EasyChair Preprint

№ 5448

Social Media Analysis for Sentiment Classification Using Gradient Boosting Machines

Pradeep Kumar and Abdul Wahid

May 4, 2021

Social Media Analysis for Sentiment Classification Using Gradient Boosting Machines

Pradeep Kumar[1] and Abdul Wahid[2]

[1,2] Department of Computer Science & Information Technology, Maulana Azad National Urdu University, Hyderabad, India
drpkumar1402@gmail.com

Abstract The Sentiment analysis deals with the emotions of users on social media discussions and reviews. Gradient Boosting Machine has shown improved results significantly on many standard classification benchmarks. This paper illustrates the process of text classification for social media to perform sentiment analysis using machine learning (ML) techniques: Gradient boosting machines (GBM), AdaBoost, and eXtreme GBM (XGBM) for analyzing online reviews. The classifiers are trained on a benchmark dataset and performance is assessed in terms of accuracy. A set of systematic experiments are conducted on a social media dataset extracted from the Kaggle. Experimental results reveal that XGBM outperforms in terms of both training and testing accuracy. Sentiment analysis would provide substantial clues about services and product reviews leading to better marketing strategies for branding the products and maximize the level of customer satisfaction and helping in policy-making decisions.

Keywords: Text mining, social media, sentiment analysis, feature selection, machine learning techniques

1 Introduction

In recent years social media has emerged as a personal communication platform to express individual opinions about specific services and products including political, social, legal, and common events of interest among the users. Handling a large number of rapidly growing digital documents has become a tedious task for the automatic categorization of text [1][2][3]. Prominently text mining (TM), machine learning (ML), and natural language processing (NLP) techniques have been applied extensively to extract useful information from vast unformatted documents [4][5]. A massive amount of these personnel views and opinions are flooded daily on the popular social media including Twitter, Facebook, YouTube, and LinkedIn [6][7].

Text classification problem deals to handle the huge number of unstructured documents such as web pages, emails, social media forum, and postings of other electronic documents. Polysemy and synonymy are other problems with text mining. Polysemy refers to the words that can have multiple meanings. Synonymy refers to the different words having the same or similar meaning. Sentiment Analysis (SA) known as opinion mining also or polarity mining deals with computational linguistics, NLP, and other text analytics methods. SA automatically extracts user sentiments from text sources such as words or phrases or complete documents. SA is found as one of the potential research areas in NLP and other diverse fields of data mining [8-10].

The explosive growth of online social media has received substantial attention in recent years to address the problem of automatic sentiment analysis. Sentiment classification attempts to measure the polarity of a given text document more precisely predicting whether the given reviews and opinions on social media are positive, negative, or neutral.

Unstructured text generated from popular social media platforms such as Facebook, YouTube, and Twitter able to provide substantial clues about services and product reviews leading to better marketing strategies for branding the products and maximize the level of customer satisfaction [11-13]. Knowledge extracted from social media can be extremely helpful because a large number of opinions expressed about a specific topic may lead to vital information related to business policy. The impact of all such opinions make them easily understandable by the majority of readers and subsequently set up a trend gradually for recommending some products or services [8][13].

Major applications of SA include predicting stock market trends, framing policies to promote the product and services, recommender systems, and managing crises. Basic methods for sentiment analysis include text statistics such as word counts, frequency, and review categories (positive and negative reviews for specific comments or remarks in the given text documents). Therefore, systematic SA of social media can help the stakeholders by providing the extraction of insightful conclusions about the public opinions and variety of topics. Moreover, it poses serious technical challenges also due to noisy, sparse, and multilingual content posted by the users on social media [16-19]. Bootstrap Aggregating (Bagging) and Boosting are popular ensemble techniques. Bagging combines the results of multiple base models to generate improved results. Bootstrapping is a sampling technique with a replacement where subsets of samples are generated from the original dataset. Bagging applies these subsets known as bags to provide the distribution of a complete set. The size of subsets may vary from the original set. (Ross 1993; Ho 1995; Breiman 2001) [20]. Gradient boosting-based classifiers apply different weak learning models like decision trees (DTs) to build up a strong prediction model. Gradient boosting models are capable of handling complex unstructured social media data quite effectively. Among various machine learning methods, Gradient Boosting Machines (GBM) has shown state-of-the-art results on many standard classification benchmarks.

The primary objective of Gradient Boosting machines is to minimize the loss function very similar to gradient descent algorithms in a neural network. In an iterative process, new weak learners are added to the model and the weights of the past learners are cemented in place, leaving unchanged samples for the new layers. GBM can be applied to multi-class classification problems and regression problems also.

The rest of this article is structured as follows: Section 2 presents the related work of sentiment analysis and classification; Section 3 explains the feature selection and pre-processing of unstructured text; Section 4 describes the research methodology applied; Section 5 presents the experimental results, and conclusion followed by future directions.

2 Related Work

Many researchers have applied supervised and unsupervised algorithms to perform text classification and sentiment analysis of social media [4][5][6]. The prior research has focused on sentiment or content classification tasks. Ordinarily, natural text cannot be utilized properly in the analysis without pre-processing. Several techniques of pre-processing have been explored in the literature. Bag-of-words (BOW) is one of the traditional approaches for sentiment classification to analyze the text features with the help of supervised learning algorithms [7-11]. Using machine learning methods, the words can be filtered from the BOW vector. For example, the appearance of each word in the text can be represented as the feature from such vectors. Other common methods include n-gram, POS tags, and negation-tags are applied to optimize the features. The N-gram and negation-tags are found effective to improvise the precision of the classifier wherein POS tags can be applied for multiple meanings of one word.

Bag of Words is a predominantly used method for sentiment analysis and classification using machine learning methods. Pang and Lee extracted features from online reviews about a movie and analyze the user's sentiments with the help of words bag and feature vectors [15]. Demitery et al., (2011) applied high order n-grams for sentiment classification of a text document. They devised an embedding mechanism of n-grams to deal with the curse of dimensionality.

Huang Zu et al. (2015) investigated the words bag method using support vector machines (SVM) and Naïve Bayes with the help of syntactic features along with part of speech (POS) tags. They constructed word dependencies and syntax trees to establish the grammatical and logistic relationship between words in sentences.

Almatarneh and Gamallo (2019) applied supervised learning techniques for class labeled training data based on automatic text classification. A predictive model is designed based on the previous text documents collected from social media and other available repositories. The effective outcomes of these machine learning models depend on several factors such as feature selection, parameter tuning, training of the model, and capability of learning in a dynamic context.

The sentence syntax tree (SST) is another alternative approach applied for designing sentiment classifiers. The sentences are parsed to build a syntax tree to establish the relationship among these words. The sentiment classification model can be built with the help of words polarity, POS features, and syntax. Dave et al. employed machine learning techniques for sentiment classification with the help of top words selected according to their generated points [11]. Mullen and Nigel Collier [4] applied SVM to analyze sentiment classification from the orientation of words point of view like topic-oriented and artist-oriented information.

Pak and Paroubek proposed a sentiment classification model using bag-of-words for Twitter data as an application of sentiment analysis in social media. Several researchers and practitioners have applied syntax trees also to establish an internal relationship between the words. Adwait Ratnaparkhi [6] used maximum entropy models for parsing the syntax trees to find the patterns behind the syntax tree. Zhan, Li and Zhu [8] improved the accuracy of parsing the syntax tree using rules and patterns. Nakagawa, Inui, and Kurohashi [9] studied the impact of sentiment dependency on words using CRF

with hidden variables. Duric and Song [10] applied the HMM-based model for the analysis of a sentence's content and sentences. Other approaches for sentiment classification includes lexicon structures for words and their sentiments generating classification rule.

2.1 Social Media Analysis

Social media analytics can be treated as a multiphase activity including capturing, understanding, and presentation. Pre-processing techniques are applied to extract prominent features required for accurate classification and prediction. Preprocessing techniques such as feature extraction, selection, grouping, and evaluating process is applied to classify the text document. Subsequently, machine learning algorithms are applied to train the classifier and then test them to categorize whether the sentiments are positive or negative [12][13]. Reduction of text dimensionality can be achieved through filtering, lemmatization, and stemming methods. Stop word filtering is the standard filtering approach to remove words from the dictionary. The main purpose of applying word filtering is to remove words that carry little information such as conjunctions, articles, prepositions, adjectives, etc bearing no particular statistical relevance [12].

Text documents usually contain various undesirable characters like punctuation marks, special characters, stop words, digits, etc which may not help to classify the text. Therefore, it has to be preprocessed before applying it to the classifier for effective outcomes. Text cleaning is the first step in any text mining problem where irrelevant details are removed from the document which may not contribute to the vital information of greater interest.

Bag of Words (BoW) model can be applied to extract the features by considering each word as a feature. Each comment on a specific product or service is treated as a bag of words. BoW creates a dictionary of all the words and their frequency in the text document or dataset used for sentiment classification [17]. Words occurrence is the number of times the word occurs in the entire corpus. BoW model ignores grammar and word orders and converts each document to numerical vectors. Widely used techniques for vector representation from the text document include word occurrence matrix, word2vec, Term Frequency (TF), and TF-Inverse Document Frequency (TF-IDF). TF-IDF is an occurrence-based numeric representation of the text document. Term Frequency-Inverse Document Frequency widely used method to transform the text into numerical presentation [13][16].

Numerically, the TF of a particular word in the text document can be computed as:

$$Term\_Frequency\_word = \frac{\text{Frequency of word in text document}}{\text{Total words in the document}} \tag{1}$$

Excluding the common words of least importance in the documents that contribute very little to provide the insights are excluded. Therefore, to reduce the impact of minimally used words in the text document, the TF-IDF of a word may be computed as:

$$Inverse\_doc\_Frequency(w) = \log\left(\frac{\text{Total Number of documents}}{\text{Number of documents containing word w}}\right) \tag{2}$$

3 Research Background

Sentiment classification can categorize an input text sequence into certain types of scores or ratings such as positive, negative, or neutral. Primary text categorization methods include feature vectors indexed by all the words of a given dictionary to represent text documents. Machine learning algorithms develop multiple models using different data sets collected from some standard repository or corpus and each model is analogous to an experience.

3.1 Empirical Data Collection

Dataset for sentiment classification is collected from Kaggle [14]. Text document for sentiment classification is a sentence extracted from social media that contains review comments about several movies. This dataset contains two field text and sentiments. The text includes the actual review comments about the movie. The sentiment is the response variable containing positive and negative sentiments. The training data contains 7086 sentences categorized as 1 with positive sentiment and 0 with negative sentiment.

3.2 Experimental Setup

In this section, we present the experimental set up to illustrate how effectively ensemble techniques (AdaBoost, GBM, XGBM) are applied for sentiment classification. Python machine learning library scikit-learn and natural language tool kit have been used to train and classify the prediction model. Each comment is counted as a record and categorized as positive or negative using machine learning algorithms [29].

3.3 Performance Evaluation Measures

Accuracy is the most common metric applied for the assessment of a classifier which can be extracted from the confusion matrix referred to as error matrix and classification table [18][19][20]. The accuracy of the proposed classification model is obtained from combining the number of true positives and true negatives classes divided by the total number of observations that provide the overall accuracy of the predictive models. Training and testing accuracy of the classifier is measured with different learning rates varying between 0 and 1.

3.4 Cross-Validation

Cross-validation is an evaluation method that attempts to generalize the outcomes quantitatively of a statistical analysis conducted on a dataset. It is conducted irrespective of the training data. Generally, using a round of cross-validation includes splitting the data into two complementary subsets training and testing, performing analysis on training data, and validation analysis using test data. The validation procedure is carried multiple times with different partitions and the mean value is taken as the results of the model to reduce the scattering.

4 Proposed Methodology

This section will examine if certain features alter the probability of unstructured documents for sentiment analysis using GBM, AdaBoost, and XGBM quantitatively. Bootstrap Aggregating (Bagging) and Boosting are popular ensemble techniques. Bagging combines the results of multiple base models to get improved results. Bagging is an ensemble technique for classification works on random subsets of the original dataset. The final prediction is achieved through voting or by taking the aggregate of individual predictions. Subsets from the dataset are taken with replacement.
Bootstrapping is a sampling technique with a replacement where subsets of samples are generated from the original dataset. Bagging applies these subsets known as bags to get the distribution of a complete set. The size of subsets may vary from the original set. Boosting is also widely used as an ensemble technique. The primary purpose of boosting is to emphasize the samples that are hard to classify accurately. Boosting builds the multiple models sequentially by assigning equal weights to each sample initially and then targets misclassified samples in subsequent models. Two popularly used algorithms are Gradient boosting and AdaBoost.

4.1 Gradient Boosting Machines

GBM is an ensemble technique that applies a decision tree as a base classifier. GBM applies boosted machine learning for e-mails extracted from the spam dataset. GBM constructs one tree at a time where each new tree helps to rectify errors caused by the earlier trained tree. Whereas using a random forest classifier the trees don't correlate with previously constructed trees [21-23]. The gradient descent algorithm is applied to minimize the error. A set of training samples $X = \{(x_1,y_{1)}, \ldots\ldots,(x_i, y_i)\}$ is taken from the spam datasets and corpus. Where $x_i \in R^n$ and $y_i \in \{+1, -1\}$ denoting the outcomes for i^{th} training sample indicating +1 as spam and -1 for non-spam e-mail. The voted combination of classifiers F(X) can be written as:

$$F(X) = \sum_{t=1}^{T} w_t f_t(x) \tag{3}$$

where $f_t(x)$: $R^n \rightarrow \{+1,-1\}$ are base classifiers, and $w_t \in R$, the weights for each base classifier in the combined classifiers. A data point (x, y) is classified according to the sign of F(X) and margin yF(X). A positive value for the margin represents spam mail, and the negative value corresponds to legitimate mail (non-spam).

4.2 AdaBoost Classifier

AdaBoost classifier applies a sequence of weak learners such as decision trees (DTs) on modified versions of the text data repeatedly. The AdaBoost method boosts the accuracy of a weak learner by simulating multiple distributions over the training samples. The Adaboost takes the majority vote of the resulting outcomes. Initially, a set of weights is applied to the training samples and then updated after each round of the training. The weights are updated in such a way that weights of the samples classified incorrectly are increased whereas the correctly classified samples are assigned lower weights. During the training process updating the weight mechanism focus the base

learner to concentrate on the harder samples. The overall prediction of the model is computed on the weighted sum of all classifiers [27][28].

$$F(X_i) = sign\left(\sum_{k=1}^{K} \alpha_k f_k(X_i)\right) \quad (4)$$

where K represents the total number of classifiers utilized, $f_k(X_i)$ is the outcome of weak classifier k for corresponding feature X_i, α_k is the weight assigned to classifier k computed as:

$$\alpha_k = \frac{1}{2}\ln\left(\frac{1-\varepsilon_k}{\varepsilon_k}\right) \quad (5)$$

where ε_k represents the error rate of the classifier that is, the number of incorrectly classified samples over the training set divided by the total number of the training set, $F(X_i)$ indicates the combination of all the weak classifiers.

4.3 eXtreme Gradient Boosting Machines

XGBM (eXtreme Gradient Boosting Machines) is an improved gradient boosting classifier. Gradient boosting algorithms can be used for classification and regression problems. GBM classifier performs significantly well on large complex datasets. While classifying social media reviews or customer behavior prediction XGBM demonstrates the importance and impact of DTs boosting as an improved classifier [18-19][21-22]. One of the main reasons behind the success of the XGBM model is its scalability. The scalability of XGBM is because of several algorithmic optimizations such as a novel tree learning algorithm used for handling sparse data. Distributed and parallel computing provides faster learning enabling quicker model investigation. However, overfitting remains a challenge to overcome that can be controlled through parameter tuning and optimization.

Optimization of Model Parameters

- Gradient Boosting Classifier – this model is generated as an additive model using arbitrary differentiable loss functions for optimizing the classifier accuracy. The main parameters tuned for generalization include loss function, learning rate, number of estimators, number of sub-samples, and error criteria to terminate the learning. Exponential and deviance loss functions are applied to optimize the model. Improved results are obtained while applying exponential loss function in Gradient boosting. The non-negative learning rate is adjusted in tune when different estimators are applied between 0 and 1 with a default value of 0.1. A large number of estimators are applied from 100 to 1000 for the better performance of the model. The number of sub-samples is taken less than or equal to 1. However, a smaller value will lead to the stochastic Gradient boosting with low variance and high bias. Different learning rates between 0 and 1 were applied to achieve optimum training and testing the accuracy of the classifier. Mean square error is the function applied to measure the quality of a split. Other error metrics, such as mean absolute error, also may be applied. Moreover, Friedman mse is generally the best parameter by default leading to a better approximation.

- AdaBoost Classifier – using AdaBoost classifier we begin with meta-estimators as base estimators where the weights are tuned such a way that difficult samples are emphasized for classification more accurately. Four vital parameters include the base estimator, the number of estimators, learning rate, and the algorithm applied. Here, two base estimators, LR, and DTs are employed to obtain the generalized results over the social media dataset. The maximum number of estimators varies from 10 to 100 and is applied for model learning. The boosting is terminated earlier also in case the perfect fit is obtained. Different learning rate between 0 and 1 is applied to achieve optimum training and testing accuracy of the classifier. However, the learning rate starts shrinking when a large number of estimators are applied. Two popular algorithms SAMME and SAMME.R, are explored to get optimal results.
- XGBM

  eXtreme Gradient Boosting Classifier provides speed enhancement through parallel and distributed computing with the help of cache awareness which makes XGBM quite faster than the original GBM. Additionally, a split finding algorithm is applied to optimize the trees to reduce the overfitting problem that leads to a faster and more accurate classifier over GBM. The GBM generates an additive model in a forward stage-wise manner that allows for the optimization of the differentiable error function. Vital parameters tuned for generalization includes maximum depth of the tree as base learners, number of parallel threads to run the XGB classifier, minimum loss reduction required to make a partition on the leaf node of the tree. The non-negative learning rate between 0 and 1 is applied in tune with different estimators with a default value of 0.1. The number of estimators utilized varies from 100 to 1000 for testing the scalability of the model. Different learning rates between 0 and 1 were applied so that optimum training and testing accuracy of the classifier could be achieved.

4.4 Analysis Results

In this section, we compare the performance of individual classifiers applied in our study by evaluating the confusion matrix of each model in terms of accuracy of training and testing with different learning rates of the model. Results achieved from the rigorous experiments conducted on social media data (online movie reviews) are presented in Table 1.

Table 1. Table captions should be placed above the tables

Classifiers	Learning rate	Accuracy	
		Training	Testing
Ada Boosting	0.050	0.857	0.860
	0.075	0.932	0.936
	0.100	0.963	0.962
	0.250	0.980	0.980
	0.500	0.988	0.986
	0.750	0.990	0.987
	1.000	0.990	0.987
	0.050	0.627	0.634

Gradient Boosting	0.075	0.674	0.689
	0.100	0.675	0.690
	0.250	0.778	0.783
	0.500	0.779	0.784
	0.750	0.780	0.784
	1.000	0.789	0.802
XGradient Boosting	0.050	0.996	0.996
	0.075	0.996	0.996
	0.100	0.996	0.996
	0.250	0.996	0.996
	0.500	0.996	0.996
	0.750	0.996	0.996
	1.000	0.996	0.996
Bagging with KNN	0.050	0.981	0.971
	0.075	0.985	0.974
	0.100	0.989	0.980
	0.250	0.984	0.971
	0.500	0.981	0.968
	0.750	0.988	0.979
	1.000	0.989	0.980

From the dataset two-third of samples are utilized for training the model and one-third of samples for testing and validation purpose. The total size of the dataset is 50000 observations divided into two categories as positive sentiments and negative sentiments.

It is observed that the higher value of a word represents greater importance in the text document. However, when corpus size is varied then large size text documents normally have more occurrences of words than smaller sized text. Term frequency normalizes the occurrence of each word within the size of a text document. If any particular term occurs in all the text documents then the inverse document frequency of that word would be computed as 0. TF-IDF is the product of term-frequency and inverse document frequency. After removing stop words stemming and lemmatization are applied to converts the words into root words. The words which appear in multiple forms having similar meanings such as improve, improved and improvement needs to be considered as one root word improve only.

5 Discussion and Observations

It is evident from the experimental observations that sentiment analysis for social media using ensemble machine learning techniques (GBM, AdaBoost, XGBM) provides a viable solution for text mining tasks and sentiment analysis to analyze user-generated reviews for specific products and services. From Table 1, it is observed that the XGBM classifier outperforms over three models (GBM, AdaBoost, and bagging with KNN) applied for sentiment analysis quantitively. Training and testing accuracy of the XGBM achieved is the maximum (0.996) irrespective of the different learning

rate applied (from 0.050 to 1.00 at regular interval of 0.250) performing consistently. However, before making any generalization similar studies need to be carried out on larger and different scalable datasets with a large number of parameters. Sentiment analysis will provide a competitive edge for the organizations to understand the behavior of their customer for products and services using social media data. This will help them to improve the product branding and maintaining better customer relationships so that revenue could be generated maximum.

SA of unstructured and uncensored modes of delivery will avoid exploiting the public sentiments which have been the common reasons for the downfall and rise of many products or services within the organizations across the globe. This will facilitate in improving the business performance and monitoring the products and services from the customer perspective. Therefore, SA can be utilized as a monitoring tool for the assessment of policy decisions or services rendered to the customers or branding their product. Ignorance may lead to dissatisfaction among the customers consequently losing the product or services or downfall in ratings.

6 Conclusion and Future Work

This paper explored sentiment analysis with automatic extraction and analyzing the reviews and opinions of messages and posts on social media using two novel approaches of machine learning (Bagging and Boosting). Sentiment analysis could be utilized for monitoring consumer opinions, products, and business intelligence as per local needs and global standards. Machine learning classifiers are found to be potential tools for all stakeholders to monitor and track their branding of products and services from the customer's view particularly in the event of a fluctuating situation of marketing trends. Knowledge extracted from social media can be extremely helpful because a large number of opinions expressed about specific topics or trends may lead to vital information related to business policy.

In this paper, it is demonstrated that ensemble learning techniques can be applied as an effective tool for insight sentiments of social media users. Future work will target reviewing comments, optimal feature selection, and comparing various machine learning algorithms for sentiment classification applied to various benchmark datasets extracted from the social media of a wide range of products and services incorporating authenticity and integrity of digital contents. We also plan to replicate and extend our study for text mining and sentiment analysis more intelligently using advanced machine learning techniques such as deep learning with clustering tweets.

References

[1] Isah, H., Trundle, P., Neagu, D.: Social media analysis for product safety using text mining and sentiment analysis. In: 14th UK Workshop on Computational Intelligence (UKCI), pp. 1-7, Bradford (2014).

[2] Mostafa, M.M.: More than words: Social networks text mining for consumer brand sentiments. Expert Systems with Applications 40(10), 4241-4251 (2013).

[3] Weiguo, F., Gordon, M.D.: The Power of Social Media Analytics. Communications of the ACM 57(6), 74-81 (2014).
[4] Almatarneh, S., Gamallo, P.: Comparing Supervised Machine Learning Strategies and Linguistic Features to Search for Very Negative Opinions. Information 10(1):16 (2019).
[5] Zou, H., Tang, X., Xie, B., Liu, B.: Sentiment Classification Using Machine Learning Techniques with Syntax Features. In: International Conference on Computational Science and Computational Intelligence (CSCI), pp. 175-179, Las Vegas, NV (2015). doi: 10.1109/CSCI.2015.44.
[6] Namugera, F., Wesonga, R., Jehopio, P.: Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. Computer Soc Networks 6(3) 2019.
[7] Balahur, A.: Sentiment Analysis in Social Media Texts. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 120–128, Atlanta Georgia (2013).
[8] Eman, M.G., Younis: Sentiment Analysis and Text Mining for Social Media Microblogs using Open-Source Tools: An Empirical Study. International Journal of Computer Applications 112(5), 44-48 (2015).
[9] Păvăloaia, V-D., Teodor, E-M., Fotache, D., Danileț, M.: Opinion Mining on Social Media Data: Sentiment Analysis of User Preferences. Sustainability 11(16), 4459(2019).
[10] Hotho, A., Nürnberger, A., Paaß, G.: A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology 20(1), 19-62 (2005).
[11] Ikonomarkis, M., Kotsiantis, S., Tampakas, V.: Text classification using machine learning techniques. WSEAS Transactions on Computers 8(4), 966-974 (2005).
[12] Mooney, R-J., Nahm, U-Y., Mooney, R-J.: Text mining with information extraction. In: Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, W. Daelemans and T. du Plessis and C. Snyman and L. Teck (Eds.), pp. 141-160, Bloemfontein, Van Schaik: South Africa (2003).
[13] Bespalov, D., Bing, B., Yanjun, Q., Shokoufandeh, A.: Sentiment classification based on supervised latent n-gram analysis. In: Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11), pp. 375-382, New York, NY, USA (2011).
[14] https://www.kaggle.com/c/si650winter11/data
[15] Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1-135 (2008).
[16] Dave, K., Lawrence, S., Pennock, D.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International conference on the world wide web (WWW-03), pp. 519-528, New York: ACM Press, (2003).
[17] Alzamzami, F., Hoda, M., Saddik, A-E.: Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation. IEEE Access 8, 101840-101858 (2020). doi: 10.1109/ACCESS.2020.2997330.
[18] Friedman, J.: Greedy Function Approximation: A Gradient Boosting, Machine. The Annals of Statistics, 29(5), 1189-1232 (2001).

[19] Hastie, T., Tibshirani, R., Friedman, J.: Elements of Statistical Learning Ed. 2, Springer (2009).
[20] Ross, Q.: C4.5: Programs for machine learning. Morgan Kaufman Publishers, San Mateo, CA (1993).
[21] Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1-47 (2002).
[22] Breiman, L.: Pasting small votes for classification in large databases and on-line. Machine Learning, 36(1), 85-103 (1999).
[23] Louppe, G., Geurts, P.: Ensembles on Random Patches. Machine Learning and Knowledge Discovery in Databases, 346-361 (2012).
[24] Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1-47 (2002).
[25] Forman, G.: An experimental study of feature selection metrics for text categorization. Journal of Machine Learning Research 3, 1289-1305 (2003).
[26] Kohavi, R.: The power of decision tables. In: The Eighth European Conference on Machine Learning (ECML-95), pp. 174-189, Heraclion Greece (1995).
[27] Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001).
[28] Freund, Y., Schapire, R.: A Decision-Theoretic Generalization of online Learning and an Application to Boosting, 1995.
[29] Pedregosa: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830 (2011).
[30] Pak, A., Paroubek, P.: Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 436-439, Los Angeles, CA, USA, (2010).