Vol. 4, No. 1 | January - June 2020

Aspect Based Sentimental Analysis of Hotel Reviews: A Comparative Study

Sindhu Abro[1], Sarang Shaikh[1], Rizwan Ali[1], Sana Fatima[1], Hafiz Abid Mahmood Malik[2]

Abstract:

The increasing use of the internet enables users to share their opinion about what they like and dislike regarding products and services. For efficient decision making, there is a need to analyze these reviews. Sentiment analysis or opinion mining is commonly used to detect polarity (positive or negative) of reviews. But it does not show the aspect or orientation of the text. In this study, we have employed state-of-art approaches to perform three tasks on the SemEval dataset. Tasks A and B are related to predicting the aspect of the restaurant's reviews, whereas task C shows their polarity. Additionally, this study aims to compare the performance of two feature engineering techniques and five machine learning algorithms to evaluate their performance on a publicly available dataset named SemEval-2015 Task 12. The experimental results showed that the word2vec features when used with the support vector machine algorithm outperformed by giving 76%, 72%, and 79% off overall accuracies for Task A, Task B, and Task C respectively. Our comparative study holds practical significance and can be used as a baseline study in the domain of aspect-based sentiment analysis.

Keywords: *Aspects Based Sentiment Analysis; Sentiment Analysis; Text Classification; Natural Language Processing (NLP); Word2Vec; Machine Learning*

1. Introduction

In recent years, there is a rapid growth of content generated by users on the internet. The web enables users to share their reviews and experiences about services and products. Moreover, it is a growing trend that customers look already available reviews before purchasing any product or service [1]. Therefore, sellers and organizations need to analyze the reviews for effective decision making. The manual process to analyze the reviews is a labor-intensive and time-consuming task. Hence, techniques like sentiment analysis or opinion analysis are commonly used to extract information from reviews. The sentiment analysis, under the domain of natural language processing, used to determine the general opinion (e.g. positive or negative) of the group of individuals regardless of topic or entity (e.g. food, price, location, etc.) [2]. Therefore, it is recommended to use aspect-based sentiment analysis (i.e. ABSA). This concerned with the decomposition of two tasks namely aspect identification and sentiment analysis [3]. In the first task, the aspect of an entity is identified and in the second task, the polarity is estimated for each identified aspect. The sentiment analysis on the aspect level performs an in-depth analysis of reviews [4].

[1]Department of Computer Science, Sukkur IBA University, Pakistan
[2]Department of Computer Science, Arab Open University, Bahrain

For example, when we look at the reviews of the restaurant, ABSA not only returns the overall sentiment of the reviews but also returns for which entity the sentiment is talking about. Such as food, price, location, service, etc. Thus, the results generated from this technique gives a better understanding of what reviewers like and dislike regarding the topic [1]. Moreover, it may help customers to decide on the purchase of the products or using the services. Additionally, ASBA enables manufacturers to improve the quality of their products and services. Therefore, in this study, we have used ABSA to identify the aspects and their polarity of the reviews related to the restaurant.

The proposed solution employed the different feature engineering techniques and ML algorithms to classify restaurant reviews under different entities, attribute, and their polarities. Regardless of this extensive amount of work, it remains difficult to compare the performance of these approaches to classify hotel reviews text. To the best of our knowledge, the existing studies lack the comparative analysis of different feature engineering techniques and ML algorithms regarding the reviews related to restaurants. Therefore, this study contributes to solving this problem by comparing two feature engineering and five ML classifiers on the standard dataset provided by SemEval. This study will serve future researchers in the field of automatic ABSA.

This rest of the paper is organized as Section 2 highlights the related works. Section 3 discusses the methodology. Sections 4, and 5 explain the experimental setup, and results. Finally, Section 6 discusses the conclusion, and future work as well.

2. Related Works

Kiritchenko et al. [5] classified the reviews using the lexicon and linguistic features. Castellucci et al. [6] used a feature based on a bag of words that have been learned from external data. Hu and Liu [7] used an association rule-based system for aspect identification. Additionally, his book [8] highlights the four methods to extract aspects namely, frequent phrases, opinion, and target relations, supervised learning, and topic models. Jakob and Gurevych [9] employed the conditional random fields for aspect term.

Bhattacharyya [11] developed the system which uses dependency parsing rules for opinion extraction. Many researchers used a hybrid approach (i.e. NLP with statistical methods) to improve the performance of the system. In SemEval 2014, Kiritchenko et al. [5] used an entity tagging system named as in-house to extract outside and aspect terms. Toh and Wang [12] used the tagging approach with Wordnet and word clusters. Socher et al. [13] employed grammatical cues with deep learning. Carrascosa [14] study showed that an ensemble learning technique can also be applied in sentiment analysis. In the Aspect Category Polarity Detection task in SemEval 2014, Mohammad et al. [15] achieved the best performance by using different linguistic features, additionally, they also used publically available sentiment lexicon.

Broadly, ABSA methods can be divided into two categories, one that uses domain-independent solutions [16] and second is to use domain-specific knowledge [4] to improve the results. There is a common approach used by researchers that they treat aspect extraction and their polarity classification independently [17], but others also trained one model to solve the two problems [18].

3. Methodology

3.1. Overview

This section represents the overall research methodology that has been followed to perform the ASBA. Fig 1 shows the steps required to train the model. As shown here, our research methodology is composed of six key steps namely data collection, data preprocessing, feature engineering, data selection, classification model construction, and classification model evaluation. The details of each step are discussed in subsequent sections.

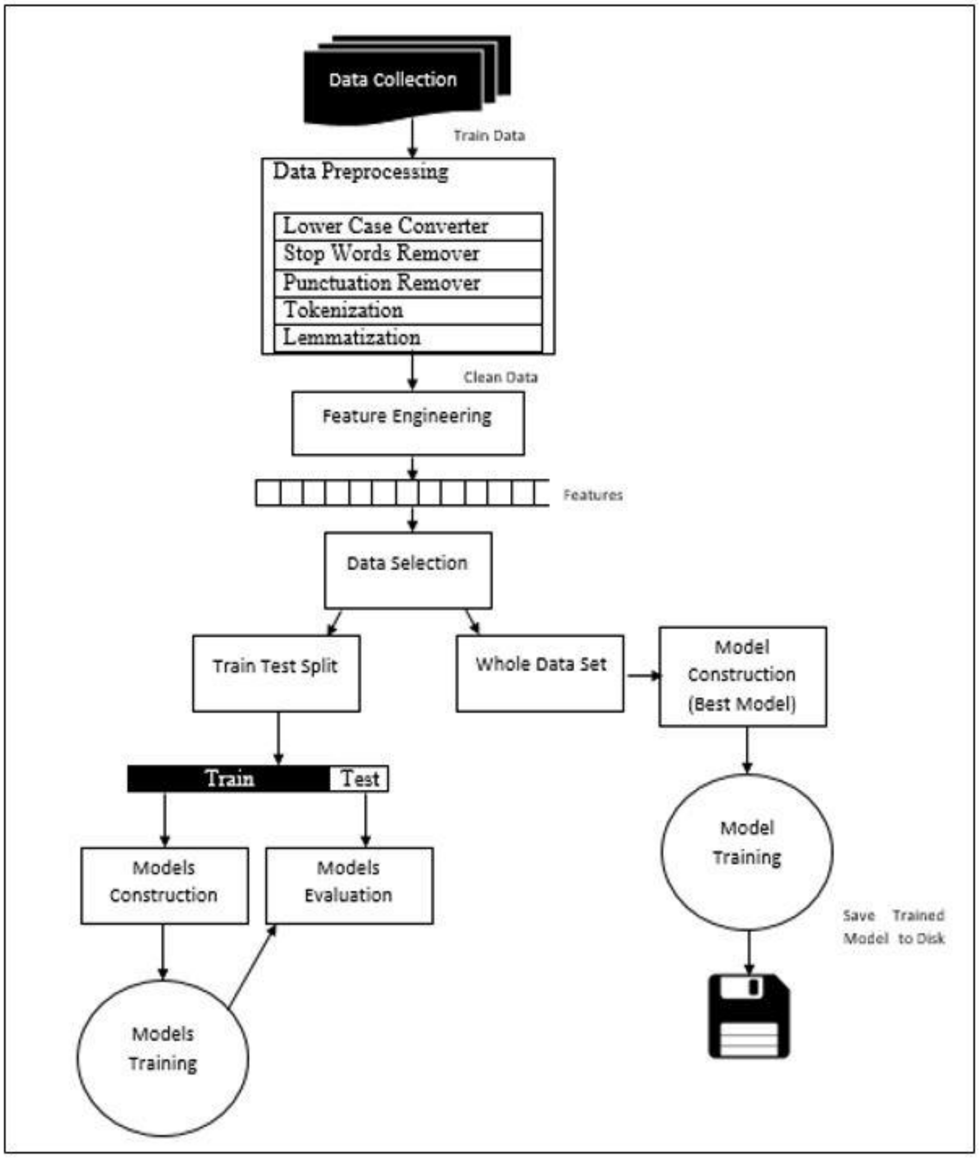

Fig 1. Overall Proposed Methodology

3.2. Data Collection

In this study, we have used publicly available data set from SemEval-2016 - Task 51[3]. This dataset contains reviews for laptops and restaurants. In this study, we will only focus on the reviews related to the restaurant. There is 3658 number of instances for the restaurant; 2799 for training and the remaining 859 for testing. In this dataset, the reviews can be categorized on the basis of aspects (i.e. category, entity, or attribute) and their polarities. By using aspect-based classification, the reviews can be labeled into six distinct classes of entity columns namely, food, restaurant, service, ambiance, drinks, and location. Additionally, the attribute can be labeled as general, quality, prices, style-options, and miscellaneous classes. However, their polarities can be positive, negative, or neutral. The distribution of reviews in training data based on entity, attribute, and polarity is shown in Fig 2, Fig 3 and Fig 4 respectively.

3.3. Text Preprocessing

Several studies show that there is a need to clean data for better classification results [19]. Therefore, we applied several preprocessing techniques to remove features from the data that are not informative. In this step, we have dropped the instances with blank values i.e. 292. Additionally, we have dropped the columns that are not required for text classification i.e. review-id, sentence-id, target, and category. After dropping the empty cells and selecting the required attributes, we converted the text (2507 remaining instance) into a lower case. Using regular expressions and pattern matching techniques, we removed white spaces, punctuation's and stop words. In addition, we have also applied tokenization and lemmatization on the preprocessed text. In tokenization, each sentence is converted into tokens or words, then words are converted to their root forms using WordNet lemmatizer e.g. posts to post

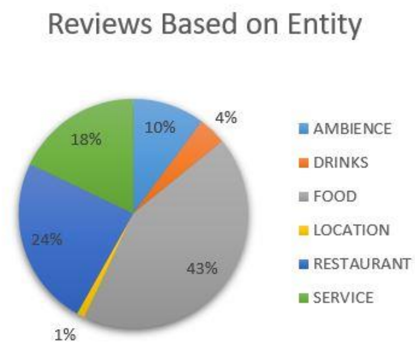

Fig. 2. Entity base distribution

Fig. 3. Attribute Base Distribution

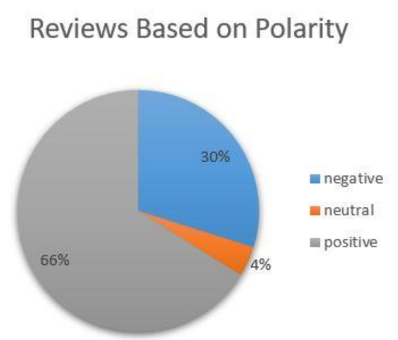

Fig. 4 Review base distribution

[3] The dataset is available at: http://alt.qcri.org/semeval2016/task5/index.php?id=data-and-tools

3.4. Feature Engineering

To learn classification rules, ML algorithms need numerical vectors because they cannot learn from raw data. Therefore, in classification one of the key steps is feature engineering. This step is used to extract the key features from raw text and represents the extracted features in numerical form. In this study, we have performed two types of features engineering techniques namely n-gram [20] with TFIDF [21], and Word2vec [22].

3.5. Data Selection

In this section, we have used two approaches to build the models named as train test split and whole data set. In the first approach, we have used the Pareto Principle. According to this principle, "80% of effects come from 20% of causes" [28]. This principle is also called an 80:20 ratios. In this study, we have split preprocessed data into a previously given ratio i.e. 80% for training and 20% for testing. Table 1, Table 2, and Table 3 show the class-wise distribution on the basis of an entity, attribute, and polarity as well as their train test splitting ratio. The training data is used to train the classification models for learning rules. However, the test data is used to evaluate the trained models.

Table 1: Approach I (Entity)

Class	Label	Total	Train	Test
Ambience	0	255	204	51
Drinks	1	99	79	20
Food	2	1076	861	215
Location	3	28	22	6
Restaurant	4	600	480	120
Service	5	499	359	90
Total		**2507**	**2005**	**502**

Table 2: Approach I (Attribute)

Class	Label	Total	Train	Test
General	0	1154	923	231
Miscellaneous	1	98	78	20
Prices	2	190	152	38
Quality	3	896	717	179
Style_options	4	169	135	34
Total		**2507**	**2005**	**502**

Table 3: Approach I (Polarity)

Class	Label	Total	Train	Test
Negative	0	749	599	150
Neutral	1	101	81	20
Positive	2	1657	1325	332
Total	3	**2507**	**2005**	**502**

In the second approach, we have used the whole data (i.e. 2507 number of instances) to train the model and for evaluation, different test data (i.e. 859 number of instances) were used. Table 4, Table 5 and Table6 show the distribution of data on the basis of entity, polarity, and attribute respectively.

Table 4: Approach II (Entity)

Class	Label	Total	Train	Test
Ambience	0	321	255	66
Drinks	1	137	99	38
Food	2	1467	1076	391
Location	3	41	28	13
Restaurant	4	796	600	196
Service	5	604	449	155

Table 5: Approach II (Attribute)

Class	**Label**	**Total**	**Train**	**Test**
General	0	1530	1154	376
Miscellaneous	1	131	98	33
Prices	2	238	190	48
Quality	3	1231	896	355
Style_options	4	236	169	67
Total		**3366**	**2507**	**859**

Table 6: Approach II (Polarity)

Class	**Label**	**Total**	**Train**	**Test**
Negative	0	953	749	204
Neutral	1	145	101	44
Positive	2	2268	1657	611
Total	3	**3366**	**2507**	**859**

3.6. Machine Learning Models

According to "no free lunch theorem" [23], any single classifier cannot outperform better on all types of datasets. Therefore, it is suggested to apply several classifiers on a master numerical vector to see which one achieves better results. Hence, we chose five different classifiers Naïve Bayes (NB) [24], Support Vector Machine (SVM) [25], Random Forest (RF) [26], Logistic Regression (LR) [27], and Ensemble in approach 1. Whereas in approach 2, we have chosen SVM and NB classifiers.

3.7. Classifier Evaluation

In this step, the constructed classifiers were used to predict the class of unlabeled text using test sets. The classifier performance is evaluated by calculating true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These four numbers constitute a confusion matrix as in Fig 5. To assess the performance of the constructed classifiers different performance metrics can be used like precision, recall, F measure, or accuracy. The details of given performance measures are given in [29]. However, in this study, we have used the most commonly used measure i.e. accuracy to evaluate the constructed classifiers. The details of this performance measure are given below.

	Predicted No	**Predicted Yes**
Actual No	TN	FP
Actual Yes	FN	TP

Fig. 5 Confusion Matrix

Accuracy

This evaluation matrix refers to the total number of instances that are correctly classified by the trained model. Refer to (1).

$$Accuracy = \frac{(TP + TN)}{TP + FP + TN + FN} \quad (1)$$

4. Experimental Setup

As mentioned in section A, the reviews can be categorized on the basis of aspects and their polarities. In this study, we have performed three tasks. In Task A, we have classified the reviews according to entity type (i.e. food, restaurant, service, ambiance, drinks, and location). In Task B, reviews are categorized according to attributes and labeled as general, quality, prices, style-options, and miscellaneous classes. Whereas in Task C, we have classified reviews according to their polarity like positive, negative, and neutral.

For all these tasks we have used two master feature representations namely n-gram (bigram) with TFIDF [21] and Word2Vec [22]. By using these master feature representations, we have followed two approaches to train the models. In approach 1, we used the train test split to train the five classifiers and evaluated their performance on test data. Whereas in approach 2, we used the whole dataset to train the models which have outperformed in approach 1 and evaluated their performance by using different test data.

5. Results

This section reports the results of all three tasks. Table 7, Table 8 and Table 9 show the accuracy using approach 1 (i.e. train test split) for Task A, B, and C, respectively. As shown in all three tables, the highest accuracy for Task A (0.71), Task B (0.69), and Task C (0.81) were obtained by SVM with word2vec.

Table 7: Approach I Results (Train-Test Split) - Task A

	Task A	
	Bigram (TFIDF)	**Word2Vec**
LR	0.59	0.70
NB	0.55	0.65
RF	0.58	0.57
SVM	**0.63**	**0.71**
Ensemble	0.61	0.67

Table 8: Approach I Results (Train-Test Split) - Task B

	Task B	
	Bigram (TFIDF)	**Word2Vec**
LR	0.60	0.67
NB	**0.61**	0.58
RF	0.54	0.56
SVM	0.58	**0.69**
Ensemble	0.57	0.66

In text-classification models, the SVM classifier performed exceptionally well among all 5 classifiers. If we evaluate the performance of all classifiers with respect to master feature representation, then we can see in Table 10 and Table 11 that for Task A and Task C the SVM classifiers with both master feature representations outperformed.

Whereas, from Table 8, Task B the NB using bigram with TFIDF (0.61) and SVM with word2vec (0.69) obtained the highest accuracy. Therefore, in approach 2, we have trained 6 models (3 Tasks x 2 master feature representations) on the whole dataset. The detail of all combinations is shown in Table 10.

Table 9: Approach I Results (Train-Test Split) - Task C

	Task C	
	Bigram (TFIDF)	**Word2Vec**
LR	0.73	0.80
NB	0.75	0.74
RF	0.73	0.74
SVM	**0.78**	**0.81**
Ensemble	0.75	0.80

Table 10: Model Selection for Approach II

Task	**Bigram (TFIDF)**	**Word2Vec**
A	NB	SVM
B	SVM	SVM
C	SVM	SVM
Ensemble	0.75	0.80

We have evaluated all these models on test data (i.e. 859). Table 11 shows the results of approach 2. It shows that word2vec obtained the best performance as compared to bigram features with TFIDF.

Table 11: Approach 2 Results for Task A, B & C

Task	**Bigram (TFIDF)**	**Word2Vec**
A	0.70	**0.76**
B	0.67	**0.72**
C	0.78	**0.79**

Furthermore, Fig 6, Fig 7 and Fig 8 show the confusion matrices of best-performing analyses. Fig 6 shows the confusion matrix of the SVM classifier using word2Vec for Task A. As shown here, out of 859 instances, 651 were correctly classified. Of these 651 instances, 47, 11, 341, 140, and 112 were classified as ambiance, drinks, food, restaurant, and service respectively. We can see that all 13 instances of location class were falsely classified.

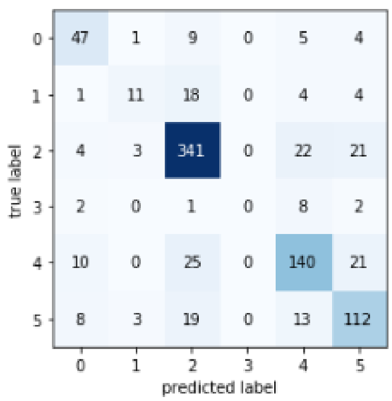

Fig. 6 Task A (Feature: Word2Vec, Classifier: SVM)

However, Fig 7 shows the confusion matrices of the SVM classifier using word2Vec features for Task B. As shown here, 621 instances out of 859 were correctly classified (i.e. General: 336 out of 376, Miscellaneous: 0 out of 33, Prices: 19 out of 48, Quality: 262 out of 335, and Style-options: 4 out of 67).

For Task C, the confusion matrix is shown in Fig 8. It shows that the SVM classifier with word2Vec features correctly classified 621 out of 859 instances, 124 as negative, and the remaining 557 as positive. As shown here, its performance was lowest in class 1 (i.e. neutral).

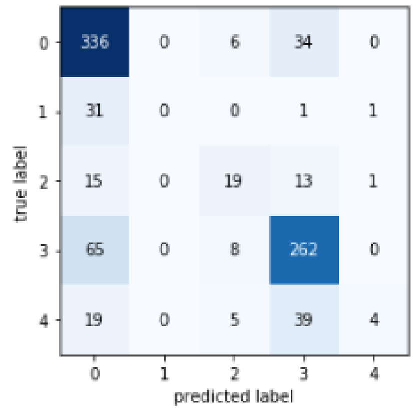

Fig. 7 Task B (Feature: Word2Vec, Classifier: SVM)

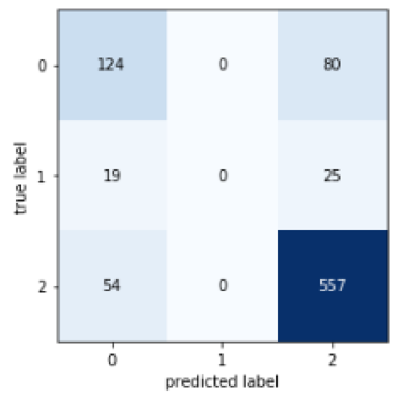

Fig. 8 Task C (Feature: Word2Vec, Classifier: SVM)

6. Conclusion

This study applied automated text classification techniques to classify the restaurant's reviews according to aspect and their polarities. Moreover, this study compared two feature engineering techniques and five ML algorithms to perform three tasks like a) classification of restaurant's reviews according to entity type, b) classification of restaurant's reviews according to their attribute and c) classification of restaurant's reviews according to their polarities. The experimental results showed that the word2vec showed better results for all tasks as compared to bigram represented through TFIDF feature engineering techniques. Moreover, the SVM algorithm showed better results as compared to NB, LR, RF, and Ensemble for all three tasks. The lowest results were observed in NB, RF, and LR for Task A, Task B, and Task C respectively. The outcomes from our study hold practical significance because these will be used as a baseline to compare future researches within different automatic text classification methods. In the future, the accuracy of the proposed system's classification can be increased by the following two strategies. First, the deep learning-based approaches will be explored and evaluated by comparing it with current state-of-the-art results. Secondly, more instances will be collected and used in the experiments for learning the classification rules efficiently.

REFERENCES

[1] Ekawati, D., and M.L. Khodra. Aspect-based sentiment analysis for Indonesian restaurant reviews. in 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA). 2017. IEEE.

[2] Wang, J., Encyclopedia of Data Warehousing and Mining, (4 Volumes). 2009: iGi Global.

[3] Schouten, K. and F. Frasincar, Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 2015. 28(3): p. 813-830.

[4] Thet, T.T., J.-C. Na, and C.S. Khoo, Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of information science, 2010. 36(6): p. 823-848.

[5] Kiritchenko, S., et al. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. in Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). 2014.

[6] Castellucci, G., et al. Unitor: Aspect based sentiment analysis with structured learning. in Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). 2014.

[7] Hu, M. and B. Liu. Mining opinion features in customer reviews. in AAAI. 2004.

[8] Liu, B., Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 2012. 5(1): p. 1-167.

[9] Jakob, N. and I. Gurevych. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. in Proceedings of the 2010 conference on empirical methods in natural language processing. 2010. Association for Computational Linguistics.

[10] Zhuang, L., F. Jing, and X.-Y. Zhu. Movie review mining and summarization. in Proceedings of the 15th ACM international conference on Information and knowledge management. 2006.

[11] Mukherjee, S. and P. Bhattacharyya. Feature specific sentiment analysis for product reviews. in International Conference on Intelligent Text Processing and Computational Linguistics. 2012. Springer.

[12] Toh, Z. and W. Wang. Dlirec: Aspect term extraction and term polarity classification system. in Association for Computational Linguistics and Dublin City University. 2014. Citeseer.

[13] Socher, R., et al. Recursive deep models for semantic compositionality over a sentiment treebank. in Proceedings of the 2013 conference on empirical methods in natural language processing. 2013.

[14] Carrascosa, R., An entry to kaggle's' sentiment analysis on movie reviews' competition. 2014.

[15] Mohammad, S.M., S. Kiritchenko, and X. Zhu, NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242, 2013.

[16] Lin, C. and Y. He. Joint sentiment/topic model for sentiment analysis. in Proceedings of the 18th ACM conference on Information and knowledge management. 2009.

[17] Brody, S. and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. in Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010. Association for Computational Linguistics.

[18] Jo, Y. and A.H. Oh. Aspect and sentiment unification model for online review analysis. in Proceedings of the fourth ACM international conference on Web search and data mining. 2011.

[19] Shaikh, S. and S.M. Doudpotta, Aspects Based Opinion Mining for Teacher and Course Evaluation. Sukkur IBA Journal of Computing and Mathematical Sciences, 2019. 3(1): p. 34-43.
[20] Cavnar, W.B. and J.M. Trenkle. N-gram-based text categorization. in Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. 1994. Citeseer.
[21] Ramos, J. Using tf-idf to determine word relevance in document queries. in Proceedings of the first instructional conference on machine learning. 2003. Piscataway, NJ.
[22] Mikolov, T., et al. Distributed representations of words and phrases and their compositionality. in Advances in neural information processing systems. 2013.
[23] Ho, Y.-C. and D.L. Pepyne, Simple explanation of the no-free-lunch theorem and its implications. Journal of optimization theory and applications, 2002. 115(3): p. 549-570.
[24] Lewis, D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. in European conference on machine learning. 1998. Springer.
[25] Joachims, T. Text categorization with support vector machines: Learning with many relevant features. in European conference on machine learning. 1998. Springer.
[26] Xu, B., et al., An Improved Random Forest Classifier for Text Categorization. JCP, 2012. 7(12): p. 2913-2920.
[27] Wenando, F.A., T.B. Adji, and I. Ardiyanto, Text classification to detect student level of understanding in prior knowledge activation process. Advanced Science Letters, 2017. 23(3): p. 2285-2287.
[28] Dunford, R., Su, Q., & Tamang, E. (2014). The pareto principle.
[29] Seliya, N., T.M. Khoshgoftaar, and J. Van Hulse. A study on the relationships of classifier performance metrics. in 2009 21st IEEE international conference on tools with artificial intelligence. 2009. IEEE.