International Journal on Data Science and Technology
2018; 4(2): 49-53
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20180402.12
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Sentiment Analysis Using Text Mining: A Review

Swati Redhu[1], Sangeet Srivastava[1, *], Barkha Bansal[1], Gaurav Gupta[2]

[1]Department of Applied Sciences, The NorthCap University, Gurgaon, India

[2]School of Mathematical Sciences, College of Natural, Applied and Health Sciences, Wenzhou-Kean University, Wenzhou, China

Email address:
swatiredhu41@gmail.com (S. Redhu), sangeetsrivastava@ncuindia.edu (S. Srivastava), barkha.bansal20@yahoo.com (B. Bansal), ggupta@kean.edu (G. Gupta)

[*]Corresponding author

To cite this article:
Swati Redhu, Sangeet Srivastava, Barkha Bansal, Gaurav Gupta. Sentiment Analysis Using Text Mining: A Review. *International Journal on Data Science and Technology*. Vol. 4, No. 2, 2018, pp. 49-53. doi: 10.11648/j.ijdst.20180402.12

Received: April 20, 2018; **Accepted**: June 20, 2018; **Published**: June 26, 2018

Abstract: Text mining and sentiment analysis have received huge attention recently, specially because of the availability of vast data in form of text available on social media, e-commerce websites, blogs and other similar sources. This data is usually unstructured and contains noise, therefore the task of gaining information is complex and expensive. There is a growing need for developing different methodologies and models for efficiently processing the texts and extracting apt information. One way to extract information is text mining and sentiment analysis, that include: data acquisition, data pre-processing and normalization, feature extraction and representation, labelling, and finally the application of various Natural Language Processing (NLP) and machine learning algorithms. This paper provides an overview of different methods used in text mining and sentiment analysis elaborating on all subtasks.

Keywords: Sentiment Analysis, Supervised Learning, Unsupervised Learning, Text Mining, Feature Extraction, Feature Representation

1. Introduction

Sentiment analysis, also known as opinion mining, in essence, is the process of quantifying the emotional value in a series of words or text, to gain an understanding of the attitudes, opinions and emotions expressed. Sentiment analysis can be applied to various sectors such as e-commerce, banking, mining social media websites like Face book, Twitter and so on. Using sentiment analysis and text mining, organizations can gain consumer insight from the response about their products and services. This can be further used to study customers' satisfaction with the services and in case of complaints and issues, finding the possible reasons for that. One of the applications of sentiment analysis is recommendation systems, for instance YouTube recommends on the basis of consumers likes, dislikes and comments provided by the user. In this paper, we extensively study various text mining and sentiment analysis techniques applied to different areas in multi lingual format and from different resources. A sentiment analysis and text mining framework typically includes following subtasks: acquiring text data, data cleaning and pre processing, data normalization, conversion of text to machine readable vectors, features selection, and finally applying NLP and machine learning algorithms. In this paper, we present a literature review on recent trends in text mining and sentiment analysis. For instance, consumer review mining and application to tourism industry are the current successful applications. Topic modelling is successfully combined with sentiment priors to generate topics and sentiment classes simultaneously. Emoji and emoticon sentiments are included in many of the studies to improve accuracy of results and so on.

2. Literature Review

In [1] Duwairi et. al mentioned that sentiment analysis determines the polarity of given text either using machine learning approach or using lexicon based approach. The classifiers applied on the datasets were Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN (k=10)) where SVM gave highest precision and KNN gave

the highest recall. Also to test the data sets 10-fold cross validation was used. They demonstrated that the precision got by SVM i.e. 75.25 was the best precision and the recall got by KNN i.e. 69.04 was the best recall. Therefore, to get better classification results, bigger data sets were required and to label them crowd sourcing was considered followed by semi supervised learning.

In [2] Kouloumpis et. al demonstrated the usefulness of linguistic features and existing lexical resources used in micro-blogging to detect the sentiments of twitter messages. From this paper the researchers concluded that micro-blogging features were more useful as compared to POS (Part-of-Speech) features and features from existing sentiment lexicon. They also concluded that if they include micro-blogging features then the training data will be of less benefit. [3] consists of a new method formed by combination of rule based classification, supervised learning and machine learning which showed the improvement in micro and macro averaged F1. To get better effect, Prabowo et. al considered semi-automatic approach. From this paper they concluded that hybrid classification was better than the classification by any individual classifier. They also concluded that reduction of rules will produce less effect on F1.

From [4], Mudinas et. al concluded that concept level sentiment analysis system (psenti) was better as compared to pure lexicon based system and pure learning based system due to more precision in polarity classification and well structured, readable results. On experimenting, they confirmed that hybrid approach was better than sentistrength. From their paper, they concluded that psenti system obtained high precision than pure lexicon based system but near to pure learning based system. It also gave well structured, readable results and more resistance to writing style of text. They also concluded that psenti system works better than sentistrength. In short, the proposed hybrid approach was capable in combining a carefully designed lexicon and a powerful supervised learning algorithm.

In [5], Lin et. al identified subjective information using automated tools and a novel probabilistic modelling framework called joint sentiment/topic model, which detects sentiment and topic together from text. They concluded that the proposed JST model was fully apart as compared to other machine learning approaches. Basically, they proposed this model on movie dataset to classify the sentiment polarity and to improve the sentiment classification accuracy. In this paper, a joint sentiment/topic (JST) model had been proposed with the help of which document level sentiment classification could be depicted and mixture of topics from text simultaneously could be extracted. On the other hand, existing approaches in sentiment classification were based on supervised learning, while the proposed JST model was fully unsupervised, hence comes up with more flexibility and could be easily combined with other applications. When the results were compared with existing supervised approaches then they found out that this model gave a competitive performance in document level sentiment classification. On other side it also had one limitation of classifying each document as a bag of words which results in ignoring the word ordering for example predicting sentiment of “not good movie” being positive and of “not bad movie” being negative. This leads them to include bigrams and trigrams in their model. Another step which would be included in future was to detect the polarity of text at several granularity levels, e.g. detecting sentiment labels for more fine-grained topics. Model performance on datasets from different domains were also evaluated.

In their paper, Li et. al [6] studied online forums hotspot and forecast using sentiment analysis and text mining approaches. First of all, to inspect the sentiment polarity for each piece of text, an algorithm was created. Afterwards to develop unsupervised text mining approach the algorithm was joined with k-means clustering and support vector machine (SVM). Described text mining approach had been used to group forums into various clusters, whose centre represent a hotspot forum within the current time span. The datasets had been taken from SINA sports forum. Experimental results showed that SVM forecasting gets high consistent results with k-means clustering. The top 10 hotspot forums given by SVM forecasting resembles 80% of k-means clustering results. Both SVM and k-means achieved the same results for the top 4 hotspot forums of the year. In this paper they had created an algorithm that automatically analyze the sentiment polarity of a text, with the help of which text values were obtained. Influential power of text was represented by absolute value and sentiment polarity by the sign of text. Previously created algorithm was then combined with k-means clustering and SVM classification to integrated approach for online sports forums cluster analysis. Unsupervised algorithm had been applied to group the forums into various clusters, whose center represent hotspot forum with the current time span. In addition to clustering the forums based on data from the current time window, forecasting for the next window was also done by them. Proof for existence of correlations between post text sentiment and hotspot distribution was given by empirical studies. Results showed that both SVM and k-means produce consistent natural groupings. Several companies could be benefited from these hotspot predicting approaches in different ways. These companies could also combine results for market basket analysis to yield comprehensive decision support information. A firm in financial sector or the financial department of a giant company might get profit from such a sentimental and text mining process. In financial market, right before a security market opens and trading begins, analysts people on sales and trading desks usually try to get an overall fix on market sentiment and for particular investments. First, algorithm design could be improved to yield a more accurate calculation of sentiment. Even for supervised learning, algorithms other than SVM, or variations of SVM, could be joined as well. Secondly, they had incorporated topic extraction. Third, a practical system, in the form of a website portal, was desired as their major future work.

In [7] a global structured model was investigated to

classify the sentiment of texts on different levels of granularity and also true solutions were confirmed by classification techniques on which assumption of the model was based with an advantage of allowing classification decisions from one level to another. On experimenting McDonald et. al realized that this model somewhat lower the classification level. In their paper, they described model with controllable assumption for sentence document analysis which gain more precision as compared to classifiers that were trained alone as well as cascaded system. It was also suggested that nested hierarchical structure would be more helpful.

[8] represented that in conveying different views of public, opinion mining was under consideration because of many web sources and it classifies the opinion either in positive or negative category. In this, Saleh et. al tested the various fields of text files either by SVM or by certain weighting schemes. The major purpose was to check the collections and for this they proposed a different collection which form a beneficial source to detect the opinion mining. They concluded that SVM was the better means to handle sentiment classification. For other tasks, they wanted to analyze the reaction on reviews and explore the outermost experience similar to Senti-word net. [9] showed that for text classification different alternatives of machine learning algorithms show large variation in their performance. They also represented that bigram show constant improvements in tasks and Naive Bayes (NB) was more preferable than SVM for small part in sentiment tasks. Also, a new SVM alternative shows constantly better results on datasets and due to this information they demonstrated NB and SVM alternatives. From this paper it was concluded that Multinomial Naive Bayes (MNB) was more preferred on sentiment analysis tasks. They concluded that SVM was more preferred on long reviews and advantages of bigrams depends on the sentiment tasks. They also concluded that NB-SVM was a strong operator and Bernoulli Naive Bayes (BNB) was worst performer than MNB.

In [10] two problems were tested: (1) determining whether the given document was a review or not, and (2) classifying the polarity of a review as positive or negative. It was also proved that review identification could be performed with more precision using only unigrams as features. Ng et. al then examined the role of four types of simple linguistic knowledge sources in a polarity classification system. Task of document level sentiment analysis were examined in this paper. Review identification and polarity classification, two problems in document level sentiment analysis were examined. They observed that review identification could be achieved with great accuracy (97-99%) with the help of SVM classifier. They then studied about the several linguistic knowledge sources in polarity classification. They found that bigrams and trigrams if chosen with respect to the weighted log-likelihood ratio as well as manually tagged information, can be quite useful.

In [11] sentiment analysis was used to do natural language processing with the help of which polarity of text document was detected. Initially only positive and negative sentiments were discriminated i.e. binary classification problem. Various machine learning techniques were matched with this problem. They used IMDB dataset to get the results which were easily reproducible. Researchers could also combine their developments which would be useful for further advancements. A simple and powerful method was proposed for sentiment analysis. They had joined three conceptually different baseline models: first one based on language models, second one based on consecutive models of sentences and the last one based on quick reweighing of BOW (Bag of Words). This paper helped in determining how to use this in standard generative language models. They included a code which is available at http://github.com/mesnilgr/iclr15.

In [12], Tripathy et. al represented that the reviews and blog datasets obtained from the social networking sites were unsystematic and need classification for a meaningful information. They could be classified as positive, negative and neutral with the help of supervised machine learning methods. In this paper, for classification of sentiments, they introduced four different machine learning algorithms i.e. NB (Naive Bayes), ME (maximum entropy), SGD (stochastic gradient descent) and SVM (support vector machine) based on precision, recall, F-measure, and accuracy. This paper helped in classifying the movie reviews using supervised machine learning algorithms which was further applied on IMDB dataset using n-gram approach. They concluded that in n-gram approach as the value of n increases, the classification accuracy decreases. It was also concluded that combination of TF-IDF and count vectorizer techniques helps in obtaining better accuracy. On further studying they also came across some limitations as small size of twitter comments, reviews or comments including punctuation symbols and words like “greatttt, fineee” as they don't have proper meaning. So, new list of words was prepared for classification after removing the stop words to select the best feature. For better accuracy hybrid machine learning techniques were also considered.

In [13] Zheng et. al paid attention to the Chinese online reviews since they were directly affected by feature selections which includes n-char-grams and n-pos-grams as potential sentiment features. To select feature subset and to count feature weight enhanced document frequency method and Boolean weighting methods were used. On experimenting with chi-square test, they concluded that the model obtained more accuracy when 4-pos-grams were used to extract features and when taking n-char-grams as feature then low order n-char-grams performed better than high order. From their paper, they concluded that accuracy decreases as the order of n-char-grams increases. Also the noun, adjectives, adverbs and verbs on combining gave better performance and if adjectives were selected as features then it would be better to follow n-pos-grams as features. They also concluded that sentiment analysis of Chinese reviews improved using document frequency method. Further research focus on sentence level since in paragraph level both

positive and negative sentences were considered and the extracting features based on grammatical structures was proposed. In social media defected products had a disastrous impact so their detection could protect the customers from losses. Earlier the automated defect discovery achieve success but still there had been no application to home appliances. So in [14] Law et. al extended their study on underperformance in large home appliances mainly the dishwashers. The domain specific sparkle and domain specific smoke term dictionaries had strong impact on dishwasher defects. Their research was very useful for improving the quality of dishwasher appliances. The authors conducted different experiments to detect the defects in the products. From first experiment they concluded that to detect the defects Afinn lexicon was used but the remaining sentiment analysis techniques performed better than unigram, bigrams and trigrams. The smoke term dictionary was very helpful in exposing the defects which were not found by sentiment analysis. From the second experiment they concluded that in discovering the defects logistic regression, neural network and decision tree classifiers performed better. Domain specific terms implies that the user was satisfied with the product category, design or its use and these terms were having high effect on all the models as compared to component specific and outcome specific features. Neural network classifiers gave the best results and the negative online reviews had adverse effect on sales, brand reputation and company profits.

In [15], Nguyen et. al proposed a new feature type to check its contribution in document level sentiment analysis. They attained best results on dataset produced by Pang and Lee (2004) containing 2000 reviews with 91.6% accuracy than by Maas et al. (2011) containing 50000 reviews with 89.87% accuracy. They also got result on dataset containing 233600 reviews with 93.24% accuracy. In this paper an experimental study on sentiment polarity classification had been conducted by them. First of all, a rating based feature had been described which was based on regression model, AND learned from external independent dataset of 233600 movie reviews. Afterwards contribution of both machine learning and rating based criteria were used to achieve accuracy of 91.6% and 89.87% on the datasets from different domains. These results showed that rating based feature was more efficient for sentiment classification on polarity reviews. Performance could also be improved by adding bigram and trigram features.

In [16], Tiwari et. al demonstrated content based approach for the online audits, film ratings etc. using sentiment analysis. These reviews were grouped by supervised machine learning strategies. For conclusions three different machine learning calculations were considered i.e. SVM, ME, NB and these conclusions were based on parameters such as accuracy, review, f-measure and precision. In this paper for classifying the film reviews of rotten tomatoes dataset using n gram method different machine learning techniques had been suggested. The authors also concluded that on comparison with other research works their output obtained better accuracy.

In [17], Arun et. al, represented sentiment analysis on tweets for demonetization. Firstly, they accessed the data and then converted it into text files as input dataset. Then sentiment analysis was performed after removing the stop words followed by determining the polarity of the words and classifying the tweets as positive and negative. So a new method was suggested for sentiment analysis on demonetization and for this process data cleaning, bigrams, polarity, sentiment scores and graphical methods were used.

In [18], a comparative analysis of different approaches for sentiment analysis and topic detection of Spanish tweets was presented with classification tasks. For classifying Spanish tweets according to sentiment and topics various experiments had been performed. Use of stemmers and lemmatizers, n-grams, word types, negations, valence shifters, link processing, search engines, special Twitter semantics (Hashtags), and different classification methods had been evaluated which was represented a detailed and complete study. The first conclusion that Anta et. al drew was that due to their brevity and lack of context tweets were very hard to deal with. These results proved that for analyzing and classifying the Spanish text it was possible to use classical methods. Best accuracy that was seen was 58% for topics and 42% for sentiments classification.

In [19], Basha et. al represented that because of the popularity of E-commerce product reviews for a product were also growing rapidly with an exponential factor. To make a decision among multiple option where time and money were precious, other people opinions would play an important role. Now most of the organizations had opinion mining and sentiment analysis as a part of their research. Also, almost every business was influenced by the social media websites and blogs which led these companies to do sentimental analysis. In this paper they had used fuzzy rule based systems (FRBS) with models, namely: Mamdani, and Takagi Sugeno Kang (TSK) using FRBS package in R. They also compared these models with other classification methods in terms of precision, Recall and F-measure, accuracy and performance of the method. Experiments on the proposed algorithm for calculating emotions and opinions regarding the product were conducted and also demonstrated the R package. Also some examples of the usage of the package and comparison to other packages had been made.

In [20], Alomari et. al represented that Arabic tweets pose a good opportunity for opinion mining research but they were delayed due to shortage of sentiment analysis resources or challenges in Arabic language text analysis. It included Arabic Jordanian twitter corpus in which either the tweets were denoted as positive or as negative and these tweets were examined using different supervised machine learning approaches. For using different weight schemes, stemming and n-grams techniques experiments were conducted which showed that SVM classifier using TF-IDF through bigrams feature was better as compared to Naive Bayesian classifier. The main objective was to examine the machine learning approach for Arabic sentiment analysis. Firstly, the authors collected a new available Arabic tweets corpus containing 1,800 tweets written in Jordanian dialect. Then they

compared the two machine learning algorithms (SVM and NB) using various n-grams with different weighting schemes and applying stemming techniques. After experimenting they finally concluded that SVM classifier using stemmer with TF-IDF weighting scheme through bigrams showed 88.72% accuracy and 88.27% f-score, their model performed better than other Arabic sentiment analyses research results.

3. Conclusion

The major applications of text mining widely include network mining, natural language dispensation, information recovery and information extraction. In this paper, we survey few representative work such as entity recognition and relation extraction and information extraction. In this paper we also discuss the sentiments of Spanish tweets, Arabic tweets and many more languages. For text mining and sentiment analysis, the major steps required are data acquirement, data conversion, feature representation, feature extraction and different machine learning algorithms. We also extensively show the results of various supervised and unsupervised sentiment analysis techniques to efficiently detect the sentiments.

Future Work

Future work includes extensive comparison of different text mining and sentiment analysis approaches on different data sets acquired from multiple resources and in multiple languages. We will also work towards finding most computationally inexpensive algorithms for various tasks and sub-tasks. Various prediction applications will also be studied.

References

[1] Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A Survey of Arabic Text Mining. In *Intelligent Natural Language Processing: Trends and Applications* (pp. 417-431). Springer, Cham.

[2] Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Icwsm*, *11* (538-541), 164.

[3] Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3* (2), 143-157.

[4] Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining* (p. 5). ACM.

[5] Lin, C., & He, Y. (2009, November). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 375-384). ACM.

[6] Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems*, *48* (2), 354-368.

[7] McDonald, R., Hannan, K., Neylon, T., Wells, M., & Reynar, J. (2007). Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 432-439).

[8] Saleh, M. R., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, *38* (12), 14799-14804.

[9] Wang, S., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 90-94). Association for Computational Linguistics.

[10] Ng, V., Dasgupta, S., & Arifin, S. M. (2006, July). Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 611-618). Association for Computational Linguistics.

[11] Mesnil, G., Mikolov, T., Ranzato, M. A., & Bengio, Y. (2014). Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *arXiv preprint arXiv:1412.5335*.

[12] Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117-126.

[13] Zheng, L., Wang, H., & Gao, S. (2018). Sentimental feature selection for sentiment analysis of Chinese online reviews. *International journal of machine learning and cybernetics*, *9* (1), 75-84.

[14] Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, *67*, 84-94.

[15] Nguyen, D. Q., Nguyen, D. Q., Vu, T., & Pham, S. B. (2014). Sentiment classification on polarity reviews: an empirical study using rating-based features. In *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 128-135).

[16] Tiwari, P., Mishra, B. K., Kumar, S., & Kumar, V. (2017). Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, *7* (1), 30-41.

[17] Arun, K., Srinagesh, A., & Ramesh, M. (2017). Twitter Sentiment Analysis on Demonetization tweets in India Using R language. *International Journal of Computer Engineering in Research Trends*, *4* (6), 252-258.

[18] Anta, A. F., Chiroque, L. N., Morere, P., & Santos, A. (2013). Sentiment analysis and topic detection of Spanish tweets: A comparative study of of NLP techniques. *Procesamiento del lenguaje natural*, *50*, 45-52.

[19] Basha, S. M., Zhenning, Y., Rajput, D. S., Iyengar, N., & Caytiles, D. R. (2017). Weighted Fuzzy Rule Based Sentiment Prediction Analysis on Tweets. *International Journal of Grid and Distributed Computing*, *10* (6), 41-54.

[20] Alomari, K. M., ElSherif, H. M., & Shaalan, K. (2017, June). Arabic Tweets Sentimental Analysis Using Machine Learning. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 602-610). Springer, Cham.