IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper*  **Volume 11, Iss 7 , 2022**

# Hybrid Classification algorithm for Feature Based Sentimental Analysis on Product Reviews

E. Sreedevi

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India.

**DOI : 10.48047/IJFANS/V11/ISS7/313**

**Abstract:**One of the main challenges of the NLP is emotional representation or perception mining (natural language processing). In this scenario, market analytics play a critical role in identifying that people want to expand their business. In fact, these people use inputs from goods that customers have used and based on their inputs and feedback give a clear-cut idea for the business people how to withstand in the present market and also gives an outstanding picture of what they should expect in the future. With few words or phrases, the results will be chosen. Thus, these individuals aim to boost their market by offering premium goods to their customers for attaining maximum benefit. Sentimental analysis has also gained a lot of interest in recent years. SA is an NLP analysis area used within a certain characteristic text to categorise opinion or perception. The data set includes a number of algorithms for machine learning, and results are compared with the Decision Tree, Naïve-Bayes classifiers that are evaluated according to such criteria as recall, precision and F-score. The article is based on a range of methods of classification in order to decide whether or not an individual is unwanted, constructive or impersonal in terms of his or her opinions, and forecasts a product's star ranking. There are also two specialised approaches such as the classification of features followed by the classification of polarisation along with test findings. Finally, a comparative study is conducted in this paper between 3 classification methods. 1) Decision Tree 2) Novel Bag-Boost algorithm of classification 3) Naive-Bayer's, where high accuracy is compared to the other two. Where the hybrid Novel algorithm gave high accuracy in comparison with the other two algorithms.

**Keywords:** Sentiment Analysis, Reviews, Machine Learning Techniques, Opinion Mining,Natural Language Processing.

## 1. Introduction

The ability for consumers to express unmediated human views is one of the main factors that has contributed to web 2.0's widespread popularity. Many users post their opinions on products and their services through social media and specialised websites. According to a survey conducted by marketing company comScore, almost 84 percent of Internet consumers have done online research on a product at least once, with up to 88 percent claiming that ratings have a huge impact on their purchasing decisions (Lipsman 2007). The consumers polled indicated that they would be able to pay up to 99 percent extra for products with excellent ratings. According to the new Global Digital Forensics report, cloud storage, social networking, smartphones, handheld devices, computer systems and many more govern the globe. Advanced smartphones will in the near future still be at the forefront of business and peculiar use. This is a primary objective for us to do. Powerful broad data collection and interpretation provides intelligent and practical insight. With the pace of Internet distribution growth, we are able to analyse vast quantities of data and to forecast the needs of customers and potential requirements. This article takes the consumer feedback as inputs i.e. where the hybrid algorithm in the machine learning compared to the other two is incredibly accurate. The key aim of the proposed approach is to provide a criterion for assessment and classification based on the text of the analysis. It is especially important to know the emotional trends and desires of consumers by comments on the online text. Such an analysis of the emotional trend of the customer was

ISSN PRINT 2319 1775 Online 2320 7876
*Research paper* **© 2012 IJFANS. All Rights Reserved**, **Volume 11, Iss 7 , 2022**

called a polarity rating challenge, to illustrate "This mobile is cool but price is extraordinary" castoff part-of-labelling of the emotional trend amongst smartphone users' opinions. In two steps, this is achieved in feature extraction and in the second, polarity is found. In order to carry out classification methods, the effectiveness of the comments and the rating based on the text review have been found. The end user has progressively shifted from a PC to a smartphone as Smartphone grew. The value of user-generated content in decision-making processes is undeniable. According to Kannan et al. (2012), due to the ever-increasing number of different goods with similar characteristics, consumers often seek credible, user-generated feedback to estimate the actual utility value of items. Furthermore, owing to the pervasiveness and intrusiveness of online advertising, users rely on user-generated content as a source of knowledge about goods and services (Chevalier and Mayzlin 2006). According to a recent analysis by Pang and Lee (2008), the interest that consumers devote to user-generated feedback is often affected by the specific product under consideration. To address this problem, we introduce a novel feature-based polarity analysis methodology that combines statistical and natural language processing techniques. First, we introduce a system for automatically identifying and recording the salient features of a device from the user's point of view. Centred on the term frequency, this study mines how much attention the consumer devotes to each function, extracting domain information in a bottom-up process. Second, for each analysis, we estimate the degree of positive or negative polarity with respect to each key attribute to model the user's sentiment. The aim is to fine-tune the models' characteristics [2]. Opinion Mining has a wide variety of applications, starting with user software preferences. Consumers' viewpoints or preferences can be accessed across a variety of platforms, regardless of whether the product or service is a product or a service. Opinion mining is useful in many aspects, for example, it tells the performance rate of a new brand introduced as well as the functionality of the device. For example, a cell phone has many colour features, camera quality, sound, video, and so on. The main goal of this paper is to imagine smart phone star ratings using different ranking methods using data from the Kaggle repository and some usage cases. As a result, smart phone companies will pick their future products and offer higher-quality products to their customers. We used two measures to approximate the rating of smart phones: 1) feature extraction, and 2) polarity classification. The aim of sentiment analysis is to derive views from textual information so that we can understand what people think about particular issues through the analysis of broad data streams, such as personal blogs and social media. A major part of sentiment analysis is an issue of classifications, i.e., a text with a viewpoint, classified as a positive or negative polarity, and techniques of machine learning have already been used to answer this question. There are 2 major approaches to perform sentimental analysis 1) lexicon based and 2) supervised methods. In the 1st approach it calculates the semantic orientation of words by obtaining the polarities of each and every word which is known as SentiWordNet[3]. Whereas the 2nd approach uses a well-known machine learning technique which is used to build a model. In which the 2nd approach is more efficient when compared to the 1st approach so in this paper we used 2nd approach. Generally Word-level polarities are more effective when compared with review-level. So there is need to bridge a gap between these two. In this article we mainly focus on Word-Level rather than review-level. We performed evaluation by using Amazon Reviews on Mobile.csv dataset for extracting the feature and then calculating the polarity of each and every review at sentence level. After so many tests we observed that Novel Hybrid bag-boost algorithm gave high accuracy when compared with the other two.

## 2. Literature Survey

ISSN PRINT 2319 1775 Online 2320 7876
*Research paper* **© 2012 IJFANS. All Rights Reserved**, **Volume 11, Iss 7 , 2022**

The major activity to test any product is to check product reviews and also to study the roles of the product. Diverse tests, for example data sets and executed algorithms, have been released for classifying mobile ratings, including versions. The way we order goods now and the rest at our home, where mobile telephone has revolutionised. More and more people are searching for a commodity when knowledge can be obtained easily. Any buyer may receive information from a customer other than the vendor's updates. These information sources include consumer reviews and information that has already been used in the customer purchase decision cycle. E-commerce firms have provided straightforward review to allow consumers to guide judgements and assessments on other judgements and reviews that relate to positive performance, and satisfy the consumer whether he/she has purchased the right product. We digitally purchase all our daily product by obtaining all information about the product based on its reviews which is available at our doorstep which revolutionized the need of smartphones. Now a days customer is looking for more and more products in online which made them as exposure or awareness becomes better. Supported user feedback and measurement measures are the simplest samples of information on smartphones which play a key role in consumer experience. Online digital promotion, in particular e-commerce, gives competing consumers the greatest opportunity for reviews and ratings, enabling them to get products from multiple customers with confidence and try to choose the product in easiest option.

To perform some statistical analysis on a product review we need some measures which is to be Considered.

1) Finding the product frequencies among assessment and judgments.
2) Find the Co-relation among price and rating.
3) Finding the best brands which have high rating
4) Finding the frequency count of words on reviews which was given by customers.
5) Finding the polarity of each and every review.
6) Finally, to propose a best classification model which produces high accuracy with less time.

It requires multiple measures to detect the general feeling, opinions and goals (Liu (2012)). Meena and Prabhakar (2007) found in a sentence-level emotion study that rules dependent on the atomic feelings of each sentence would help to determine a general feeling of a sentence. However, only adjectives and verbs in Meena at al. function were considered as features that implies only the opinions target can be associated

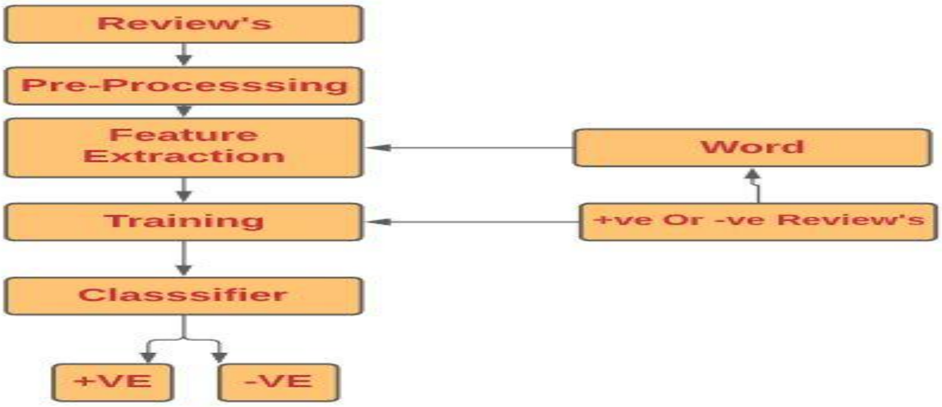


**Figure 1**System Architecture

Figure 1 shows how the results of an enlightenment study include device elements such as removing activities from other social networks, submitting data, presenting tasks and collecting training set. The training set consists of certain tests determined by positive or negative or neutral. By Vader Sentimental analysis the decisions made on the training set and, gives the polarity of each review. This article examines the star's prediction of smartphones and categorizes "reviews" into positive, negative, neutral.

## 3. Hybrid Approach

In particular, this paper aims to determine the smartphone rating, which you can choose from the user feedback from various online social networks. It also uses features like stability to enhance our forecast. In this paper, we contributed and tested whether the neutral positive or negative function was present. In this aspect, the emotional approach of Vader's classification evaluates text and sentence levels. The polarity has been classified and test data have been trained, we created a confusion matrix and measure the precision, recall, and F1score for data accurateness. We used various Machine learning algorithms like (1) Navie-bayes. (2) Random Forest 3) Novel Hybrid Bag-Boost algorithms for this purpose[30](1), Combinations are typically useful to test bias and determine the way you learn[8]. Any solution presents a related problem [9] that divides the statement into two classes, both positive and negative [10]. We first analyse the data in order to remove undesired data from our system. The Lexicon dictionary then describes polarity and its learning algorithm. There are different ways to determine whether a positive, negative or neutral statement is present[30]. On the basis of rules, lexicons, and machine learning procedures, the hybrid paradigm applies a single structure that uses the efficient arenas of separable classifiers while attempting to circumvent their limits. Let's look at a scenario to further understand the concept of ensemble learning. Assume you're a mobile device maker looking to launch a smartphone with certain exclusive features. Currently, you want to build a dummy piece with preliminary feedback before making it official (ratings). When would you be able to do that?

A: You might be prompted to have one of your friends score the phone for you. You'll be just enjoying the guy you chose now and you don't want to break your heart by giving the poor work a 1 or 2-star rating.

B: As a second choice, you might ask 5 of your product's friends to rate it: You could get some genuine feedback for this choice.

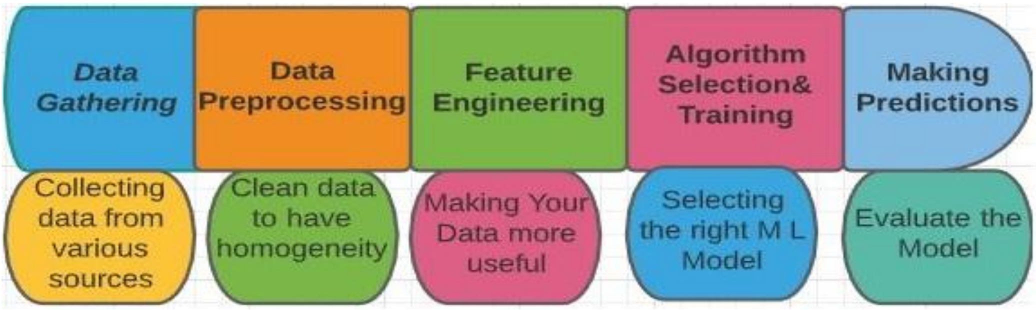


**Figure 2**. Shows how our designed model can be estimated

Bagging = Bootstrap aggregating Bootstrap sampling: given a set X containing N training examples: Create Xj by drawing N examples uniformly at random with replacement from X Expect Xj to omit 37% of examples from X Bagging: Create L bootstrap samples X1,..., XL Train classifier dj on Xj Classify new instance x by majority vote of learned classifiers (equal weights) 4 / 19

For each model m in the b

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper*  **Volume 11, Iss 7 , 2022**

Do k times: (where 'k' is some constant)

Randomly divide the training dataset into two datasets: A, and B.

Train m with A

Test m with B

Select the model that obtains the highest average score

Result: An ensemble of classifiers.

Call XGBoost()

Output of learner is binary hypothesis dj

Compute error pj (dj) = error of dj for some sample inputs taken from X according to pj (can compute exactly)

4 Create pj+1 from pj by decreasing weight of instances that dj predicts accuracy.

## 4. Experimental Results

Classification Techniques for Sentiment Analysis: Many classification techniques were available for predicting the opinion on a review, including feature extraction and polarity classification. As a result, the grouping techniques used in this article are as follows: The Decision Tree, the Naive-Bayes algorithm, and the Novel Hybrid Bag-Boost algorithm are all examples of classification algorithms used in this paper.

### 4.1 Naïve-Bayes Algorithm

Naïve-Bayes algorithm is considered to be one of the easiest and simplest algorithms to implement for getting accurate results. Because of its flexible nature this algorithm is used in wide applications. The main root for this algorithm is Bayes theorem which is based on probability and statistics. According to literature survey, NB is one of the best classifiers which is used in datamining technique. Apart from Uncomplexity, it is known to be one of the best classification algorithms. Bayes theorem provides a way to calculate P(c) of P(c ) , P(x) and P(x) of the posterior probability[18,19,20].

We try to estimate the value of Y which can be obtained by maximizing P(Y=y|X=x) It assumes conditional independence among various features. i.e., Y= [T, F]

P (X1=x1, X2=x2………...Xn=xn | Y=y)

= P (X1=x1|Y=y),P( X2=x2|Y=y)…………...P(Xn=xn | Y=y)

= πi = 1 to n P(Xi=xi | Y=y)

According to Bayes theorem

$$P(y|X) = \frac{P(X|y)\,P(y)}{P(X)}$$

### 4.2 Decision Tree

In Random Forest, a wider range of decision trees is an extension of the common decision tree algorithm. It seeks to reduce the difference between the new decision conventions . The decision tree is built by choosing a random variable set (features). Finally, a random trees set is known as a Random Forest or short RF. Thanks to their higher classification accuracy, RF is one of the most accurate classification algorithms. Another feature of RF is its importance in comparison to other alternative approaches for unbalanced and incomplete data.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

*Research paper*  **Volume 11, Iss 7 , 2022**

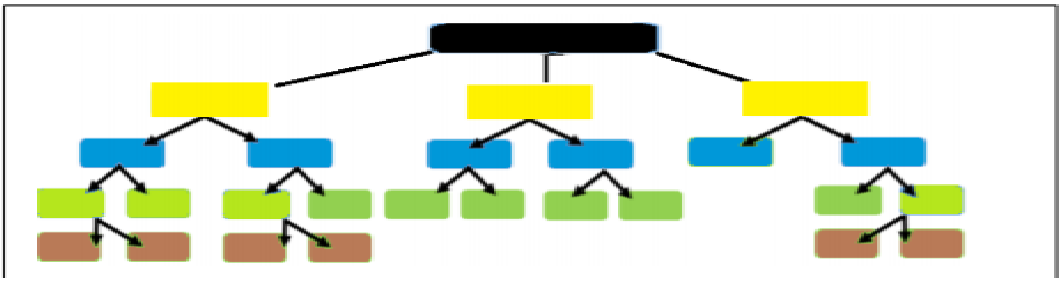

**Figure 3**. Shows RF Tree Feature Selection Technique

A sorting strategy was applied to increase the classification accuracy. This is aimed at reducing the dataset's dimension. In other words, it tries to filter the most relevant features affecting the classification rather than taking into account all variables of the results. Here we have chosen a Sentimental Analysis Feature Selection Technique called SAFSE which seeks to remove the most minor features from the data until the pre-specified number of significant features has been obtained. Due to its effectively definable characteristics, It is simple to configure and manage, and the prediction of the goal variable is significantly related.

4.3 Evaluation Process

In any research problem apart from classification of data, accuracy is the main measure in this paper. In order to quantify the accuracy of classifying sensitivity, specificity and precision were taken as performance indicators which is based on confusion matrix which is as shown in the Figure 4.

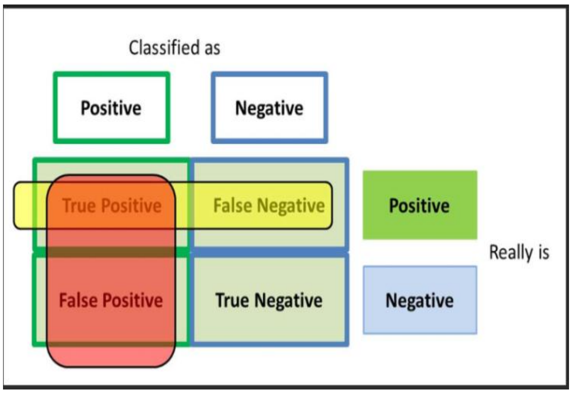


**Figure 4**. Shows Evaluation Process

The assessment measures (Like, Precision, Recall, f1-score etc.) of the classifier can be easily extracted from the obtained Confusion Matrix. Which one in turn used to find the model's accuracy [8, 9, 18], as shown in the figure 9.

Sensitivity represents the percentage of real positive events that the model accurately detected. It focuses on the model's capacity to accurately identify positive cases. The percentage of actual negative events that the model accurately detected is known as specificity. The ability of the model to appropriately reject negative instances is its main concern. The fraction of positive cases that are actually anticipated to be positive is known as precision. It focuses on how well the model predicts the future in a positive way.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TP + FP}$$

ISSN PRINT 2319 1775 Online 2320 7876
**Volume 11, Iss 7 , 2022**

$$Precision = \frac{TP}{TP + FP}$$

Where TP= true positive FN=false negative TN=true negative FP=false positive

Apart from these measures Accuracy and Error are another 2 measures for classification. The total accuracy of the model's predictions is gauged by classification accuracy. It determines what percentage of the total number of instances were correctly categorised. The total accuracy of the predictions made by the model is gauged by its error rate. It determines what percentage of the total number of instances is misclassified.

$$Classification\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN}$$

The research is carried out in evaluations using a hybrid methodology. Other algorithms such as the Naïve Bayes, the decision tree etc. The results are related. In comparison to Decision Tree and Naïve Bayes, the Novel Hybrid Bag Booster Algorithm has a high 95% accuracy. At the start it could be concluded that the hybrid algorithm is more accurate, than other 2 algorithms.

**Table 1**Illustrations the accurateness of Decision Tree Algorithm

| Star rating | Prec | Rec | fl-score | Support |
|---|---|---|---|---|
| 1 | 0.84 | 0.64 | 0.73 | 231 |
| 5 | 0.82 | 0.87 | 0.85 | 995 |
| Microwave | 0.84 | 0.87 | 0.88 | 1226 |
| Macro_avg | 0.88 | 0.81 | 0.84 | 1226 |
| Weighted_avg | 0.88 | 0.87 | 0.87 | 1226 |

**Table 2**Illustrations the accurateness of Naïve Bayes Algorithm

| Star rating | Prec | Rec | fl-score | Support |
|---|---|---|---|---|
| 1 | 0.60 | 0.21 | 0.31 | 231 |
| 5 | 0.84 | 0.97 | 0.90 | 995 |
| Microwave | 0.83 | 0.83 | 0.83 | 1226 |
| Macro_avg | 0.72 | 0.59 | 0.61 | 1226 |
| Weighted_avg | 0.86 | 0.87 | 0.87 | 1226 |

**Table 3**Novel Hybrid Bag-Boost Algorithm Accuracy

| Star rating | Prec | Rec | fl-score | Support |
|---|---|---|---|---|
| 1 | 0.88 | 0.84 | 0.82 | 231 |
| 5 | 0.92 | 0.90 | 0.94 | 995 |

ISSN PRINT 2319 1775 Online 2320 7876
*Research paper* **© 2012 IJFANS. All Rights Reserved**, **Volume 11, Iss 7 , 2022**

| | | | | |
|---|---|---|---|---|
| Microwave | 0.9457 | 0.9356 | 0.942 | 1226 |
| Macro_avg | 0.890 | 0.910 | 0.934 | 1226 |
| Weighted_avg | 0.950 | 0.950 | 0.950 | 1226 |

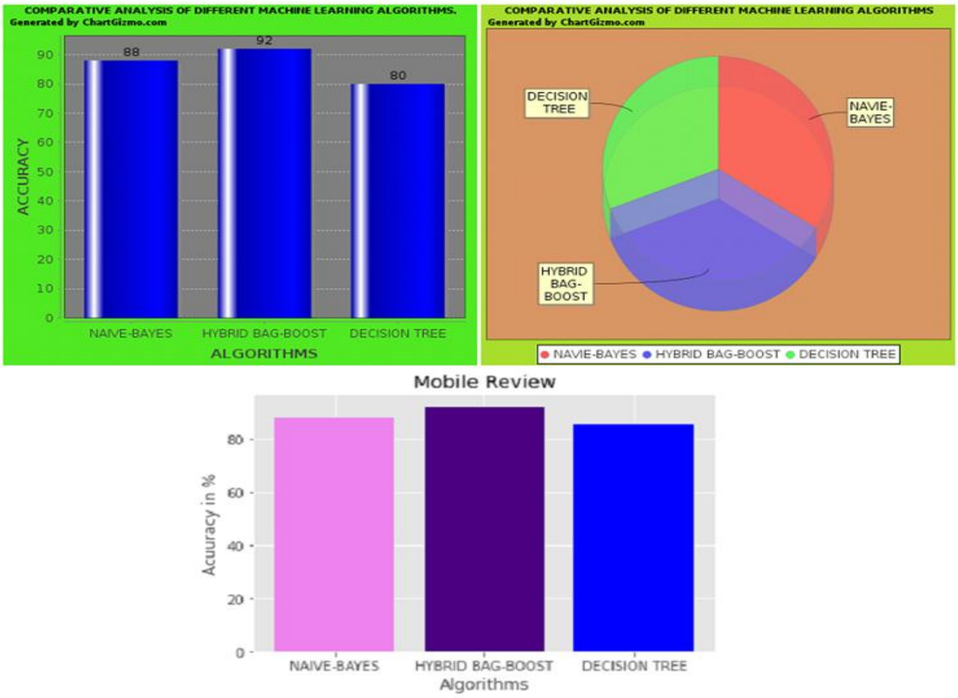


**Figure 5**. Accuracy Comparison

## 5. Conclusion and Future Work

Study of the feeling shows that reviews and feelings are pleasant. These algorithms can be used with computer study algorithms such as Naïve Bayes for the classification of facets and product empathy of polarity. The decision tree and hybrid Bag Boost are available. The experimental results show that the proposed technology has accomplished approximately 95 percent exactness compared to the other 2 algorithms, and its mission is very optimistic. We assume that by using larger databases with customer reviews accessible on the Internet, extension and simple application of this software can be strengthened. And the same algorithm is applied to GPU to increase the relative acceleration ratio

## References

1. Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., ... & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences, 97(1), 262-267.
2. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.
3. D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin. "Learning sentiment-specific word embedding for Twitter sentiment classification." In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '14), 2014.
4. Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent Data Analysis, 1(3), 131-156.

5. F. Fernández-Navarro, C. Hervás-Martínez, J. A. Gómez-Ruiz, J. M. Sánchez. "Analysis of Preprocessing vs. Hybrid Methods for Credit Scoring." In Proceedings of the International Symposium on Intelligent Data Analysis (IDA '09), 2009.
6. G. Qiu, B. Liu, J. Bu, C. Chen. "Opinion word expansion and target extraction through double propagation." Computational Linguistics, 37(1), 9-27, 2011.
7. Guyon, I., &Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157-1182.
8. H. Liu, H. Motoda (Eds.). "Feature Selection for Knowledge Discovery and Data Mining." Springer, 1998.
9. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1-2), 273-324.
10. L. Ma, S. Kwon, H. Cao, Y. Liu, J. Han. "Target-dependent Twitter sentiment classification." In Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data (SIGMOD '17), 2017.
11. Liu, H., &Motoda, H. (Eds.). (2008). Feature selection for knowledge discovery and data mining. Springer Science & Business Media.
12. M. Hu, B. Liu. "Mining and summarizing customer reviews." In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), 2004.
13. S. Kiritchenko, X. Zhu, C. Cherry, S. Mohammad. "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets." In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval '13), 2013.
14. S. Zhou, F. Ouyang, W. Li, G. Xu, S. Li. "Hybrid Classification Algorithm Based on SVM and Decision Tree for Chinese Text Sentiment Analysis." In Proceedings of the 5th International Conference on Computational Intelligence and Industrial Application (PACIIA '12), 2012.
15. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-2517.