced sentiment learning using twitter hashtags and smileys. Proceedings of the 23rd International Conference on Computational Linguistics 241–249, Association for Computational Linguistics.

[22] Hanhoon Kang, Seong Joon Yoo , Dongil Han," Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews". Expert Systems with Applications 39 (2012) 6000–6010.

[23] Rui Xia , Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1152.

[24] M. Rushdi Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez,L.A., Ureña-López", Experiments with SVM to classify opinions in different domains" Expert Systems with Applications 38 (2011) 14799–14804.