

Sentiment analysis on online social network

Vijaya Abhinandan

vijaya.abhinanda@gmail.com

Department of Computer Science and Engineering
Nalla Malla Reddy Engineering College

ABSTRACT

A large amount of data is maintained in every Social networking sites. The total data constantly gathered on these sites make it difficult for methods like use of field agents, clipping services and ad-hoc research to maintain social media data. This paper discusses the previous research on sentiment analysis.

1. Introduction and State of Art

A large amount of data is maintained in every Social networking sites. The total data constantly gathered on these sites make it difficult for methods like use of field agents, clipping services and ad-hoc research to maintain social media data [2]. It is essential to employ tools able of analyzing social media mainly the characteristics of social media.

The mining review which uses machine learning and semantic orientation has investigated [3]. The approach which is used to classify the movie review uses the supervised classification.

A corpus is formed to represent the data in the documents and all the classifiers are trained using this corpus. Thus the proposed technique is more efficient. The machine learning approach uses supervised learning, the proposed semantic orientation approach uses “un-supervised learning” prior data is not required to mine the data.

To investigate the effectiveness of classification of the documents machine learning techniques are used [4]. By experimenting machine learning techniques are much better than human produced baseline for sentiment analysis on a review data. This classification uses features which are based on unigrams and bigrams.

Zhu[5] proposed aspect based opinion polling free from textual customer reviews. Opinion polling uses aspect based segmentation model which segment the multi-aspect sentence to single aspect units.

A sentiment analyzer to extract opinions has proposed [6] related to a subject from online data documents. Sentiment analyzer uses natural language processing techniques. The sentiment analyzer finds out all the references on the subject and sentiment polarity of each reference is determined. The sentiment analysis conducted by the researches utilized the sentiment lexicon and sentiment pattern database for extraction and association purposes.

Alekh Agarwal[7] proposed a machine learning method incorporating linguistic knowledge gathered through synonymy graphs, for effective opinion classification. This approach shows the degree of influence among relationships of documents have on their sentiment analysis. This is brought about by the use of graph-cut technique and opinion words got through synonym graphs of word net. The proposed approach also improves the accuracy of predictions in classification task. Experiments using the system have given results with the accuracy over 90 percent with an advantage of decreased in processing time, with the difference of minimum in the final accuracies.

The proposed methodology from the authors resulted in the following conclusions:

1. Automated mining of linguistic information is possible, so demonstrated with the structure of links in word net.
2. Generic method of using graph-cut technique for efficient opinion classification.

Ahmed Abbasi[8] proposed sentiment analysis methods which are used to classify the forum opinions in web which are in multiple languages. To enhance the performance of the classifier the entropy weighted genetic algorithm is incorporated. Using movie review data set the experiments stated that the used techniques are efficient. Machine learning has many application for security and text application for many area [9-13]

Anidya et al., [9] ranked the product reviews based on customer-oriented and manufacturer ranking mechanism. The expected helpfulness of the review is used for the ranking and also

ranking is based on the expected effect on sale. The proposed methods identify the reviews which have the most impact.

To all the customer reviews of a product Miniqing Hu[10] performed mining and summarization process. The process is carried out in three steps:

1. The product characteristics commented by the customer in the review are mined. Natural language processing and Data mining techniques are used for mining.
2. The opinions in the review are identified and the opinions are divided as positive or negative. Set of adjectives words called opinion words are identified. To identify the semantic orientation and opinion orientation of each sentence is decided WordNet is used.
3. Summarize the results.

The main objective is to perform summary of a customer reviews of a product which are sold online.

Qui[11] analyzed the problems which are related to opinion mining such as the expansion of lexicon and target of the opinion. Good, Bad, Excellent, Poor are the list of opinion words which are used to indicate positive and negative sentiments. Based on boot strapping the link between opinion word and targets relations are identified. Bootstrapping process is started using the initial opinion lexicon. Semi-supervised methods are used in this process.

Lei Zhang[12] analyzed opinion words which are domain dependent. Sentiment context of each noun feature which is positive or negative. For identifying the noun product feature two steps are used which are positive or negative.

Xiaowen Ding [13] proposed a holistic lexicon-based approach which uses External indications. This approach has the advantage that opinion words which are context dependent are easily handed. Linguistic Patterns are used in this algorithm to deal with special words, phrases.

Sentiment classification:

Sentiment classification is the classification task of naming an document as positive or negative opinions. Machine learning algorithms are used for sentiment classification.

Machine Learning Algorithms:

Machine learning algorithm is defined as a system which is able of getting and merging the knowledge automatically. The systems that learn from analytical observation, training, experience, and other means, results in a system that can reveal self-improvement, effectiveness and efficiency. Knowledge and a corresponding knowledge organization are usually used by a machine learning system to test the knowledge acquired, interpret and analyse.

Supervised learning generates a function which connects input to expected outputs which are also called as labels.

Semi-supervised learning generates an suitable function or classifier in which both labelled and unlabelled examples are combined[16,18].

Sentiment Analysis Tasks

Sentiment classification consists of the polarity classification of a given document text. Positive, negative, neutral express the sentence level opinions.

The sentiment analysis can be done at three levels which are: the document level, Sentence level and feature level.

Document Level sentiment classification: In document level sentiment analysis main challenge is to extract informative text for inferring sentiment of the whole document. The learning methods can be confused because of objective statements are rendered by subjective statements and complicate further for document categorization task with conflicting sentiment[17].

Sentence level sentiment classification:

The sentiment classification is a fine-grained level than document level sentiment classification in which polarity of the sentence can be given by three categories as positive, negative and neutral. The challenge faced by sentence level sentiment classification is the identification features indicating whether sentences are on-topic which is kind of co-reference problem.[17]

Feature sentiment classification: Product features as product attributes of components. Analysis of such features for identifying sentiment of the document is called as feature based sentiment analysis. In this approach positive or negative opinion is identified from the already extracted features. It is a fine grained analysis model among all other models[16].

Text Classification

The large volume of online text is available through the news feed from internet, mails, databases, websites and the digital libraries. The problem is to arrange text documents from large databases.

Naive bayes classifier: The naïve bayes classifier is well known as probabilistic classifier which describes its application to text.

The naive bayes classifier was build in order to incorporate the data which is unlabelled. The main aim of learning generative model is to estimation of the parameters which uses labelled training data. The algorithm uses estimated parameters for the classification of new documents by the calculation to which class it belongs to.

The navie Bayesian classifier works as follows:

By considering samples of a training set ,with the class labels T.

$C_1, C_2, C_3, \dots, C_k$ are the classes .Each sample contains n-dimensional vector, $X = \{x_1, x_2, x_3, \dots, x_n\}$ which represents n measured values of the n attributes named as $A_1, A_2, A_3, \dots, A_n$ respectively.

The classifier classifies the X which is the given sample which belongs to the class which has the highest probability. Thus X is belonged to the class C_1 if and only.

$$P(C_i/X) > P(C_j/X) \text{ for } 1 \leq j \leq m, j \neq i.$$

The value of $P(C_i/X)$ which is maximized is called as the maximum posterior hypothesis. By Bayes theorem,

$$P(C_i/X) = \frac{P(X/C_i) P(C_i)}{P(X)}$$

$P(X)$ is same for all classes value so only the value of $P(X|C_i)P(C_i)$ needs to be maximized.

The estimation of priori probabilities of the class $P(C_i)$ is

$$P(C_i) = \text{freq}(C_i, T)/|T|$$

$$P(X|C_i) \approx \prod_{k=1}^n P(x_k|C_i)$$

The expectation maximization algorithm oscillates between the steps of guessing the model which depends on probability of distribution over completions of missing data (E-step) and the re-estimation of the model parameters using these completions (M-step).

E-step is used only to compute the expected statistics over completions rather than the probability distribution forming over completions.

Where M-step consists of re-estimation of the model which can be maximization of the estimated log-likelihood the data.[51,52].

2.Sentiment Analysis Methods:

This section provides a brief description of the eight sentiment analysis methods investigated in this paper. These methods are the most popular in the literature (i.e., the most cited and widely used) and they cover diverse techniques such as the use of Natural Language Processing (NLP) in assigning polarity, the use of Amazon's Mechanical Turk (AMT) to create labeled datasets, the use of psychometric scales to identify mood-based sentiments, the use of supervised and unsupervised machine learning techniques, and so on. Validation of these methods also varies greatly, from using toy examples to a large collection of labeled data.

Emoticons

The simplest way to detect polarity (i.e., positive and negative affect) of a message is based on the emoticons it contains. Emoticons are popular in these days, to the extent that some (e.g. <3) are now included in English Oxford Dictionary [19]. Emoticons are primarily face-based and represent happy or sad feelings, although a wide range of non-facial variations exist: for instance, <3 represents a heart and expresses love or affection.

To extract polarity from emoticons, we utilize a set of common emoticons from [20, 21, 22] as listed in Table 1. This table also includes the popular variations that express the primary polarities of positive, negative, and neutral. Messages with more than one emoticon were associated to the polarity of the first emoticon that appeared in the text, although we encountered only a small number of such cases in the data. As one may expect, the rate of OSN messages containing at least one emoticon is very low compared to the total number of messages that could express emotion. A recent work has identified that this rate is less than 10% [23]. Therefore, emoticons have been often used in combination with other techniques for building a training dataset in supervised machine learning techniques [24].

SentiStrength

Machine-learning-based methods are suitable for applications that need content-driven or adaptive polarity identification models. Many key classifiers for identifying polarity in OSN data have been proposed [26,27,28]. The most comprehensive work [28] compared a wide range of supervised and unsupervised classification methods, including simple logistic regression, SVM, J48 classification tree, JRip rule-based classifier, SVM regression, AdaBoost, Decision Table and Naive Bayes. The core classification of this work relies on the set of words in the LIWC dictionary [29], and the authors expanded this baseline by adding new features for the OSN context. The tool named SentiStrength, which implements a combination of learning techniques that produces the best results and the best training model empirically obtained [28].

3. Conclusion

A large amount of data is maintained in every Social networking sites. The total data constantly gathered on these sites make it difficult for methods like use of field agents, clipping services and ad-hoc research to maintain social media data. This paper discusses the previous research on sentiment analysis. Future direction of analysis online social network will involve many application as mentioned in [36-41].

References:

- [1] Kim, P: "The Forrester Wave: Brand Monitoring" Forrester Wave, 2006.
- [2] Pang, B., Lee, "Proceedings of the International Conference on Computational Linguistics 2008.
- [3] Lina Zhou, Pimwadee Chaovalit, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", 2005.
- [4] P.S. Dodds and C.M. Danforth. Measuring the happiness of large-scale written expression:
- [5] Zhu, Jingbo Wang, Huizhen Zhu, Muhua Tsou, Benjamin K. Ma, Matthew, "Aspect-Based Opinion Polling from Customer Reviews". 2011.
- [6] Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack: "Sentiment Analyzer: Extracting Sentiments about Given Topic using Natural Language Processing Techniques", 2003

- [7] Alekh Agarwal & Pushpak Bhattacharyya, Sentiment analysis: “A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified”, 2005.
- [8] Ahmed Abbasi, Hsinchun Chen, And Arab Salem, “Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums”.
- [9] Yaqoob, I., et al., The rise of ransomware and emerging security challenges in the Internet of Things. *Computer Networks*, 2017.
- [10] Mujtaba, G., et al., Email Classification Research Trends: Review and Open Issues. *IEEE Access*, 2017.
- [11] Mujtaba, G., et al., Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PloS one*, 2017. **12**(2): p. e0170242.
- [12] Al-garadi, M.A., et al., Using online social networks to track a pandemic: A systematic review. *Journal of biomedical informatics*, 2016. **62**: p. 1-11.
- [13] Zerdoumi, S., et al., Image pattern recognition in big data: taxonomy and open challenges: survey. *Multimedia Tools and Applications*, 2017: p. 1-31.
- [14] Anindyam Ghose, Panagiotis G. Ipeirotis: “Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews”.
- [15] Mingqing Hu and Bing Liu, “Mining and Summarizing Customer Reviews”.
- [16] Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen. "Opinion Word Expansion and Target Extraction through Double Propagation".
- [17] Lei Zhang and Bing Liu. "Identifying Noun Product Features that Imply Opinions.", 2011
- [18] Xiaowen Ding. A Holistic Lexicon-Based Approach to Opinion Mining.
- [19] Wellman, Barry & S.D. Berkowitz, *Social Structures: “A Network Approach*. Cambridge: Cambridge University Press”, 1988.
- [20] Linton (1977). “A set of measures of centrality based on betweenness”.
- [21] Zhong wu zhai, Bing Liu, Hua Xu & Hua Xu, “Clustering product features for opinion mining”.

- [22] V.S.Jagtap & karishma pawar, "Analysis of different approaches to sentence-Level Sentiment Classification".
- [23] "Bing Liu. Sentiment Analysis and opinion Mining", 2012.
- [24] Graphic symbol for love (and that exclamation) are added as words.
- [25] List of text emoticons: "The ultimate resource. www.cool-smileys.com/text-emoticons".
- [26] Msn messenger emoticons. "<http://messenger.msn.com/Resource/Emoticons.aspx>".
- [27] Yahoo messenger emoticons.
<http://messenger.yahoo.com/features/emoticons>.
- [28] J.Park, V.Barash, C. Fink, and M. Cha: "Interpreting differences in emoticons across cultures". 2013.
- [29] J. Read. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification".
- [30] Y.R.Tausczik & J. W. Pennebaker "Liwc and computerized text analysis methods", 2010.
- [31] A.Bermingham & A.F.Smeaton: "Classifying Sentiment in Microblogs".
- [32] G.Paltoglou & M.Thelwall: "Unsupervised Sentiment Analysis in Social Media".
- [33] M.Thelwall: "Sentiment strength detection in the social web with sentiment strength".
- [34] Al-garadi, M.A., K.D. Varathan, and S.D. Ravana, Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 2016. **63**: p. 433-443.
- [35] Albrecht, C.C., C.O. Albrecht, and M. Zimbelman, *Fraud examination*. 2008: Cengage Learning.
- [36] Al-garadi, M.A., K.D. Varathan, and S.D. Ravana, Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method. *Physical A: Statistical Mechanics and its Applications*, 2017. **468**: p. 278-288.
- [37] Berthon, P.R., et al., Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy. *Business horizons*, 2012. **55**(3): p. 261-271.
- [38] Khan, M.S., et al., Virtual community detection through the association between prime nodes in online social networks and its application to ranking algorithms. *IEEE Access*, 2016. **4**: p. 9614-9624.

- [39] Al-Garadi, M.A., et al., Identifying the influential spreaders in multilayer interactions of online social networks. *Journal of Intelligent & Fuzzy Systems*, 2016. **31**(5): p. 2721-2735.
- [40] Corley, C.D., et al., Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*, 2010. **7**(2): p. 596-615.
- [41] Yang, C.C., et al. Social media mining for drug safety signal detection. in *Proceedings of the 2012 international workshop on Smart health and wellbeing*. 2012. ACM.