Mining Public Opinion on Plastic Ban in India

Nandini Tomar[1], Ritesh Srivastava[2], Veena Mittal[3]

[1, 3]Manav Rachna International Institute of Research and Studies (MRIIRS) Haryana, India
[2]GCET, Greater Noida, India
[1]tomar.nandini@gmail.com, [2]ritesh21july@gmail.com,
[3]veena.mittal06@gmail.com

Abstract. Every product available in our environment has a shelf-life, but plastic is the only material that is non-degradable. The complex polymer presents in the plastic make it durable and non-degradable. As a result, it found in different forms on the earth for a long time. People have become used to of plastic made product in day to day life like carrying bags, disposable cutlery, food packaging and many more. Extensive quantities of plastic waste have accumulated in the nature and landfills and have posed an alarming hazard to the environment and now it reached a crisis point. Currently, India is ranked as the top four producers of plastic waste in the world. Though there is a law against the use of plastic in India but the usage of plastic made products is still high as the ban is not implemented completely and effectively. In this paper, we propose a framework for analyzing the opinion of Indian population on the plastic ban with the help of sentiment analysis technique on Twitter textual data. We train and test a machine learning classifier on different combination of datasets achieving 77.94% classification accuracy. The result obtained will help to understand how effective and successful polybags ban scheme will be when entirely implemented in India.

Keywords: Sentiment Analysis, Supervised Machine Learning, Natural Language Processing.

1 Introduction

Twitter, one of the most popular micro blogging websites among people all over the world, provides a podium to their registered users to express their feeling on the topic of interest effortlessly. People share their opinion on different social/political issues or product/services by writing a short message called as tweet having a maximum 140 characters. The informal structure and free format of tweets make it easy for a novice user to post a message with ease, and it results as an increasing number of users worldwide. All the tweets posted on Twitter wall carry user's emotions, feelings or feedback about the entity discussed which if can be mined appropriately can act as a crucial resource for different service industries and government organizations to track their status in the competitive market. Sentiment Analysis process is a part of Natural Language Processing which computationally identify and categorize the opinions expressed in a piece of text.

1.1 Plastic Pollution

Plastics also called as wonder material are cheap, lightweight, robust, durable, corrosion-resistant and long-lasting materials. They are now used in almost in every aspect of our day to day life including transport, telecommunications, clothing, footwear, medical equipment, food/product packaging, disposable cutlery, electric equipment, etc. As a result, the production and consumption of plastics has increased significantly over the last few years from around 0.8 million tons in 1955 to over 270 million tons today, and this rate of production is proliferating [1, 2].

Though, plastics are convenient in our day to day use and help to gain societal benefits and offer future technological and medical advances, however, excessive consumption and careless disposal of plastic material results as the high accretion of plastic waste in the natural environment and in the landfills, which now has posed an alarming hazard to the environment.

Use of plastic and accumulation of plastic waste harms our environment in many ways:

- Mutilation Sanitary System: The residues of plastic like polybags, pet bottles, food wrappers, etc. block pipes causing the fear of many waters contaminated diseases.
- Ocean Pollution: Throwing away the plastic wastes like plastic bags, bottles, etc. deposited in the ocean and pollute and disturb the natural eco-system of the ocean.
- Environmental Imbalance: Excessive use of plastic made goods, and careless disposal of plastic waste triggers the water, soil and air pollutions disturbing various ecosystem of the Earth.

Countries around the world are articulating ways to get rid of plastic waste disposal and plastic pollution. Some of the countries listed in the below Table 1, where the plastic use law is in place and strictly followed [2].

Table 1. List of Country and their Law for The use of Plastic

Country	[1] Law
France	‘Plastic Ban’ law is passed in 2016 which states all plastic plates, cups, and utensils will be banned by 2020. France is the first country to ban all the plastic made products used in daily life.
Rwanda	Rwanda implemented a complete ban on plastic bags and this country is plastic bag free since 2008.
Sweden	This country is known as one of the world’s best recycling nations which follow the policy of ‘No Plastic Ban, Instead More Plastic Recycling.’
Ireland	In this country a tax scheme is implemented in the use of plastic bags.
China	In 2008 China passes a law to deal with plastic crisis. In China it is illegal to give plastic bag to the customers for free.

1.2 The ban on Plastic Bags in India

In August 2017 in India, the manufacturer, stock, sale and use of plastic bags have been banned in 17 states and Union Territories including Bengaluru, Maharashtra, Delhi, Punjab, Rajasthan, Goa, West Bengal, etc. However, use of plastic carry bags has been partially banned in 11 states which are pilgrimage centers, tourist and historical places. The name of states where the use of plastic carry bags is partially banned are Andhra Pradesh, Arunachal Pradesh, Assam, Goa, Gujarat, Karnataka, Odisha, Tamil Nadu, West Bengal, Uttar Pradesh and Uttarakhand [2].

National Green Tribunal in Delhi NCR announced a ban on disposable plastic items like food packaging, cutlery, polybags and other plastic items. The ban on plastic items came into effect on January 1, 2017. The ban on plastic affects the whole National Capital Territory (NCT) area of Delhi [3].

On March 23, 2018, in India's second-populous state Maharashtra, the government banned the manufacture, usage, sale, transport, distribution, and storage of plastic bags with or without handle, and disposable products made out of plastic and thermocol [4] [5].

Due to the massive dependence on plastic and lack of alternatives to the banned products its effective enforcement is an issue. In this work, we try to find out people's reaction on the plastic ban by analyzing tweets posted by Twitter users so that government can get a clear picture of its effective implementation and appropriate corrective actions can be taken on time.

1.3 Sentiment Analysis

Sentiment Analysis is the area of Natural Language Processing (NLP) and text mining which help to analyze the opinion of a person (called an opinion holder) towards an entity from given written text.

1.4 Techniques of SA

Numerous methods have been devised so far to perform sentiment analysis process [6-12]. Each method has their applicability and strength, so the selection of suitable technique depends upon the type of textual data used in the analysis process and the area where analysis has to be performed. Techniques used to perform SA process are categorized into following two major categories [13]:

1. Machine Learning (ML) Approach
2. Lexicon Based Approach

Rest of this work is confined in the following sections; section 2 gives the brief overview of the work done in the area of sentiment analysis by various researchers. Next, in section 3, all the tools and techniques used to execute this work along with approach opted for data collection is specified. Section 4 graphically shows and defines the systematic flow and implementation of this work. Section 5, gives a detailed discussion of the results obtained by executing this work. Finally, in section 6, we conclude this work along with future perspectives defining the scope of improvement.

2 Literature Review

Analyzing user's opinion from written text with improved accuracy is a quite challenging task. Researchers are unceasingly working on finding new methods and strategies to improve time, speed and efficiency of the sentiment analysis process. Some of the significant work done in this area by various researchers is given in subsequent part of this section.

Tweets posted on Twitter are used as a corpus of positive and negative opinion words to perform sentiment analysis [14]. A method is proposed which collect an arbitrary large corpus of positive and negative opinion words automatically. It also collects a corpus of objective texts. With this collected corpora sentiment analysis is performed. Domain-specific features extraction is one of the important tasks in sentiment analysis process. The work of [15] performed SA by using the strategy of selecting most frequent noun as a feature from collected blog or document with the help of association mining. They use nearby opinion words to find out infrequent features.

A distance-based method [16] is suggested to identify compact features of a product, e.g., camera type, battery life, functions, etc. are the compact features of a mobile phone. In [17], the author defines the existing approaches for opinion-oriented information retrieval system including blogs and micro blogging sites. A corpus of opinion words is created from web blogs [18]. Emoticons are also used to indicate the user's mood. In this work SVM and CRF, classifiers are trained using collected corpora of opinion words to perform sentiment analysis at the document level. A detailed discussion on sentiment analysis and its tools and techniques are described in [13].

To make analysis more precise and accurate careful feature selection processes are also discussed. Use of emoticons smiley and sad symbolized as ":-)" and ": - (" is suggested to form a training data set for training the classification model [19]. Emoticons are collected from Usenet newsgroups. Dataset to train the classifier contains two samples one is text with positive opinion words and smiley emoticons, and another one is text with negative opinion words and sad emoticons.

OPINE [20] uses web-based search for calculating Pointwise Mutual Information (PMI) score between the phrase and a part of differentiator related to product class using equation (1), where f is fact and d is a differentiator.

$$PMI(f,d) = \frac{(hits(f \wedge d))}{(hits(d)hits(f)} \qquad (1)$$

The semantic orientation of adjectives is one of the important factors to consider carrying out sentiment analysis process. Remarkable work is done in [21], where they propose an algorithm to cluster all positive or negative objectives by their calculated semantic orientation. This algorithm is based on conjunction hypothesis and is designed for isolated adjectives.

3 Experimental Setup and Data Collection

3.1 Experimental Setup

Execution of this work is done on a machine with Intel Core i3, 2.4GHz, 3 MB L3 cache processor with 4GB of RAM. A script written in R language is executed to scrap out the tweets in which people express their feelings for a plastic ban on Twitter. We also utilize the online open resource SentiWordNet to perform subjectivity test on collected tweets. Finally, sentiment analysis on the collected dataset is performed using a powerful machine learning tool WEKA [24] . We can directly apply readily available machine learning algorithm on collected data or can use java code to initiate the execution.

3.2 Tweets Extraction

A key search is performed by entering few plastic ban related keywords mentioned in the harvest list given below. A script, written in 'R' language, is executed to scrap dawn the tweets from Twitter open API. Around 31,634 tweets posted on the Twitter platform from 15th June 2018 to 22nd June 2018 (i.e., of one week) were extracted.

Harvest List= {Plastic Ban, Polythene Ban, Plastic Pollution}

3.3 Subjectivity Analysis

Tweets extracted from Twitter open API may contain some tweets with no sentiments or opinion within called as objective tweets. Subjectivity analysis filters out statements that include no opinion. To perform this analysis an openly available lexicon resource, SentiWordNet [22] is used which contain positive/negative sentiment word. Each tweet is labeled manually according to the class they belong to, i.e., either positive or negative.

After subjectivity test and manual annotation on all the collected tweets (i.e., 31,634 Tweets), we get 12,729 tweets with positive polarity and 11,943 tweets with negative polarity rest 6962 tweets was objective and so are discarded. To balance the collected, labeled dataset, we have selected 11,000 positive tweets and 11,000 negative tweets.

3.4 Training Data Preparation

The next step immediately after data extraction is training data preparation. In order to make our training data rich in opinion words and vast here we encapsulate in domain dataset with out of domain dataset available online. The training dataset is constructed from the following two sources:

1. Manually Annotated Dataset.
2. International Movie Database (IMDb) [23].

Manually Annotated Data. The manually annotated dataset is the collection of labeled positive and negative tweets which are specific to the domain (in this work Plastic Ban). After applying subjectivity analysis and manual annotation on all collected tweets, we obtain 12,729 positive and 11,943 negative tweets out of which we have selected 11,000 positive polarity tweets and 11,000 negative polarity tweets to possess the uniform tweets selection throughout this work. The statistics of our dataset is given in Table 2.

IMDb Reviews. IMDb reviews are labeled collection of 1000 positive and 1000 negative movie reviews which are highly polar, that is all positive reviews comprise strictly positive opinion words and in the same way negative reviews contain strictly negative opinion words. Encapsulating these reviews with our collected dataset will enrich the semantic knowledge base which will help in better polarity classification.

Table 2. Training Dataset Statistics

Sources	Positive Instances	Negative Instances
IMDb Reviews	1000	1000
Manually Annotated Dataset	11,000	11,000
Total Instances	12,000	12,000

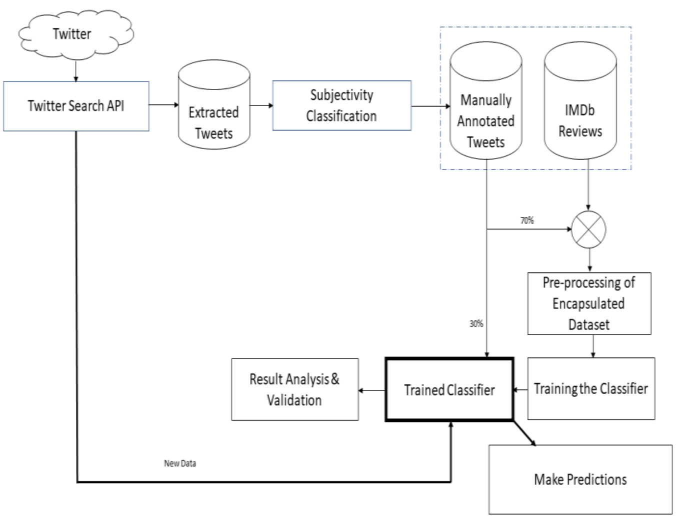

Fig. 1. Proposed System Architecture

4 PROPOSED APPROACH

The approach chosen to execute this work is given in below-mentioned Fig 1. Following steps need to follow to perform sentiment analysis on collected data:

a) Data Collection
b) Preprocessing
c) Training the Classifier
d) Polarity Classification

4.1 Data Collection

Tweets in which users have posted their views about plastic ban are collected from Twitter REST API from 15th June 2018 to 22nd June 2018, i.e. of one week. A connection needs to be established between the user program and Twitter open API before the data extraction process. To build this connection first, we need to create a Twitter account and then get credentials (they are API Key, API Secret, Access Token and Access Secret) on the Twitter developer site.

A script written in R language is executed to get the authentication and related tweets from Twitter open API.

4.2 Preprocessing

All the tweets collected from Twitter may contain some undesirable data including Twitter specific terms like '@' tags, '#' tags, URLs, symbols etc. and some other terms like punctuations, stop words, abbreviations, etc. called as noise which needs to be removed from the collected dataset before passing to the. Following pre-processing steps are followed to make the dataset clean and smooth:

- Tokenization
- Normalization
- Case Sensitivity
- Stemming

4.3 Training Machine Learning Classifier

In this work, we opted for supervised machine learning classifier, Support Vector Machine (SVM).

Based upon the combination of training and testing dataset used to train the SVM classifier following three models are introduced:

Model-1. Model-1 is trained and tested over highly polar IMDb reviews using k fold cross-validation process for k= 10. Cross-validation is a resampling process used to evaluate machine learning models on a limited data sample. In k-fold cross-validation testing process starts by partitioning the given dataset into k number of equal size sub-datasets, after that a single sub-set is selected as the validation dataset, and rest (k-1) sub-datasets are used for training the classifier. This process is repeated k times, with each k subsamples used exactly once as the validation dataset. Final result is produced by averaging the k calculated results.

It's a standard approach to follow to verify the correctness of the out-of-domain data set used (IMDb reviews in this work) in combination with the in-domain dataset.

Model-2. Model-2 is trained with IMDb reviews and tested over manually annotated tweets using batch setting. This model is proposed to demonstrate that if we use a classifier trained solely with different domain dataset may misclassify few statements due to the absence of some domain-specific opinion words in classifiers knowledge base. Results shown in the subsequent section will validate this statement.

Model-3. Model-3 is our core model which is trained and tested over summed up dataset (i.e., Domain Specific Tweets + IMDb reviews) using 10-fold cross-validation process. We believe that the accuracy of the classifier can be improved by adding more opinion words in its knowledge base and here, we achieve it by encapsulating data from two sources (IMDb reviews and Domain Specific Twitter Data) to train the machine learning classifier.

Table 3. Confusion matrix obtained for Model-1

Predicted→ Actual↓	**Negative Tweets**	**Positive Tweets**	**Accuracy Obtained**
Negative Tweets	782(TN)	218(FP)	78.2%
Positive Tweets	203 (FN)	797 (TP)	79.7%
Total Tweets	985	1015	78.95%

Where, TP=True Positive, TN=True Negative, FP=False Positive, and FN=False Negative Instances

Table 4. Confusion matrix obtained for Model-2

Predicted→ Actual↓	**Negative Tweets**	**Positive Tweets**	**Accuracy Obtained**
Negative Tweets	6784(TN)	4216(FP)	61.67%
Positive Tweets	4205(FN)	6795(TP)	61.77%
Total Tweets	11,989	12,011	61.72%

Table 5. Confusion matrix obtained for Model-3

Predicted→ Actual↓	**Negative Tweets**	**Positive Tweets**	**Accuracy Obtained**
Negative Tweets	18,689(TN)	5311(FP)	77.87%
Positive Tweets	5264(FN)	18736(TP)	78.06%
Total Tweets	23953	24047	77.97%

5 Results Analysis and Discussions

This section comprises a detailed discussion on the performance obtained with each proposed model to make it easy to compare and visualize the difference between them. This work of finding public opinion on plastic ban is carried out by using our core model-3 which is trained and tested with the dataset build by encapsulating data from two different sources. Classifier's result is received in the form of a matrix called as a confusion matrix which determines the correctness and accuracy of the SVM classification model on the labelled validation dataset. It is used for Classification problem where the output can have two or more classes. The result is analyzed regarding precision value, recall value, F-curve and overall accuracy attained.

5.1 Results Obtained with Model-1

Model-1 is trained and tested over IMDb dataset which contains 2000 highly polar reviews out of which 1000 are positive and rest 1000 are negative reviews. Complete execution is performed in WEKA using k-fold cross-validation process where k=10. Confusion matrix obtained with this model is shown in above-mentioned Table 3. With model-1 78.95% classification accuracy is attained.
We propose this model to verify the correctness of the out-of-domain data set used (IMDb reviews in this work) in combination with the in-domain dataset (collected tweets related to plastic ban).

5.2 Results Obtained with Model-2:

Model-2 is trained with IMDb and validated over annotated domain specific tweets (containing 22,000 labeled tweets) using batch setting. With this model, we attain 61.72% classification accuracy. Accuracy with this model dropped because classifier is trained and validated over two different domain datasets which result as misclassification of few instances. Table 4 shows the predicted instances with this model.

5.3 Results Obtained with Model-3:

In model-3, we enrich our training dataset by encapsulating data from two different sourced, i.e., one is manually annotated tweets (22,000 Tweets) scrapped from Twitter API, and another one is IMDb reviews comprising 2000 highly polar reviews. This change in strategy results in improved classification accuracy. Results obtained with this model show the significant improvement in the performance of the classifier. Table 5 shows all the predicted values and accuracy obtained corresponding to the given dataset.

5.4 Performance Obtained with Each Suggested Model

Performance of the classifier is analyzed regarding precision value, recall value, F-curve and overall accuracy attained. From Table 6, we can observe the performance corresponding to each model, and it is evidently seen that our core model-3 outperform when trained and tested with the encapsulated dataset.

Table 6. Performance Measures Obtained with Model-1, Model-2 & Model-3

Model	**Model-1**	**Model-2**	**Model-3**
Model Description	***Trained & validated over IMDb reviews***	***Trained with IMDb reviews & validated over the domain-specific dataset***	***Trained & tested over encapsulated dataset (IMDb reviews+domain specific labeled tweets)***
Precision	50.47%	50.04%	50.06%
Recall	79.7%	61.77%	78.07%
F-Measure	61.80%	55.29%	61.00%
Accuracy	78.95%	61.72%	77.97%

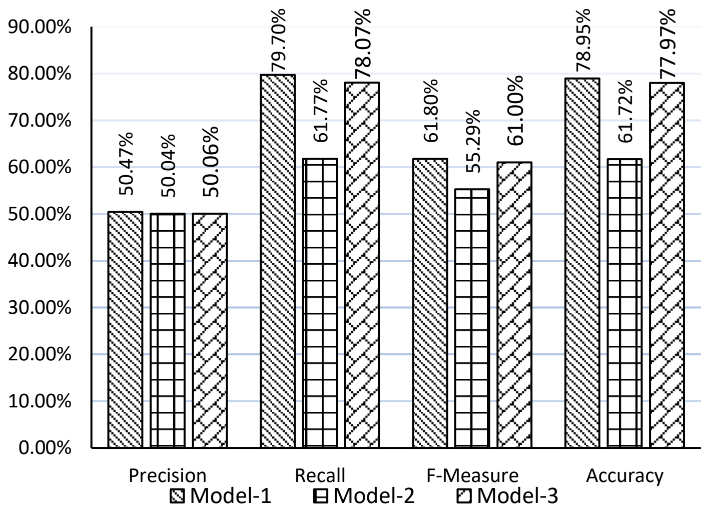

Fig. 2. Graphical Representation of Performance Measures Obtained with Each Model

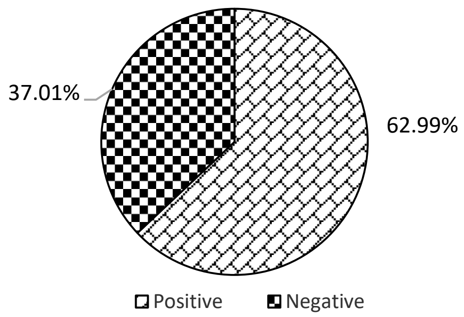

Fig. 3. Polarity classification percentage

5.5 Performing SA on Domain-Specific Subjective Tweets

Finally, we use our core Model-3 to perform classification on time, and domain-specific dataset (i.e., total 1,50,374 subjective Tweets) in which issue related to plastic ban was discussed. The classifier classifies 94735 tweets as positive polarity and remaining 55639 tweets as the negative polarity. The pie chart given in Fig 3 shows the percentage share of classified positive and negative tweets. Our observation on collected Twitter data shows that around 62.99% people carry positive opinion about Plastic Ban in India.

6 Conclusions and Future Work

We have successfully built a model with 77.94% classification accuracy and performed sentiment analysis on time specific and domain-specific Twitter data posted on the Twitter wall. Our observation shows that 62.99% people are quite positive towards the law in polybags ban in India and rest 37% people shows a negative opinion about it. Furthermore, we can explore other more accurate classification methods for getting more accurate results.

In this work, only two class classification is performed, i.e., positive class and negative class. However, we can also extend this work by defining more classes according to the behavior people show about this law like neutral, anger, anticipation, etc. Also, we can improve the accuracy and efficiency of the classifier by applying more pre-processing steps.

7 References

1. *Plastic Bag*. Available: https://en.wikipedia.org/wiki/Plastic_bag
2. *Plastic Ban: What India Can Learn From Other Countries*. Available: https://swachhindia.ndtv.com/plastic-ban-india-can-learn-countries-6161/
3. "India just banned all forms of disposable plastic in its capital, Independent", Available: https://www.independent.co.uk/news/world/asia/india-delhi-bans-disposable-plastic-single-use-a7545541.html
4. "*Plastic ban in Maharashtra: What is allowed, what is banned*", Available: https://indianexpress.com/article/india/plastic-ban-in-maharashtra-mumbai-from-june-23-what-is-allowed-what-is-banned-all-you-need-to-know-5228307
5. "Why have laws to completely ban plastic bags failed in India?." Available: https://scroll.in/article/872612/why-have-laws-to-completely-ban-plastic-bags-failed-in-india
6. Srivastava, R. and Bhatia, M.P.S., 2013, August. Quantifying modified opinion strength: A fuzzy inference system for sentiment analysis. In 2013 *International Conference on Advances in Computing, Communications and Informatics (ICACCI)* pp. 1512-1519. IEEE.
7. R. Srivastava and M. Bhatia, "Ensemble methods for sentiment analysis of on-line micro-texts," in *international conference on Recent Advances and Innovations in Engineering (ICRAIE),* 2016, pp. 1-6: IEEE.

8. R. Srivastava and M. Bhatia, "Offline vs. Online Sentiment Analysis: Issues With Sentiment Analysis of Online Micro-Texts," *International Journal of Information Retrieval Research (IJIRR),* vol. 7, no. 4, pp. 1-18, 2017.
9. R. Srivastava and M. Bhatia, "Real-Time Unspecified Major Sub-Events Detection in the Twitter Data Stream That Cause the Change in the Sentiment Score of the Targeted Event," *International Journal of Information Technology and Web Engineering (IJITWE),* vol. 12, no. 4, pp. 1-21, 2017.
10. R. Srivastava and M. Bhatia, "Challenges with Sentiment Analysis of On-line Micro-texts," *International Journal of Intelligent Systems and Applications,* vol. 9, no. 7, p. 31, 2017.
11. R. Srivastava, M. Bhatia, H. K. Srivastava, and C. Sahu, "Exploiting grammatical dependencies for fine-grained opinion mining," in *international conference on Computer and communication technology (iccct),* 2010, pp. 768-775: IEEE.
12. R. Srivastava, H. Kumar, M. Bhatia, and S. Jain, "Analyzing Delhi assembly election 2015 using textual content of social network," in *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015*, 2015, pp. 78-85: ACM.
13. V. B. Vaghela and B. M. Jadav, "Analysis of various sentiment classification techniques," *Analysis,* vol. 140, no. 3, pp. 22-27, 2016.
14. A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *LREc*, 2010, vol. 10, no. 2010, pp. 1320-1326.
15. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168-177: ACM.
16. M. Hu and B. Liu, "Mining opinion features in customer reviews," in *AAAI*, 2004, vol. 4, no. 4, pp. 755-760.
17. B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.
18. C. Yang, K. H.-Y. Lin, and H.-H. Chen, "Emotion classification using web blog corpora," in *Web Intelligence, IEEE/WIC/ACM International Conference on*, 2007, pp. 275-278: IEEE.
19. J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proceedings of the ACL student research workshop*, 2005, pp. 43-48: Association for Computational Linguistics.
20. A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural language processing and text mining*: Springer, 2007, pp. 9-28.
21. V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997, pp. 174-181: Association for Computational Linguistics.
22. A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC,* 2006, vol. 6, pp. 417-422: Citeseer.
23. B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 2004, p. 271: Association for Computational Linguistics.
24. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10-18, 2009.