Subjectivity Detection for Sentiment Analysis on Twitter Data

Sindhu Chandra Sekharan
SRM Institute of Science and Technology

Subjectivity Detection for Sentiment Analysis on Twitter Data

C. Sindhu, Binoy Sasmal, Rahul Gupta, and J. Prathipa

Abstract With the quick increment in the quantity of web clients, the Internet has an enormous measure of data produced by the clients. Many people share their views regarding a topic on social media platforms such as Facebook and Twitter and give their feedback or review about a product on e-commerce web sites such as Amazon and Flipkart which leads to a huge amount of data. The identification of subjective statements from the data is known as subjectivity detection. To automate the analysis of such data, sentiment analysis is used. The aim is to find the opinionative data and classify it according to its polarity, i.e. positive, negative or neutral feedback, known as sentiment classification and then analysing it which is known as sentiment analysis. However, before performing sentiment examination, the information is exposed to different pre-processing procedures which finally give the desired optimized output. This allows us to get to know about the public's mood or opinion about a particular topic. This summarization helps the concerned organization or public to improve their product or service based on the feedback received.

Keywords Twitter · Sentiment analysis · Subjectivity detection · Opinion · Corpus

1 Introduction

With the enormous development of number of web users, the quantity of tweets every day on Twitter has likewise expanded definitely. Mining of sentiment from these tweets is helpful for the organizations and associations. For instance, it very well may be utilized as a sub-module in suggestion motors and so forth.

Sentiment analysis [1] is relevant mining of content that plans to group content into positive, negative and impartial. Sentiment analysis is an issue which incorporates different NLP sub-problems which are to be settled which incorporate mockery identification, element acknowledgement and subjectivity recognition and so forth.

C. Sindhu (✉) · B. Sasmal · R. Gupta · J. Prathipa
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India
e-mail: sindhucmaa@gmail.com

D. J. Hemanth et al. (eds.), *Artificial Intelligence Techniques for Advanced Computing Applications*, Lecture Notes in Networks and Systems 130,
https://doi.org/10.1007/978-981-15-5329-5_43

Subjectivity detection [2] has picked up significance with the quick development of data created via web-based networking media which requires the identification of subjective data (opinion) and objective data (fact). Subjectivity identification can be substantially more testing than polarity recognition; however, it has been under-explored because of the supposition that most of the information via web-based networking media is objective. For example, “My favourite pair of shoes is sold out” is an objective statement because it is a fact, and “This pair of shoes is very good” is a subjective statement because it tells about the opinion of the person.

Subjectivity detection helps to get to know about the opinion of the users about a particular product and topic which indeed helps the concerned organization or public to improve their service or product dependent on the feedback received.

2 Related Work

Systematic literature audit process is used in this overview. First we scanned for some related papers, research reports that are comprehensively worried about subjectivity identification or opinion mining from the content.

Detection of user’s opinion and classifying its polarity, i.e. positive, negative and neutral, is known as polarity detection (PD) [3]. Previously done work in sentiment analysis was either knowledge-based or sentiment-based. But recently there have been various studies that utilize various machine learning techniques to classify the text. Supervised machine learning techniques are comparatively better than unsupervised machine learning techniques in performance but it is expensive to acquire the huge amount of labelled data required for supervised learning, whereas it is comparatively easy and less expensive to acquire unlabelled data for unsupervised learning.

Numerous specialists are putting their endeavours to identify the best technique for subjectivity identification. Albeit, a portion of the algorithms give great outcomes such as support vector machine (SVM), maximum entropy, Naïve Bayes [4] and so forth; however, no technique can resolve every one of the difficulties. The vast majority of the researchers detailed that SVM has high precision [5] than different algorithms. The different algorithms and the data sets used in different papers have been mentioned in Table 1.

3 Subjectivity Detection Approach

A framework was implemented in which the first step is to classify messages as subjective and objective tweets (subjectivity detection). The second step is to classify the subjective tweets into positive and negative (polarity detection).

Usually, a purely objective sentence does not carry any sentiment, and a purely subjective sentence usually tends to lean towards a positive or a negative sentiment.

Table 1 Subjectivity detection tasks in sentiment analysis

Paper	Data set	Algorithms	Features
[6]	Own Twitter data	SVM	Meta-features (part of speech (POS), polarity-MPQA)
[7]	Manually annotated tweets	corpus-based, dictionary-based, log-liner regression	WordNet, POS
[8]	Own Twitter data	SVM	unigrams, emoticons, hashtags, lexicon [30]
[9]	Own Twitter data	FCA	WordNet, OpenDover
[10]	SemEval-2013	clustering-based word sense disambiguation (WSD), lexicon-based classifier	WordNet, SentiWordNet
[11]	SS-Tweet	Senti strength	polarity, negations, emphatic lengthening
[12]	Twitter data	Unigram, bigram, uni-bigram	POS, SentiWordNet
[13]	Movie reviews	Statistical, maximum entropy	Emoticons, negations
[14]	Starbucks twitter data set	Dynamic architectural artificial neural networks	Polarity, hashtags
[15]	Twitter data	SVM, Naïve Bayes, maximum entropy, hybrid approach	Unigram feature
[16]	Spatiotemporal social (STS), healthcare reform (HCR) data set	LexRatio, maximum entropy, LProp, N-gram, hashtags, emoticons, lexicon [3]	Emotion detection, Twitter follower graph
[17]	Customer review Twitter sata set	Naive Bayes, maximum entropy, SVM	WordNet, sentiment classification

Though there are a few exceptions, for example, “The food made me sick” is an objective sentence with sentiment, and “I believe he came to the college yesterday” is a subjective sentence with no sentiment. Classifying a sentence as objective or subjective is done by using libraries such as TextBlob created by Steven Loria, and tools such as Opinion Finder (http://mpqa.cs.pitt.edu/opinionfinder/). A filtering mechanism is also implemented to have a control on the level of subjectivity in the training set by using a subjectivity threshold.

Another approach that was previously implemented, which was later scraped due to inconsistencies in the results and lack of accuracy, was from a given tweet, we map its POS using a POS dictionary (http://wordlist.sourceforge.net/pos-readme). POS tags are used to indicate sentiment tagging in a tweet. Objective messages usually consist of adjectives or interjections. We get the prior subjectivity and polarity

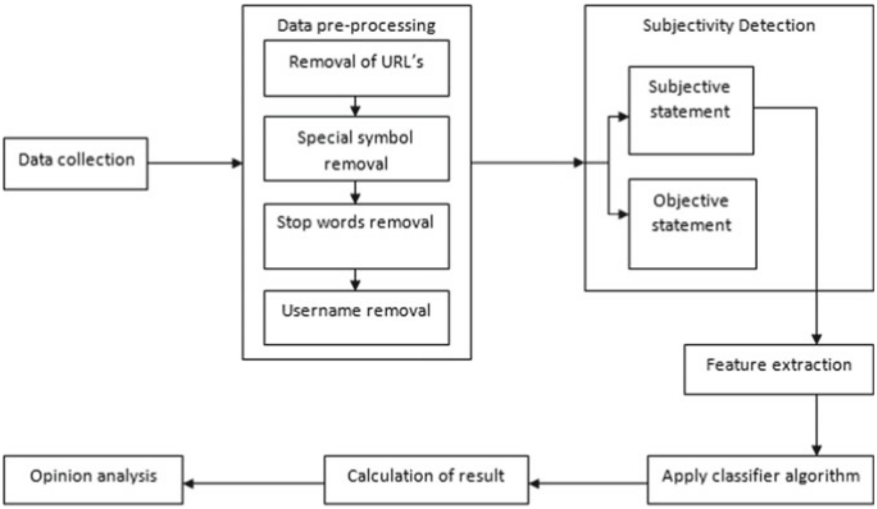

Fig. 1 Subjectivity detection for sentiment analysis on Twitter data

Table 2 Data set annotation scheme

Sentiment	Annotation
Positive	4
Neutral	2
Negative	0

information from the subjectivity lexicon used in [18]. The following methodology in Fig. 1 is adopted for the subjectivity detection.

3.1 Data Collection

The data set that is used consists of around 1.6 million tweets for training and 5000 tweets for testing [19]. The tweets in the data set are categorized into positive, negative and neutral. The data set is very versatile and consists of various categories such as company, movie, location, person, product, event and misc. The emoticons were removed for the training and the test data (Table 2).

3.2 Data Pre-processing

The data extracted from Twitter contains various contents which do not contribute to the sentiment of the user; therefore, it has to be first pre-processed. Pre-processing

[20] includes four basic steps—removal of URL, removal of special symbols, removal of stop words and removal of username. In removal of URL, any kind of link which is tweeted by the user and does not contribute to the sentiment analysis is removed. Removal of special symbol step deals with removing various symbols which do not have any actual sentiment, e.g. full stop (.), punctuation mark (!) and so forth. Stop words [21] removal step removes the stop words, words such as a, the which do have no effect on sentiment analysis should be removed and the conversion of emoticons to its equivalent word. Finally, in the username removal step every user's username starts with @ which has no effect on the sentiment analysis is removed, e.g. @username.

3.3 Subjectivity Detection

As previously mentioned, the first step is to classify the tweet into subjective and objective and remove the tweets based on their subjectivity scorekeeping only the tweets having score higher than the specified threshold.

This step is introduced to achieve higher accuracy. The pre-processed data is taken and is classified into subjective or objective statement using a subjectivity classifier. All the tweets having a subjectivity score lesser than the specified threshold are filtered out, and the classifier is trained with only the remaining tweets. It is observed that as the subjectivity threshold is increased, significant amount of tweets gets filtered out.

3.4 Feature Extraction

A data set contains numerous ascribes that add clamour to the data and influence exactness. The commotion likewise bit by bit expands the time required to assemble the model. Feature extraction basically combines ascribes into a reduced feature set. The selected features and their blend assume a significant job for identifying the sentiment of the text.

Selection of features [22, 23] from the extracted features can possibly improve the arrangement exactness, restricted in on a key feature subset of opinion discriminators and give more prominent understanding into habitually happening ascribes and qualities.

The extracted features focus on a document vector whereupon machine learning strategies are applied to group the extremity of the content utilizing the got document vector.

3.5 Apply Classifier Algorithm

There are generally three approaches which include: **Supervised learning** [24, 25] is a sort of learning in which we train the machine with the information which is well labelled. The machine learning approach pertinent to sentiment examination, for the most part, belongs to supervised classification. In machine learning-based methods, two sets of records are required: training set and a test set. Machine learning techniques such as naïve Bayes, SVM, maximum entropy and so forth are used. **Unsupervised learning** [26] is a sort of learning wherein we train the machine with the information which is not labelled. Classification is performed by comparing the features of a given text with sentiment lexicons whose sentiment values are determined prior to their use. Clustering methods such as k-means, mean shift clustering and so forth are used. **Reinforcement learning (**RL) [27] is the field that reviews the problems and procedures that attempt to retro-feed its model to improve. To achieve this, RL needs to be able to "sense" signals, consequently choose an activity and afterwards look at the result against a "reward" definition. RL attempts to make sense of what to do to boost these prizes, yet it does this without any support.

Supervised learning methods for classification by using machine learning [28] algorithms such as Naive Bayes, SVM and maximum entropy have been found to give good accuracy. SVM was used as vast majority of researchers claimed it to be more accurate than the other algorithms, so we decided to use SVM to build the classifier.

3.6 Calculation of Result

Calculating the polarity of the user's statement using the approach described. The most generally used assessment measurements are accuracy, recall, precision and *F*-score. The confusion matrix is shown in Table 3.

Table 3 Confusion matrix showing the performance of a sentiment analysis method

	Is positive	Is negative
Positive prediction	TP	FP
Negative prediction	FN	TN

4 Evaluation

First we see the effects of the subjectivity threshold parameter. From the results obtained, it can be clearly observed that the tweets get filtered out to an ever-increasing extent with an increase in subjectivity threshold parameter as shown in Fig. 2.

Here we see the tweets remaining after the filtering process from TextBlob and Opinion Finder tool. TextBlob utilizes a function that finds a tweet's subjectivity level, whereas Opinion Finder tool denotes which segments of the message are subjective. This helps us to find the level of subjectivity of a tweet. SentiOutlook is used, which we created, to find the best-suited filtering for our experiment.

$$\text{Subjectivity Level} = \frac{\text{Length of subjective parts}}{\text{Total length of the tweet}} \tag{1}$$

For the experiment, we pick an optimal threshold value of 0.5, factoring that the model should be trained on a progressively conventional data set and the subjectivity level can be calculated using (1). The relation of the accuracy with subjectivity threshold can be seen in Fig. 3.

Using the SVM classifier, we obtained satisfactory accuracy, precision, recall and *F*-score as shown in Table 4.

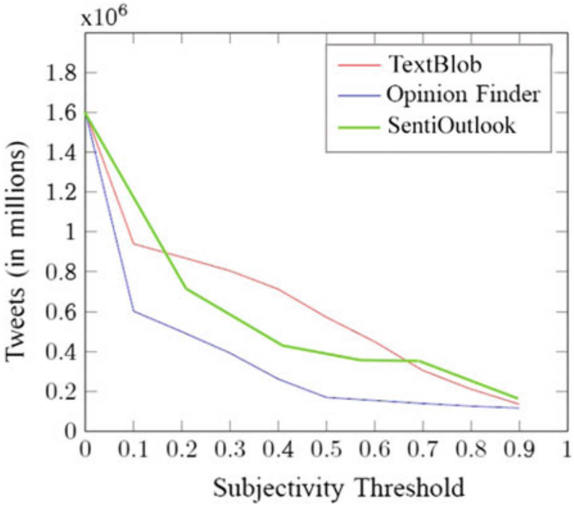

Fig. 2 Remaining tweets with subjectivity threshold

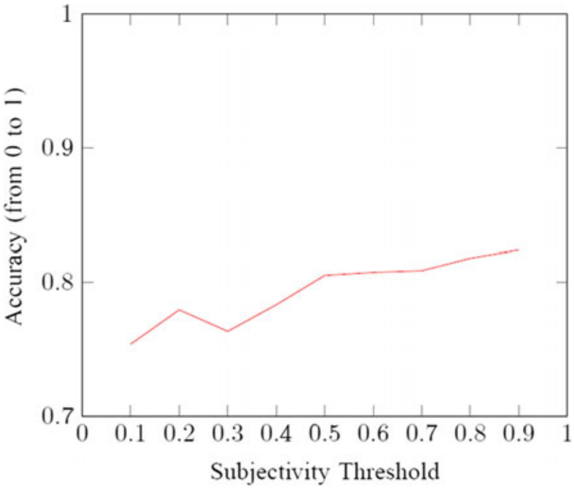

Fig. 3 Accuracy against subjectivity threshold

Table 4 Performance matrices using SVM

Matrices	Values	Formula used
Accuracy	82.66	$\frac{\text{TN+TP}}{\text{TN+FN+TP+FP}}$
Precision	78	$\frac{\text{TP}}{\text{TP+FP}}$
Recall	86	$\frac{\text{TP}}{\text{TP+FN}}$
F-score	84	$\frac{2\times\text{precision}\times\text{recall}}{\text{precision+recall}}$

5 Discussion

Twitter has a large amount of data in the form of tweets which includes the comments, opinions and reviews of the public regarding a particular product or service. Therefore, sentiment analysis comes into play to mine the opinion of the users. Many researches have been done on this but there is still a lot of scope in increasing the accuracy of the system. We came across various techniques which can be used to improve the accuracy but hardly any work is accomplished [29] on them such as oxymoron words, misspelled words, etc., and these problems should be considered in any future work done on this topic. Also, the example we took in the introduction "This pair of shoes is very good" is actually a subjective statement; however, our system detects it as an objective statement. It will also be quite interesting to go beyond just the positive and negative and extract more information and patterns from these data.

6 Conclusion

The growth of social data is exponential, which has given rise to new aspects, such as the subjectivity detection. Subjectivity detection is a natural language processing task that consists of differentiating subjective data (opinions) from objective data (facts). By using subjectivity detection, we can filter out the tweets that are objective and find the tweets that are subjective and carry out sentiment analysis only on the subjective data. The accuracy of the sentiment analysis can further be increased by implementing methods to fix the misspelled words and correcting the use of any abbreviated or shortened words. The use of oxymoron words is another factor that can affect the accuracy of sentiment analysis which can be corrected by using a new model to detect and replace the oxymoron words with equivalent words that are effectively analysed by the sentiment analyzer.

References

1. Neri F, Aliprandi C, Capeci F, Cuadros M, By T (2012) Sentiment analysis on social media. In: IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 919–926
2. Satapathy R, Chaturvedi I, Cambri E, Ho SS, Cheon Na J (2017) Subjectivity detection in nuclear energy tweets. Computacion y Sisttemas 21(4):657–664
3. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the human language technology conference and the conference on empirical methods in natural language processing (HLT/EMNLP), pp 347–354
4. Parveen H, Pandey S (2016) Sentiment analysis on Twitter dataset using Naive Bayes Algorithm. In: IEEE trans 2nd international conference on applied and theoretical computing and communication technology (iCATccT), pp 416–419
5. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing, vol 10, pp 79–86
6. Barbosa L, Feng J (2010) Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd international conference on computational linguistics: posters (COLING'10). Association for Computational Linguistics, Stroudsburg, pp 36–44
7. Kumar A, Sebastian TM (2012) Sentiment analysis on Twitter. Int J Comput Sci 9(4):372–378
8. Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B (2011) Combining Lexicon-based and Learning-based Methods for Twitter sentiment analysis. Technical report, HP Laboratories
9. Kontopoulos E, Berberidis C, Dergiades T, Bassiliades N (2013) Ontologybasedsentiment analysis of Twitter posts. Expert Syst Appl 40(10):4065–4074
10. Ortega R, Fonseca A, Montoyo A (2013) SSA-UO: unsupervised twitter sentiment analysis. In: Proceedings of the 7th international workshop on semantic evaluation—2nd joint conference on lexical and computational semantics (SemEval'13). Association for Computational Linguistics, pp 501–507
11. Thelwall M, Buckley K, Paltoglou G (2012) Sentiment strength detection for the socialweb. J Am Soc Inform Sci Technol 63(1):163–173
12. Gurkhe D, Rishit B (2014) Effective sentiment analysis of social media datasets using Naive Bayesian classification
13. Duric A, Song F (2012) Feature selection for sentiment analysis based on content and syntax models. Decision Support Syst. 53:704–711

14. Saif H et al (2016) Contextual semantics for sentiment analysis of Twitter. Inf Process Manag 52:5–19
15. Bahrainian SA, Dengel A (2013) Sentiment analysis and summarization of twitter data. 2013 IEEE 16th international conference on computational science and engineering (CSE) IEEE
16. Speriosu M, Sudan N, Upadhyay S, Baldridge J (2011) Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the first workshop on unsupervised learning in NLP (EMNLP'11). Association for Computational Linguistics, Stroudsburg, PA, pp 53–63
17. Gautam G, Yadav D (2014) Sentiment analysis of Twitter data using machine learning approaches and semantic analysis. In: 2014 seventh international conference on contemporary computing (IC3), IEEE
18. Riloff E, Wiebe J (2003) Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 conference on empirical methods in natural language processing (EMNLP-03), pp 105–112
19. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, p 12
20. Jianqiang Z, Xiaolin G (2017) Comparision research on text preprocessing methods on twitter sentiment analysis. IEEE Trans 5:2870–2879
21. Saif H, Fernandez M, He Y, Alani H (2015) On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: Proceedings of 9th Language Resources Evaluation Conference (LREC), Reykjavik, Iceland, 2014, pp 80–81
22. Mansour R, Hady MFA, Hosam E, Amr H, Ashour A (2015) Feature selection for twitter sentiment analysis: an experimental study. Computational linguistics and intelligent text processing: 16th international conference, CICLing 2015, Cairo, Egypt, April 14–20, 2015, Proceedings, Part II, Springer International Publishing, pp 92–103
23. Chandrasekhar G, Sahin F (2014) A survey on feature selection methods. Comput. Elect. Eng. 40:16–28
24. Dhanalakshmi V, Dhivya B, Saravanan A (2016) Opinion mining from student feedback data using supervised learning algorithms. 1–5. https://doi.org/10.1109/icbdsc.2016.7460390
25. Read J (2005) Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of ACL-05, 43nd meeting of the association for computational linguistics. Association for Computational Linguistics
26. Ko Y, Seo J (2000) Automatic text categorization by unsupervised learning. In: Proceedings of the 18th conference on computational linguistics, vol 1. Associations for computational Linguistics, pp 453–459
27. Frenay B, Verleysen M (2016) Reinforced extreme learning machines for fast robust regression in the presence of outliers. IEEE Trans Cybern. 46(12):3351–3363
28. Chaudhari M, Govilkar S (2015) A survey of machine learning techniques for sentiment classification. IJCSA 5(3):13–23
29. Patil H, Atique M (2015) Sentiment analysis for social media: a survey, pp 1–4. https://doi.org/10.1109/icissec.2015.7371033
30. Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of the conference on web search and web data mining (WSDM)