SentiReview: Sentiment Analysis based on Text and Emoticons

Ms. Payal Yadav
Computer Engineering Department
Sarvajanik College of Engineering and Technology
Surat, India
payalyadav501@gmail.com

Prof. Dhatri Pandya
Computer Engineering Department
Sarvajanik College of Engineering and Technology
Surat, India
dhatri.pandya@scet.ac.in

Abstract— **Social networking has gradually become a routine for people to post their opinions, views and comments on any product or person. People share their feelings online in a very informal language. Thus, it is very difficult task to analyze exact sentiments attached with that natural language. Sentiment Analysis is a study of people's attitude, opinions, and emotions to classify whether it is positive, negative or neutral. Use of emoticons on social media has increased rapidly in recent years. Hence, we have focused more on how emoticons play an important role in sentiment analysis. Various factors that affect sentiment analysis are discussed briefly in this paper. Also various issues like sarcasm detection, multilingualism, handling acronyms and slang language, lexical variation and dynamic dictionary handling are discussed.**

Keywords— Social Networks; Natural Language Processing; Sentiment Analysis; Emoticons

I. INTRODUCTION

Natural Language Processing (NLP) has vast contribution in the area of data mining [10]. It is study of automatic analysis and representation of human language [26]. NLP consists of various tasks such as Word Sense Disambiguation, Coreference resolution, Part-of-Speech (POS) Tagging, Information extraction and Sentiment Analysis [11]. Sentiment Analysis is a study of people's attitude, opinions, and emotions to classify whether it is positive, negative or neural [12] [13]. It can be applied on various forms of data such as text, emoticons, images, audio, and video. People post their views, opinions and emotions on social networking sites such as Facebook [22] [23] and Twitter [24] with wide use of emoticons [3]. Twitter users widely use hashtags and smileys for emphasizing their views behind any idea or a person [21]. This paper mainly focuses on Sentiment Analysis based on text and emoticons and how emoticons plays crucial role in analyzing sentiments. Sentiment Analysis can be applied on three levels: Document Level, Sentence Level and Entity Level [14] [25]. In this paper, we discuss about importance of emoticons in sentiment analysis and how their presence affects the sentiment analysis results. Sentiment analysis approaches such as Machine learning and Lexicon based approach are discussed in brief. Machine learning technique consists of Support Vector Machine (SVM), Maximum Entropy and Naïve Bayes Algorithms [16]. Lexicon based approaches are further classified as corpus based and dictionary based lexicon approaches [7]. After literature review of existing work done in this area, we came across many factors that affect sentiment analysis. Some issues that were identified are discussed in brief and some issues which are needed to be handled are also discussed. Rest of the paper is organized as follows: In section II, importance of emotions in sentiment analysis is discussed. Section III summarizes the existing work done in sentiment analysis field and gives a brief overview of them. Sentiment Analysis based on important attributes such as time, location, gender and age is discussed in section IV. In section V, text preprocessing and its importance in sentiment analysis is described. Section VI states Feature extraction and selection techniques. Later, section VII enlists various factors that affect sentiment analysis results when text as well as emoticons is considered as cues for analyzing sentiments. Section VIII briefly discusses some major issues in sentiment analysis field.

II. TEXT VS EMOTICONS

Users on social media use variety of emoticons. Some of the emoticons such as :) and :(are widely used to express emotions [7] [9]. In face-to-face communication, sentiment can often be concluded from visual cues like smiling or crying or laughing. However, in plain text communication, such visual cues are lost [3]. Thus, People use emoticons as an alternative to face-to-face visual cues. Emoticons are powerful units to change the polarity of a statement. When a smiling emoticon is added in a positive statement, the orientation of statement becomes more towards positive. E.g. Loved the food and Loved the food :) . Few sentences are ambiguous and do not carries any sentiment. In those situations, visual cues are used to estimate actual emotion behind the sentence. Such as, "I feel like a fool today :)" and "I feel like a fool today :(" differs because of emoticon attached at the end of it.

978-1-5090-5960-7/17/$31.00 ©2017 IEEE

III. RELATED WORK

There have been abundant studies done on sentiment analysis in recent years and much work has been done to focus on relationship between text and emoticons to classify the polarity of text in exact manner. Figure 1 shows general phases of sentiment analysis which includes basic steps to be followed to analyze sentiments from input dataset.

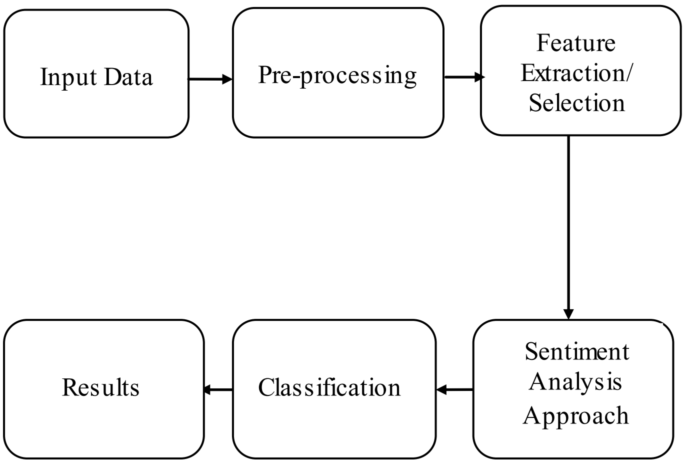

Fig.1 Phases of Sentiment Analysis

A. Lexicon Based Approach

Lexicon is the vocabulary of a person, language, or branch of knowledge [27]. Lexicon based sentiment analysis approaches use lexicons for calculating polarity of individual words and aggregate their scores to determine overall polarity of text [7]. Lexicon based approaches are further classified into two sub approaches: (1) Corpus based approach, (2) Dictionary based approach. Corpus based approach [2] is not word oriented, but context oriented. Whereas dictionary based approach is used by combining existing seeds present in the dictionary to build more dictionaries [1]. In paper [1], M. Datar and P. Kosamkar proposed a novel approach for sentiment analysis using dictionary based lexicon approach and represented it on a unique figure called "emoticon-graph". They also considered punctuation marks such as ! and ? for extracting exact emphasis of them on sentences. Hogenboom et al. proposed an approach [3] for sentiment analysis by maintaining separate dictionaries for text and emoticons and assign score to them. After assigning score to both text and emoticons separately, results were aggregated and final score of input statement was calculated with weighted average technique [3].

B. Machine Learning

Machine learning is the subfield of computer science that gives computers the propensity to learn without programming them explicitly [28]. Machine learning algorithms like Support Vector Machine (SVM), Naïve Bayes, Maximum Entropy, and Neural Networks have been used extensively for sentiment analysis [9] [15] [16]. Although Machine learning algorithms can outperform simple lexicon based methods, they require large training databases to be effective and efficient. Another challenge in sentiment analysis through machine learning is its domain specific adaptation [7]. It works accurately only for the domain in which it has been trained. In paper [7], H. Wang and A. Jorge proposed a system using deep learning algorithm Word2Vec to represent words, including emoticons in the data set, and generate feature matrix [7]. In referred work [10], authors used naive bayes classifier using incremental learning technique to perform analysis in hourly, weekly and monthly pattern. Summary of existing techniques in sentiment analysis is shown in Table 1.

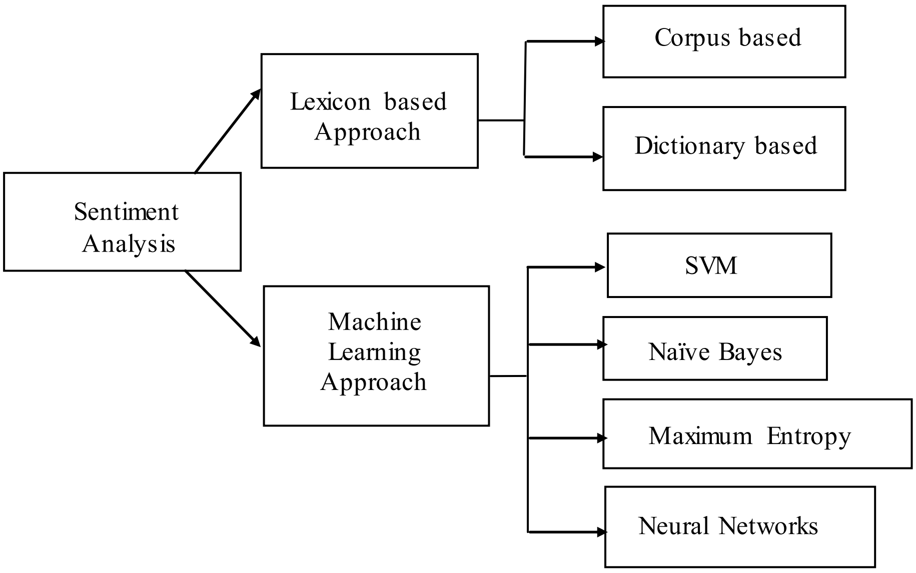

Fig.2 Sentiment Analysis Approaches

Table 1. Summary of existing Techniques

Referred work	Data Sources	Pre-Processing	Technique used
A novel Approach for Polarity Determination Using Emoticons: Emoticon-Graph [1]	Online Survey on application reviewing	Blank space and punctuation removal	Dictionary based Lexicon Approach
Localized Twitter Opinion Mining using Sentiment Analysis [2]	Twitter Oauth	POS Tagging	Corpus based Lexicon Approach
Exploiting Emoticons in Sentiment Analysis [3]	Dutch Tweets	Segmentation	Dictionary based Lexicon Approach
MoodLens: An Emoticon-Based Sentiment Analysis System for Chinese Tweets [4]	Weibo Tweets	Bag of Word	Incremental Naïve Bayes Classifier
Monitoring System for Potential Users with Depression Using Sentiment Analysis [5]	Social Network Posts	Filtering using Machine Learning (SVM)	Corpus based Lexicon Approach
A :) is worth a Thousand Words: How people attach sentiment to emoticons and words in tweets [6]	Twitter API	Case Conversion, Remove Non alphabetic characters, Stemming	Pointwise Mutual Information
Issues of Social Data analytics with a new method for sentiment analysis of social media [7]	Tweets related to public transportation	Segmentation	Adaptive Fuzzy Inference method with linguistics processors
Multilingual Sentiment Analysis using Emoticons and Keywords [8]	Geek forums	Stemming, Stop word Removal	K-Nearest Neighbor, SVM, Logistic Regression, Multinomial naïve bayes classifier
Sentiment Expression via Emoticons on Social Media [9]	Twitter Decahose API	Filtering based on frequency	Word2Vec (Deep Learning) and K-Means clustering

IV. CONTEXT BASED SENTIMENT ANALYSIS

In recent years, much research has been done on analyzing sentiments based on attributes such as gender, age, location, time, etc. These attributes helps to figure out correct sentiment based on context. Sentiment Analysis on product review data highly depends on gender, age and location of person. People posts their views for a product based on their personal experiences which vary from location to location [2]. Analyzing mood of an individual from his/her social network profile is also a context based sentiment analysis where age and gender of person acts as important parameters to extract their sentiments [4]. Mood of an individual may change in hour, day and weekly basis [5]. Thus analysis in terms of time is an important aspect for context based sentiment analysis. In such cases, emoticons used by people of different age, gender at different location and on different time varies.

V. SIGNIFICANCE OF PREPROCESSING

Most of the existing system removes special characters and numbers in text pre-processing step. Due to this, characters such as :,), (, - and ! which makes up emoticons are removed and system classify sentiments only on basis of textual cues given in data. Preprocessing is an important step in sentiment analysis which helps to improve efficiency as well as effectiveness [15]. Preprocessing tasks includes Stop word removal, Stemming, Lemmatization, Part-of-Speech (POS) tagging [2] , Expand Abbreviations and Chunking [11]. Proper preprocessing leads to better results in classifying data into positive, negative and neutral classes. It also makes the task lighter and efficient to process. Data that do not consist of any sentiments are removed during pre-processing stage itself so that the task does not get bulky in later stages of sentiment analysis.

VI. Feature Extraction and Selection

Feature extraction techniques [15] reduce the feature vector length by transforming all the features in lower-dimensional feature vector. It maps the high-dimensional data on lower-dimensional space.

A. Unigrams

Unigram features are bag-of-words (BoW) [4] features extracted by removing extra spaces and noisy characters between two words. E.g. "The movie was awesome." Here, words "The", "movie", "was", and "awesome" are all unigram features.

B. Bigrams

Bigrams are the features composed of every two consecutive words in the text. For the above example, "The movie", "movie was" and "was awesome" are the bigram features. These features have potential of including some contextual information.

C. Bi-tagged

Bi-tagged features are selectively extracted using part-of-speech (POS) based fixed patterns [16]. Bigrams containing mostly adverbs and adjectives are considered more sentiment bearing.

Feature selection techniques [16] select the minimum notable features and represent the class attribute in the reduced feature space. Feature selection techniques can significantly improve the classification accuracy and also provide better insight into important class features, resulting in a better understanding of sentiments.

A. Term Frequency (TF): Specifies number of times a feature appears in document

B. Feature Presence (FP): Determines whether a feature appears in a document or not. It gives value in 0 or 1. Where, 1 indicates presence and 0 indicates absence of feature.

C. Term Frequency- Inverse document Frequency (TF-IDF): Signifies importance of a particular word in a document.

VII. Factors Affecting Sentiment Anlaysis

Listed below are factors that affect sentiment analysis when both text and emoticons are taken into consideration:

A. Level of Analysis

When sentiment analysis is applied on sentence level, words and emoticons which are not a part of that sentence do not affect the overall result [14]. In cases where document level analysis is done, all words and emoticons are taken into consideration to analyze overall sentiment.

B. Domain Dependency

Few words and emoticons are ambiguous in nature. They mean something in a domain and something other in different domain. For example "Cheap" is considered as positive word for product reviews but it is considered as negative for movie reviews [7].

C. Intensifiers

The use of exclamation marks and dots are very much in trend these days. They are called "Amplifiers" and add more sentiments to words associated with them [1]. For example "It was nice meeting you" and "It was nice meeting you…!!!" differs due to presence of intensifiers.

D. Count of Emoticons

People use emoticons in wide way to express their feelings. Many have habit of repeating same emoticon [6] more than once to emphasis more on their feelings. For example "What a pleasant surprise :) :) :) " carries more happiness than "What a pleasant surprise :) " .

E. Emoticon Position

When we deal with paragraphs, we find different emoticons attached with different sentences. For Example "The camera quality of phone is not that good, but battery life has forced me to buy it :) ". Here the customer is showing a positive opinion towards the battery life of phone and not to the camera quality.

F. Classification

Classification of data in mostly done into three common classes: Positive, negative and Neutral [18]. Few researchers have also classified their work into sub classes such as Angry, Sad, Happy, Joyful and much more [2] [4] [6]. As more is the level of classification, better results are generated.

VIII. Major Issues

The field of sentiment analysis is based on very informal language and consists of many flaws and ambiguities which we as humans can easily percept but it is difficult for a machine or a system to recognize.

A. Sarcasm

Sarcasm [20] is taunting or an ironical expression. As humans, we can easily recognize sarcasm in verbal communications as well as textual communications, but it is difficult for a machine to identify whether it is a sarcastic statement or a negative statement [17]. E.g. I wish I could see you with a personality :). The statement consists of no negative words and no positive words as well. But the emoticon attached at the end cues that it is an ironic statement.

B. Multilingualism

Most of researchers work on English language for sentiment analysis [8]. As social networks and search engines now support multiple languages, there is need for sentiment analysis on other languages like Greek, Chinese, Hindi and French.

C. Dynamic Handling of words and emoticons

Majority of work done in lexicon based sentiment analysis uses static dictionary for maintaining scores of text and emoticons [19]. In such cases, system fails to classify the terms which are not present in dictionary or not trained by a machine learning algorithm.

D. Misspells

Humans type misspellings most of the times while they post their views or comments on social media. A misspell is easily understood by a human reader but a machine would fail to recognize that word even if it is present in dictionary or been trained by learning algorithm.

E. Negation Handling

Negation words are those which change the polarity of a sentence when appears. E.g. "The movie was good" and "The movie was not good" are two oppositely oriented statements. The word "not" switched the positive statement into negative one.

F. Acronyms and Langugage Slangs

Informal Language consists of many language slangs and acronyms such as gr8, ni8, HAHAHA, LOL, f9, TTYL and many more. In most of the existing work, numbers are removed during preprocessing step and system do not classify words like "gr8" that holds sentiment. Expanding abbreviation is also not done by most of existing methods which tends to failure in recognizing sentiments behind words like "LOL".

G. Lexical Variation

Some words are written in more than one style E.g. "awesome" is "awsum", "awesum", "awsom" and many more. Dealing with these variation of lexicons is a challenging task in sentiment analysis.

IX. CONCLUSION

In this paper, importance of emoticons in sentiment analysis has been shown by different examples. Factors that affect sentiment analysis are discussed in brief. The paper also summarizes existing approaches for sentiment analysis. Text pre-processing, feature extraction and feature selection plays an important role for analyzing sentiments efficiently. Various issues such as sarcasm, dynamic dictionary handling, acronyms and language slangs, lexical variations have been discussed in brief. Among existing techniques in sentiment analysis, Machine learning techniques are domain specific and work well for a specific domain (Movie or product reviews) but not in general applications such as sentiment analysis on social networking data or twitter dataset. Lexicon based approach are convenient for all domains as it emphasis on part of text present in the lexicon. Various machine learning techniques and lexicon based techniques can be combined to form a hybrid approach which may result into more accurate sentiment analysis.

REFERENCES

[1] M. Datar and P. Kosamkar, "A Novel Approach for Polarity Determination Using Emoticons: Emoticon-Graph," In Advances in Intelligent Systems and Computing, 2016, pp. 481-489.

[2] S. Hridoy, M. Ekram, M. Islam, F. Ahmed and R. Rahman, "Localized twitter opinion mining using sentiment analysis," Decision Analytics, 2015, p.1.

[3] A. Hogenboom, D. Bal and F. Frasincar, "Exploiting Emoticons in sentiment analysis," In Proceedings of the 28th Annual ACM symposium on Applied Computing, 2013, pp. 703-710.

[4] J. Zhao, L. Dong, J. Wu and K. Xu, "MoodLens: an emoticon-based sentiment analysis system for chinese tweets," In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12, 2012, pp. 1528-1531.

[5] R. Rosa, G. Schwartz and I. de Campos Ribeiro, "Monitoring system for potential users with depression using sentiment analysis," In International Conference on Consumer Electronics (ICCE) , 2016, pp. 381-382.

[6] M. Boia, B. Faltings, C. Musat and P. Pu, "A :) Is Worth a Thousand Words: How People Attach Sentiment to Emoticons and Words in Tweets," In International Conference on Social Computing (SocialCom), Alexandria, VA, 2013 pp. 345-350.

[7] Z. Wang, V. Joo, C. Tong and D. Chan, "Issues of social data analytics with a new method for sentiment analysis of social media data," In 6th International Conference on cloud computing technology and science, 2014, pp. 899-904.

[8] G. Solakidis, K. Vavliakis and P. Mitkas, "Multilingual Sentiment Analysis Using Emoticons and Keywords,"In IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014.

[9] H. Wang and J. Castanon, "Sentiment expression via emoticons on social media," In IEEE International Conference on Big Data, Santa Clara, CA, 2015, pp. 2404-2408.

[10] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," In Computational Intelligence Magazine, 2014, vol. 9, no. 2, pp. 48-57.

[11] R. Dudhabaware and M. Madankar, "Review on natural language processing tasks for text documents," In Computational Intelligence and Computing Research (ICCIC),Coimbarote, 2014, pp. 1-5.

[12] V. Singh and S. Dubey, "Opinion mining and analysis: A literature review", In 5th International Conference The Next Generation Information Technology Summit (Confluence) , Noida, 2014, pp. 232-239.

[13] K. Ahmed, N. Tazi and A. Hossny, "Sentiment Analysis over Social Networks: An Overview,"In International Conference on Systems, Man and Cybernetics (SMC), Kowloon, 2015, pp. 2174-2179.

[14] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, 2014, vol. 5, no. 4, pp. 1093-1113.

[15] E. Haddi, X. Liu and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," Procedia Computer Science, 2013, vol. 17, pp. 26-32.

[16] B. Agarwal and N. Mittal, "Machine Learning Approaches for Sentiment Analysis," In Springer International Publishing Switzerland , 2016, pp. 193-208.

[17] P. Carvalho, L. Sarmento, M. Silva and E. Oliveira, "Clues for detecting irony in user-generated contents: oh...!! it's so easy;-)," In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion , 2009, pp. 53-56.

[18] T. Yokoi, M. Kobayashi, and R. Ibrahim, "Emoticon Extraction Method Based on Eye Characters and Symmetric String," In International Conference on Systems, Man, and Cybernetics (SMC), 2015, pp. 2979-2984.

[19] H. Ameur, and S. Jamoussi, "Dynamic construction of dictionaries for sentiment classification," In 13th International Conference on Data Mining Workshops, 2013, pp. 896-903.

[20] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in twitter," Language resources and evaluation 47, 2013, pp. 239-268.

[21] D. Davidov, O. Tsur and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," In Proceedings of the 23rd international conference on computational linguistics, 2010, pp. 241-249.

[22] J. Akaichi, "Social networks' Facebook'statutes updates mining for sentiment classification," In International Conference on Social Computing (SocialCom), 2013, pp. 886-891.

[23] J. Akaichi, Z. Dhouioui and M. Pérez, "Text mining facebook status updates for sentiment classification," In 17th International Conference System Theory ,Control and Computing (ICSTCC), 2013, pp. 640-645.

[24] E. Kouloumpis, T. Wilson and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011, pp. 538-541.

[25] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou and P. Li, "User-level sentiment analysis incorporating social networks," In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 1397-1405.

[26] Wikipedia. (2016). Natural Langugage Processing. Available: https://en.wikipedia.org/wiki/Natural_language_processing

[27] Wikipedia. (2016). Lexicon. Available: https://en.wikipedia.org/wiki/Lexicon

[28] Wikipedia. (2015). Machine Learning. Available: https://en.wikipedia.org/wiki/Machine_learning