SENTIMENT ANALYSIS OF PRODUCT REVIEWS IN THE ABSENCE OF LABELLED DATA USING SUPERVISED LEARNING APPROACHES

Waqar Muhammad[1*], Maria Mushtaq[2], Khurum Nazir Junejo[3], Muhammad Yaseen Khan[4]

[1,2]Saint Louis University, St. Louis, Missouri, United States

[3]Ibex CX, Karachi, Pakistan

[4]Centre for Language Computing, FoC, Muhammad Ali Jinnah University, Karachi, Pakistan

E-mail: waqar_m@hotmail.co.uk[1*] (corresponding author), maria mushtaq@slu.edu[2], junejo@gmail.com[3], yaseen.khan@jinnah.edu[4]

DOI: https://doi.org/10.22452/mjcs.vol33no2.3

ABSTRACT

With the growing pace of internet usage, there is a vast variety of diverse individual opinions and thoughts avail-able online. Consumer reviews can act as a feedback and as well as a pool of ideas for which they can be of immense importance to any business. With the growth and popularity of opinion-rich resources such as online review sites and personal blogs, people now can and do, actively use information technology to seek out and un-derstand the opinions of others to decide whether to buy a product or not. Social media websites such as Facebook, Twitter, and e-commerce websites such as eBay, Amazon, etc. are being widely used to communicate viewpoints effectively. Assigning a positive or a negative sentiment to these reviews can help companies understand their users and also help users to make better decisions. Sentiment analysis being a challenging task can be tackled using supervised machine learning techniques or through unsupervised lexicon based approaches if labelled data is unavailable. In this study, we show that in absence of labelled product reviews of a particular website, labelled product reviews from a different website can be effectively used to train the supervised techniques to achieve a comparable performance to the unsupervised lexicon based approaches. This approach also benefits by covering all of the product reviews which the lexicon based approaches fail to do so. We deduce this by comparing five su-pervised approached and three lexicon based approaches on iPhone 5s reviews gathered from Amazon, Facebook, and Reevoo blog. Furthermore, we also found that unigram features combined with bigram features give the best results, and the effect of varying the training data size on the performance of ML classifiers in some cases was significant whereas in other cases it did not have any effect. Our results also suggest that reviews from Amazon are easiest to classify, followed by reviews from Reevo, and Facebook reviews are the hardest to classify.

Keywords: Sentiment Analysis, Product Review Mining, Machine Learning, Lexicon based approaches, Sentiment Dictionary

1.0 INTRODUCTION

Internet has not only changed the way we work, but has also changed how we interact with others, express our-selves, and even how we buy stuff. More and more people are adopting online shopping, and this industry is growing day by day. People from around the globe have an easy access to the online social networks (OSN), encouraging the freedom of speech, especially amongst youngsters to express their opinions. OSN such as Twitter, and Facebook and e-commerce websites such as Amazon, and eBay are open platforms used by numerous people to interact and share their viewpoints about various products. Although they cover a wide range of aspects but are often unorganized and unstructured thus resulting in the loss of their semantic meaning. These opinions still hold a variety of sentiments that are diverse in nature reflecting either positive or negative attitude towards the products. Classifying these reviews as positive or negative automatically can summarize and quantify the sentiments of the reviewers. This quantified sentiment can then help new buyers benefit from experience of previous buyers without having to read through the reviews. Extremely negative reviews can also be identified to help improve product performance or design. These insights also help the sales and marketing teams to identify what buyers like or dislike about their products. These insights can provide a company the necessary edge over its competitors, thus increasing its sales.

Quantifying the sentiments in a product review is not an easy task. Predominantly there exist two types of approaches for semantic analysis (SA): supervised machine learning approaches, and (unsupervised) lexicon based approaches. Supervised approaches are based on machine learning (ML) algorithms such as Naïve Bayes, Support Vector Machines, Logistic Regression, etc. They require a labelled set of product reviews (referred to as training data) to learn the predictive model. On the other hand, lexicon based approaches use predefined lexical dictionar-ies [1], thus not requiring labelled training examples. These lexicon based approaches can also be thought of as an expert system or knowledge-based approach. Supervised approaches have the advantage of achieving higher accuracy, but their reliance on labelled data is a bottleneck as it requires a tedious process of reading through each review and labelling them as positive or negative. Due to the high cost of labelling the data, lexicon based approaches were developed as they do not require labelled data and thus can be applied directly without learning any training model. But, they suffer from a problem of their own, they may not cover all the reviews i.e. they fail to assign a sentiment label to every review. To reach the best of both worlds, we propose using labelled data from one source to train supervised algorithms and then apply it to the product reviews of the domain of interest. The proposed solution demonstrates performance similar to that of lexicon based approach with the benefit of hundred percent coverage. Thus, this strategy empowers us to use supervised approaches even in the absence of labelled data from the domain of interest. The contribution of this study can be summarized as follows:

- For the first time, a detailed comparison of supervised ML and lexicon based approaches is presented on same set of product reviews to determine which will perform better under which scenarios.
- A technique is proposed to train supervised ML approaches in the absence of labelled product reviews from the domain of interest, and its effectiveness through empirical evaluation is also established.
- The first comparison of Amazon, Reevo blog, and Facebook product reviews is presented.

Our comparison is based on three lexicon based approaches namely SentiStrength, SenticNet, and Happiness Index, and five supervised ML approaches, namely Naïve Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), K-NN, and Random Forests (RF). The training data was built by extracting the product reviews of iPhone 5s from Amazon. The test set was constructed by extracting the reviews from Amazon, Reevoo blog, and Facebook for the same product. Using these training and test sets; we compare the above mentioned approaches in terms of prediction performance, and coverage. Our findings suggest that in cases where labelled training data is unavailable, supervised ML classifier trained on data from a different source should be used if coverage is important, otherwise lexicon based approaches may be used for similar performance but with lesser coverage.

The rest of the paper is organized as follows; we provide a literature review in Section 2. Description of our methodology in presented in Section 3. Dataset details, evaluation setup, and performance metrics are also described in this section. Section 4. presents the results and the discussion about the performance of the various models. Finally, we conclude and state our future direction in Section 5.

2.0 RELATED WORK

Various strategies have been applied by researchers in the struggle to improve SA. ML and lexical/dictionary based approaches have been the predominant of these approaches. A lexical approach utilizes a dictionary containing words along with their manually assigned polarity. The polarity of a word can be either positive or negative. For each review, two scores are calculated using these polarities. If a word is present in the dictionary, and its polarity is positive then its polarity value is added to the positive polarity score, whereas words with negative polarity are added to the negative polarity score. A max (or some other) function is used to determine the final sentiment of the document from the two scores. This approach has been carried out by [2] and [3] to determine the sentiment (or polarity). [4] subtracted the association strength of a word from the association strength of a set of negative words to a set of positive words to compute its semantic orientation. They were able to achieve an accuracy rate of 82%using two different semantic orientation statistic metrics. Still there exists weakness in lexicon based approaches in terms of coverage and the necessity to redefine sentiment lexicons for each domain [5] [6]. A very comprehensive study was carried out by [1] to compare existing lexicon methods through their newly developed web toll iFeel [7]. They compared eight existing lexicon based SA approaches, namely linguistic count and word count (LIWC), Happiness Index, SentiWordNet, SailAil Sentiment Analyzer (SASA), positive and negative affect schedule for Twitter (PANAS-t), Emoticons, SenticNet and SentiStrength. Seven of these eight methods were combined into a new ensemble approached named iFeel. It successfully achieved the highest level of coverage and F-measure.

Supervised ML techniques generally outperform lexicon based approaches in terms of both accuracy and coverage, because they have the advantage of learning from training data. [8] perform SA on movie reviews by discriminating positive and negative sentiments. They collected the reviews from Internet Movie Database (IMDb). Movie ratings present in the form of stars or numeric digits were converted into labels. They found that ML algorithms outperformed the human-produced baselines (data was human annotated). The results suggested that SVM yielded relatively higher accuracies based on different features as compared to Naïve Bayes. [9] performed sentiment classification using distant supervision i.e. the training data consisted of emoticons which served as noisy labels. The corpus comprised of tweets from Twitter that were gathered using query terms including hash tags of products, companies, and people. They used three ML techniques namely Naïve Bayes, Maximum Entropy, and SVM. Unigram, bigram, and part of speech (POS) tags were used as features. Their results showed that SVM performed better than the rest, and that unigram were the most effective among the different feature types thus obtaining an accuracy of 82.9%. A similar work was carried out by [10], in which they retrieved text messages of popular newspaper and magazines such as New York Times, Washington Post, etc. but they classified texts as objective, positive, and negative. Their classifier was based on multinomial Naïve Bayes with n-gram and POS-tagged features.

Only a few studies have been carried out for SA of product reviews. [11] conducted experiments on product reviews gathered from CNet and Amazon. They used complex strategies such as dependency parsing, negating modification, etc. of which stemming improved unigram baseline but linguistic features hurt the performance. They showed that n-grams up to trigrams can improve performance but it was not clear whether performance for n-grams on large scale dataset remains consistent. [2] performed classification using the semantic orientation of words based on product attributes. [6] compared various SA approaches to determine the best features for opinion extraction from reviews. Using relaxation labelling technique to assign polarity, they computed point-wise mutual information based on the search engine hit counts. On the other hand, [12] performed comparative experiments on sentiment classification with higher order n-gram using three algorithms Winnow, a generative model based on language modelling, and a discriminative classifier named Passive-Aggressive algorithm. The experiment was conducted on a very large corpus of over twon hundred thousands online reviews available on Froogle, and the discriminative model significantly outperformed the other two.

Recently, there have been a number of studies on SA using Hadoop that address the problem of processing large, complex, and unstructured data along with achieving high classification accuracy. Hadoop is a tool designed to process large data sets using divide and rule methodology. It consists of a Hadoop distributed file system (HDFS) for storing data, and MapReduce program to process data (both inspired by google technologies on MapReduce and GoogleFileSystem) [13]. The authors obtain real-time data from twitter using streaming API. They achieved an overall accuracy of 72.27% by using different lexical sources such as open NLP, wordnet, and sentiwordnet. Words not covered by these lexicons were assigned sentiment values by the PMI-IR 2 algorithm.

[14] proposed their own MapReduce program for SA, but they could only obtain average results. Their main aim was to improve text classification and to resolve the problem of dat sparsity. [15] proposed a dictionary based technique for classifying text using Hadoop. The data set comprised of large number of tweets collected from twitter using Twitter API, and twitter4j. These tweets were stored into HDFS using single node configuration. The polarity of negation words was reversed and blind negation words were labelled as negative thus achieving an accuracy of 75%. [16] collected product reviews from Amazon, and used Hadoop's multimode cluster with MapReduce program. Mapper involved parsing files and information extracting, where as the Reducer consisted of stop word removal, polarity reversal of negation words and a polarity calculator to calculate sentiment score of the reviews based on separate dictionaries of positive and negative, and stop words. Their experiment yielded relatively small error rate for both negative and positive reviews i.e 5.89% and 2.38% respectively.

Although a number of techniques have been developed for SA but very little attention has been given to the task of empirically comparing supervised ML and lexicon based approaches. [17] survey the sentiment analysis techniques developed over the last century and discuss how they have evolved into current day techniques. [18] in addition to a comparison of sentiment analysis techniques also discuss the tasks and the challenges ahead of the sentiment analysis community. [19, 20, 21, 22] on the other hand compare sentiment analysis techniques empirically, out of which [19] and [20] only compare lexicon based approaches, and [21] compare supervised approaches. Whereas [22] compare the classification performances of six pre-processing methods using four supervised ML classifiers on five Twitter datasets. [23] like us, evaluate the SA approaches in a cross domain setting, however they only try to determine which feature selection provides better results in this setting for supervised ML classifiers. We were only able to find one research article that compares ML and lexicon based SA approaches [24]. They compare the techniques on the movie review dataset collected from

IMDb. However, our study compares ML and lexicon based approaches for product reviews collected from various OSN in a cross-domain setting.

3.0 METHODOLOGY

The prediction of discrete valued sentiment (target) variable is referred to as a classification problem. It aims to determine the sentiment (or class/category) of an instance. In this paper, the target is to predict whether a review is positive or negative. Therefore, the problem is a binary class classification problem, in which the class label P refers to reviews containing positive sentiment, whereas class label N refers to reviews containing negative sentiment. For this study, we treat the class P as the positive class.

The binomial classification problem of predicting the sentiment contained in a text document (product review) can be formally defined as follows. Given a set of labelled reviews $L = \{\langle x_i, c_i \rangle\}_{i=1}^{|L|}$ where $c_i \in C = \{P, N\}$ denotes the category of document $\mathbf{x}_i$ and $|L|$ is the total number of labelled documents; learn a classifier that assigns a category label from P or N to each review in the set $U = \{\langle \mathbf{x}_i \rangle\}_{i=1}^{|U|}$. This is a supervised learning setting in which it is assumed that the joint probability distribution of reviews and sentiment categories is identical in sets U and L (although this is not guaranteed in practice). In other words, the task is to learn to approximate the unknown target function $\Phi' : U \rightarrow \{Y, N\}$ by the classifier function $\Phi : U \rightarrow \{Y, N\}$ such that the number of reviews in U for which $\Phi(\mathbf{x}_j) \neq \Phi'(\mathbf{x}_j)$ is minimum. It is important to note that for lexicon based approaches, the classifier function $\Phi : U \rightarrow \{Y, N\}$ is directly derived from lexicon dictionary and hence there is no learning of classifier function $\Phi : U$ from the labelled set L. A review is represented as a boolean (or in some cases real) valued vector $\mathbf{x}_i = \langle x_{i1}, x_{i2}, \ldots, x_{i|A|} \rangle$ where x_{ij} indicates the value of the attribute j for review i. The integer $|A|$ is the number of attributes used in L and U (after standard pre-processing). The category labels are assumed to be just symbolic labels without semantics and no additional knowledge of a procedural or declarative nature is available.

In the remainder of this section, we describe the lexicon and ML based approaches used in this paper. We also describe the data and its cleaning process in this section. Evaluation setting and metrics are also described in this section.

3.1 Lexicon Based Approaches

All the lexicon based approaches have in common a dictionary of words having some score that hints towards their polarity. They differ in terms of the source of these words, dictionary size, and the methodology used to assign them a score. The process of building such a lexicon is subjective, therefore all these dictionaries only have a small overlap. Similarly, a word may be deemed by one expert to have a positive sentiment, whereas it may be deemed by others as neutral or even negative. Furthermore many words inherently do not contain a positive or negative orientation, but it is the context in which they are used that makes their polarity positive or negative. Due to the aforementioned reasons, there is no standard sentiment lexicon. Often times there might be a tweet, or a blog, that does not contain a word that has a polarity, in which case, it is said that the lexicon does not cover that particular piece of text and thus no score is assigned to it. This problem is referred to as the coverage problem. Lexicon based approaches have a major benefit of not requiring any data for training, and thus can be used as off the shelf solution. Below we describe the three most widely used sentiment lexicons.

3.1.1 SentiStrength

SentiStrength is a lexicon consisting of 2,310 words exhibiting sentiments based on the Linguistic Inquiry and Word Count (LIWC) dictionary [25]. Each word has a human-assigned sentiment score from −5 to 5. Words having a score from −5 to −1 are considered as negative, whereas words having a score from 1 to 5 are regarded as positive words. SentiStrength splits the given text into words and separates out emoticons and punctuation, but in our study, we already remove these artefacts during the preprocessing phase. Each word is then mapped with the associated sentiment score defined by SentiStrength. The overall classification of the sentence is marked as positive if the total positive score is greater than the total negative score and vice versa. SentiStrength is a combination of supervised and unsupervised classification methods implemented by authors after a comprehensive comparison of the several methods [26].

3.1.2 Happiness Index

Happiness Index proposed by [27], calculates frequency and average psychological scores for the Affective Norms for English Words (ANEW) dictionary [28]. ANEW is a set of 1,034 words bearing scores for psychological valence (good−bad), arousal (active−passive), and dominance (strong−weak), and their semantic differentials. Based on these dimensions words are assigned a happiness score on the scale between 1 to 9. In our study, we consider the words with scores between 1 to 4 as negative words, whereas words with scores between 5 to 9 are regarded as positive words.

3.1.3 SenticNet

SenticNet is a semantic measure for concept-level SA. It uses common-sense knowledge by exploiting both artificial intelligence and semantic web techniques in order to recognise, process, and interpret natural language opinions over the Web. While identifying concepts, it assigns a sentiment score ranging from −1 to 1 [29]. We consider words with values below 0 as negative words, whereas words with scores above 0 are considered as positive words. [30] also compare the capacity of SenticNet with SentiWordNet for detecting opinion polarity over a collection of 2,000 patient opinions and conclude SenticNet to be more accurate.

3.2 Supervised Approaches

The performance of supervised approaches is very much dependent on the quality and quantity of the data. Given sufficiently large enough data, the difference in performance on different supervised approaches tends to be in-significant in many cases. Since we have limited data, different algorithms tend to perform differently. Some algorithms are too slow, especially for text classification, because the feature space can easily grow to hundred thousand features. Therefore we do not experiment with the computationally expensive techniques such as deep learning, instead we use state of the art supervised ML algorithms that have shown very good performances on reasonable sized datasets. The five algorithms used are described below.

3.2.1 Naïve Bayes

Naïve Bayes [31] is a probabilistic algorithm that is based on the Bayes Rule. Abstractly, Naïve Bayes is a conditional probability model determining the probability of class C given the evidence that event X will occur, where the attributes in event X are assumed mutually independent, therefore:

$$P(C_k|X) = \frac{P(C_k)\ P(X|C_k)}{P(X)} \tag{1}$$

Where C_k is the specific class, X is the text vector for classification, preprocessed and then tokenized, thus $X = \{x_1, x_2, x_3, \ldots, x_n\}$, $P(C_k)$ is the $prior$ probability of class k, $P(X|C_k)$ is the $likelihood$, the probability of text appearing in the given class k, $P(X)$ is the $evidence$ (normalising factor), probability of text X.

Carrying with the above equation, the $evidence$ is often shunned, thus, we assume the classifying function reduced as:

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{argmax}\ P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \tag{2}$$

3.2.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) are non-probabilistic binary linear classifiers. They project the data into a higher dimensional feature space where a hyperplane is learned to discriminate between the points of the two classes [32]. The hyperplane is such that it maximises the margin between the closest points of the two classes, thus achieving a better generalization over the unseen data. It uses the following discriminant function:

$$g(x) = w^T f(x) + b \tag{3}$$

Where w is the weights vector, b is the bias term, and f(x) represents a non-linear mapping from input space to high-dimensional feature space.

3.2.3 Logistic Regression

Logistic Regression (LR) is an alternative to linear discriminant analysis and can be seen as analogous to linear regression. However, it is based on quite different assumptions. Firstly, the conditional distribution is a Bernoulli distribution rather than a Gaussian distribution. Secondly, the predicted values are probabilities through a logistic function. It nicely measures the relationship between the categorical dependent variable and one or more inde-pendent variables. It can outperform SVM and NN on problems with a large number of features, thus making it a successful classifier for text classification. Its decision function can be given as:

$$P(C_k) = \begin{cases} \frac{1}{1+exp^{w_0+\sum w_i x_i}} & if, k = 1 \\ \\ \frac{exp^{w_0+\sum w_i x_i}}{1+exp^{w_0+\sum w_i x_i}} & otherwise \end{cases} \quad (4)$$

Where X is a vector of real-valued features $\{x_1, x_2, \ldots, x_n\}$, W is a vector of weight calculated by LR $\{w_0, w_1, \ldots, w_n\}$ and C_k is the Boolean class.

3.2.4 K-Nearest Neighbour

K-Nearest Neighbour (K-NN) is an instance-based learning approach that compares new instances with instances seen in training instead of performing generalization. Given a new instance for prediction, the distance or simi-larities are computed between the instance and the training data to make a decision [33]. Each instance represents a point in an n-dimensional space. The K-NN algorithm searches for k nearest neighbours in the pattern space where. The majority vote of these k nearest neighbours determines the label of the unknown instance.

3.2.5 Random Forest

Random forests (RF) are an ensemble approach proposed by [34]. They are based on notion of bootstrapping and selection of random subset of features i.e. they operate by constructing a multitude of decision trees at training time on a random subset of the training examples. At each candidate split in the learning process, a random subset of the features is selected. The mode of the classes of the individual trees is the predicted class label for a record. RF have been recently performing very well on many text classification problems. RF have been shown to be more robust to overfitting than other decision tree algorithms.

3.3 Dataset

For this study, we use product reviews of iPhone 5s crawled from three different Internet sources. The first dataset consists of reviews collected from Amazon. It is the largest Internet-based retailer in the world. Amazon was chosen based on the number of reviews, the length of the reviews, and their quality providing real-time data. Moreover, along with the reviews, Amazon provides a scalar rating (number of stars out of five) for every product review. We arbitrarily consider three stars and above as positive reviews and the remaining as negative. Our study, however, does not deal with the neutral reviews. Practicing the same criteria, we crawled data from Reevoo, a company that provides ratings and reviews as a service for multinational brands and retailers. It assures the availability of genuine reviews to those who shop online and has a significant number of comments. Lastly, we extracted sufficient data in the form of comments from Facebook pages using its Graph API. Since Facebook comments have no star ratings, each comment was read and then labelled by at least three people; and comments with two or more positive labels were annotated as the positive label and vice versa.

3.4 Preprocessing

We preprocess the data for any artifacts that can effect the performance of the sentiment classifiers. First hyperlinks and URLs are removed as they do not contain any sentiment value. Reviews exceeding more than 10 lines are considered fake and, hence, discarded. Irrelevant data collected from Facebook is omitted. Special characters and emoticons (like `:)` indicating happiness, `:(` indicating sadness or sorrow, `:p` indication happiness and amusement etc) are removed because the sentiment lexicon used by us do not provide polarity scores for special characters and symbols. Hashtags and name anchors (`#` and `@` followed by string literal without space and special symbol respectively) do not contain any meaningful sentiment, therefore are removed from comments. This also helps us to identify duplicates that needs to be removed from the data. Comments bearing tags solely are deleted altogether. Facebook data consisted of a large number of neutral comments that are not useful, these comments were removed from the dataset. Furthermore all reviews were converted to lowercase and numbers were removed.

Table 1: Distribution of positive and negative reviews in the dataset

	Positive	Negative	Total	Positive Ratio in %
Amazon	2731	2729	5460	50.01
Reevoo	994	219	1213	81.94
Facebook	354	272	626	56.54

Facebook is the largest OSN where people, especially teenagers who purchase products, frequently interact with each other and have a very high churn out percentage. Although the amount of data available on the Twitter is immense and a number of studies have been reported to use data on Twitter [35] [10], we observed that the sentiments seen on Facebook were more diverse than on any other OSN with very less number of research involving Facebook data. Overall, we managed to accumulate around twenty thousand reviews, 2209 posts and 31231 comments related to iPhone 5s from Amazon, Reevoo, and Facebook. The distribution of positive and negative data after the preprocessing step is shown in Table 1.

3.5 Evaluation Setup

Since lexicon based approaches do not use training data, a separate test set is used for all the results. 30% of the data from Amazon, Reevo, and Facebook is held out as testing data. The performance of the ML and as well as the lexicon based approaches are reported on this test set and is not used by ML algorithms during training. We used multiple settings for comparison. In the first setting, training data of Amazon, Reevo, and Facebook was used to learn separate ML models and then evaluate their performance on their respective test sets. Whereas in the second setting, ML algorithms were trained on the Amazon training set only, and then same ML models were applied on the test set of Amazon, Reevo, and Facebook to examine the performance of the classifier on reviews from a different domain.

RapidMiner implementation was used for evaluating the five ML algorithms. The Naïve Bayes algorithm does not have any parameters for tuning, whereas default parameters were used for LR. However, different parameters values were experimented for K-NN and RF. K=3 was chosen as the parameter values for K-NN classifier. For RF, the number of trees and the maximum depth were chosen as 80 and 100, respectively. Classifiers are trained using 10-fold cross-validation on our training data that consist of 50% positive and 50% negative reviews.

3.6 Evaluation Metrics

We use various metrics to evaluate the performance of the classifiers. The first measure used is accuracy, it is the percentage of total correct predictions for both the classes. It is formally defined as follows:

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}$$

where $T_P = |c_i = P, p_i = P|, T_N = |c_i = N, p_i = N|, F_P = |c_i = N, p_i = P|$, and $F_N = |c_i = P, p_i = N|$, c_i, and p_i are the actual and predicted label of the i^{th} example, respectively. We have treated P as positive class

referring to a positive review, where N refers to a negative review. T_P, T_N, F_P, and F_N are referred to as true positives, true negatives, false positives, and false negatives, respectively.

An ideal sentiment analysis would have 100% accuracy, but for real world problems, it's seldom the case. When the accuracy drops below 100%, the errors could be of two types (F_P and F_N). It is important to know what type of errors our prediction system is making because the misclassification cost is may not be the same for the two type of errors; the consequences of not identifying and resolving a problem indicated in a negative review can be more than classifying a positive review as negative. Therefore we use other measures, precision, and recall to evaluate efficiency of the prediction algorithms. Both, precision, and recall are defined below.

$$Precision = \frac{T_P}{T_P + F_P}$$

$$Recall = \frac{T_P}{T_P + F_N}$$

When dealing with lexicon based approaches, these three measures are also not enough, because these approaches do not assign a label to every review. If the review does not contain any word from the dictionary of a specific lexicon, then no score or label is assigned to that review. Therefore, a fourth measure, called coverage, is also needed to evaluate the performance of a lexicon based approach. Coverage is the percentage of reviews predicted either positive or negative from the entire test data. It is extremely important to obtain high coverage, otherwise a lot of reviews do not get any label.

4.0 RESULTS

In this section, in addition to describing the results, we also present the significant insights extracted after a thorough evaluation of the results.

4.1 Lexicon Based Approaches

Table 2 presents the performance of the unsupervised approaches. The coverage of the three lexicons is presented in Table 3. The best performances are highlighted in bold. From the two tables, it can be seen that the lexicon based approaches perform very well on atleast one data and thus there is no clear winner. For the Amazon dataset, SentiStrength has the highest accuracy and precision, whereas SenticNet achieved the highest recall, thus suggesting that SentiStrength as a winner. Whereas, on a closer look it can be seen that SenticNet has far superior coverage than SentiStrength. Therefore, SenticNet can be declared as the best performer for the Amazon dataset.

Table 2: Results using Unsupervised Learning Approaches

	Amazon			**Reevoo**			**Facebook**		
Approach	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SenticNet	84.03	85.10	**98.06**	89.28	91.00	**97.80**	73.60	74.90	95.10
SentiStrength	**87.49**	**94.19**	91.13	75.79	**95.64**	76.76	**85.70**	**87.80**	94.30
Hapinnes Index	83.00	86.60	95.10	**90.22**	92.00	97.70	79.50	80.60	**96.40**
Average	84.84	88.63	94.76	*85.09*	*92.88*	90.75	79.60	81.10	*95.26*

Table 3: Coverage of Unsupervised Approaches

	Amazon	**Reevoo**	**Facebook**
SenticNet	**97.34**	**87.69**	**72.49**
SentiStrength	80.24	59.12	58.30
Hapinnes Index	78.32	66.97	31.50
Average	*85.3*	71.26	54.09

Sentistrength, which achieved the best accuracy over the Amazon dataset, performs the worst over the Reevo dataset. It has the lowest accuracy and coverage. It also has the highest precision but at the expense of recall. On the other hand, Happiness index has the highest accuracy, with SenticNet not far behind. Since SenticNet has a very high coverage (more than 97%) and is slightly behind Happiness Index, therefore SenticNet can be declared as the best performer over the Reevo dataset.

For the Facebook data, Happiness index has very poor coverage (lower than 32%). SenticNet on the other hand has the highest coverage but lowest accuracy. SentiStrength has more than 14% lower coverage than SenticNet but has more than 12% higher accuracy, therefore SentiStrength can be declared as the best performer over the Facebook dataset.

Amazon dataset seems to be the easiest dataset for classification using the lexicon based approaches. It has the highest coverage and an average accuracy of 84.84% (only 0.25% behind the average accuracy on Reevo dataset). Facebook data seems the most difficult to classify with the least accuracy and coverage. The large difference between the precision and recall of the Facebook data suggests that majority of the misclassified reviews are false positives (FP). Precision on Amazon dataset is lower than precision over the Reevo dataset, whereas it has a higher recall than the Reevo dataset. This indicates that for the Amazon dataset, more negative reviews are misclassified as positive as compared to that for the Reevo dataset.

If the prediction of majority class majority class is considered as the baseline classifier, than the average performance gain in terms of accuracy over the baseline classifier is more than 34%, 3%, and 23% for the Amazon, Reevo, and Facebook datasets, respectively. This suggests, that overall, the lexicon based sentiment analysis tech-niques performed very well on the three datasets.

4.2 Supervised Approaches

For the supervised setting, the 70% of the Amazon data was used for training, and hence a supervised model was learned through various ML algorithms. This same model was used to classify the remaining 30% of the Amazon data, and the whole of the Reevo, and Facebook data. Therefore, for the Reevo, and the Facebook, the test data has a different distribution than the training data (from Amazon) over which the model was learned. This is done to ascertain whether a labelled data from a different source can be used to successfully classify Reevo and Facebook dataset or not. The results for these supervised algorithms are presented in Table 4.

Table 4: Results using Supervised Learning Approaches trained on 70% of Amazon dataset.

Classifiers	**Amazon**			**Reevoo**			**Facebook**		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
NB	62.59	73.98	38.75	30.34	5.87	13.70	31.67	27.72	9.86
LR	73.71	68.99	92.84	85.51	8.79	81.28	**73.92**	26.13	93.31
K-NN	78.84	77.42	75.89	58.51	7.86	45.66	68.91	20.08	51.06
RF	49.96	49.95	40.00	42.06	27.83	33.80	42.06	27.83	9.86
SVM	**90.45**	**95.5**	**97.38**	**90.08**	**90.08**	**100**	72.81	**72.81**	**100**
Average	*71.11*	*73.16*	*68.97*	61.30	28.08	54.88	57.87	34.91	52.81

SVM stands out by performing far better than the lexicon based approaches on the Amazon and the Reevo dataset. For the Facebook dataset, the performance of SVM is not far behind the lexicon based approaches because, SentiStrength even though has an accuracy of 85.70%, its coverage is below 60%, whereas the coverage of SVM (and all the supervised approaches) is 100%. Though, in general, the result of the supervised approaches is quite disappointing. With the exception of LR and SVM, all the classifiers performed very poorly. The average accuracy, precision, and recall for the lexicon based approaches are far superior to that of supervised approaches. It is aston-ishing to see that even for the Amazon dataset, for which the training and test data is from the same distribution, the lexicon based approaches performed better than the supervised approaches. SVM with exceptional performance shows that the supervised models learned on one dataset can be used to successfully classify datasets from different domains. Using this observation, we can build models for classifying product reviews for websites for which we do not have labelled training data (like Reevo and Facebook) and achieve a significant performance gain over the baseline prediction of the majority class (see Table 1).

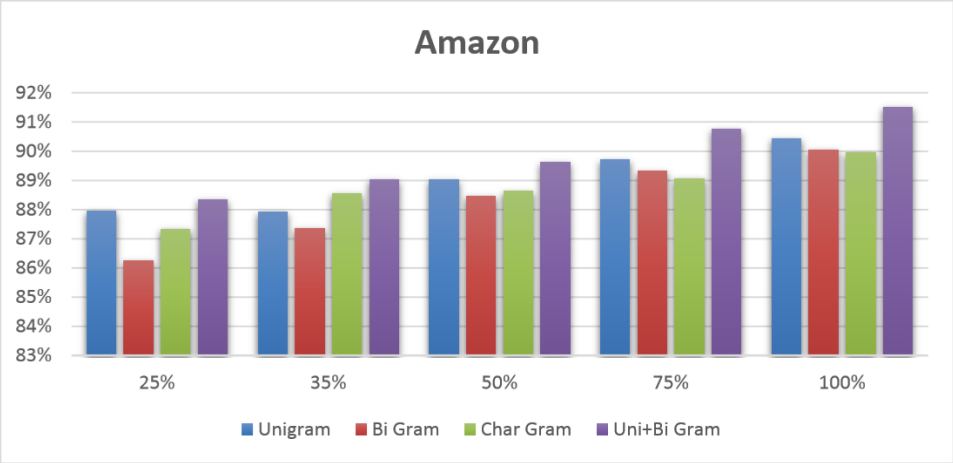

Fig. 1: Amazon training and Amazon testing over 25%, 35%, 50%,75% and 100% dataset size for SVM

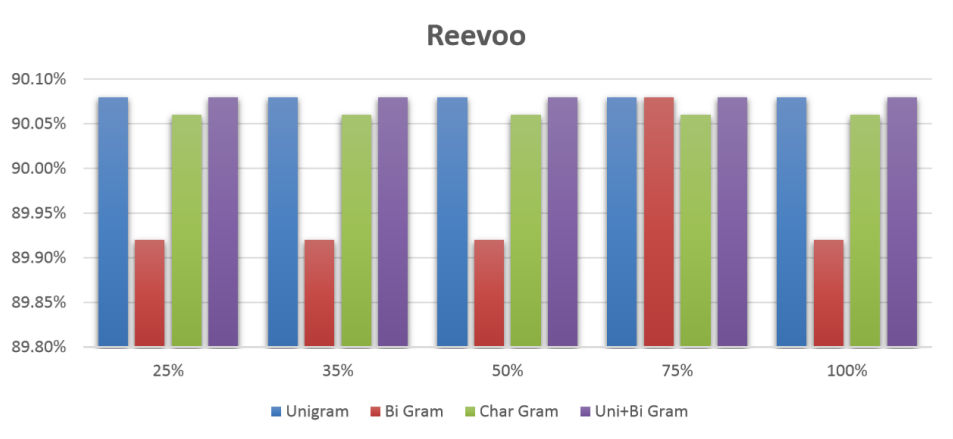

Fig. 2: Amazon training and Reevoo testing over 25%, 35%, 50%,75% and 100% dataset size for SVM

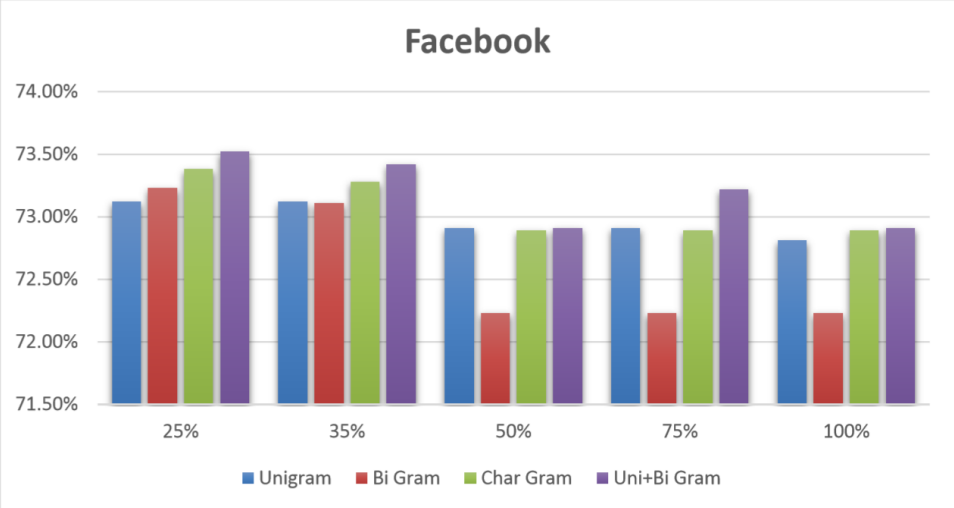

Fig. 3: Amazon training and Facebook testing over 25%, 35%, 50%,75% and 100% dataset size for SVM

4.3 Feature Set for SVM

SVM's superior results lead us to further explore its performance variation using different feature sets and training data sizes. In addition to the unigram approach (on which previous results are based), we also experimented with bigrams, character-grams, and the combination of unigrams and bigrams. In order to ascertain how many examples are necessary for training data to achieve good results, we varied the size of the training data from 25% to 100%of the training set used in Section 4.2. The results are depicted in Figures 1, 2, and 3. In general, the performance of the supervised model increases with the increase in the size of the training data. This trend is confirmed by the Amazon data (Fig. 1), its accuracy increased by 4% by the increase in training data size. Whereas, there is no significant change for the Reevo dataset (Fig. 2), and even a slight decrease (of about 0.5%) over the Facebook data (Fig. 3).

With regards to the choice of feature set, the combined unigram and bigram feature set results in the best perfor-mance. Character-gram performance is better than the bigram and somewhat comparable to the performance using the unigram feature set. The performance using bigram is relatively significantly lower for the Reevo and the Face-book dataset, this could be attributed the fact that the model was learned over the Amazon training data, therefore the chances of a two word combination (a bigram) from Amazon data being matched to a sequence of two words in Reevo, and Facebook data would be much lower than the chances of matching single word (a unigram).

4.4 Same Source Supervised Classification

To determine how much was lost in performance by using training data from a different source for the Reevo, and Facebook dataset, we divided each of the two datasets into training and test sets also. 70% of the data was used for training while the remaining 30% was used to form the test data such that the ratio of the positive and negative reviews remains the same. These results are shown in Table 5. As expected, on average, the performance of the supervised algorithms when trained on the data from the same source is significantly greater than when the data from a different source is used to train them. NB, k-NN, and RF seem to be the major beneficiaries, with their performance increasing dramatically. Whereas, unexpectedly, the performance of LR has decreased. SVM, on the other hand did not benefit much over the Reevo dataset, but its performance significantly increased over the Facebook dataset. Despite the good accuracy of NB, LR, and k-NN over the Reevo dataset, their low precision and recall suggests that these classifiers are relatively performing better at correctly classifying the negative class i.e. TN (true negatives) than the positive class i.e. TP (true positives).

Table 5: Results Using Training and Test Sets from the Same Source.

Approach	**Reevo**			**Facebook**		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
NB	62.14	7.76	25.76	73.31	51.72	35.29
LR	72.25	15.70	40.91	71.70	48.72	67.06
K-NN	85.82	11.11	6.06	72.03	48.44	36.47
RF	90.05	90.05	100	72.67	72.67	100
SVM	90.92	90.92	99.50	80.55	82.43	92.92
Average	*80.23*	43.10	54.44	74.05	*60.79*	*66.34*

4.5 Comparison

Figures 4 and 5 compare the performance of lexicon based approaches with the performance of supervised learning models trained using the data from the same source as the test data, and from a source different than the test data, respectively. It can be seen that when the ML model is learned from a training data belonging to the same source as of the test data, its performance exceeds that of the lexicon based approaches (Fig. 4). Whereas when the training and test set belong to different sources (as is the case for the Reevo and Facebook data in Fig. 5), the performance of supervised ML models varies greatly. In this setting SVM outperforms the lexicon based approaches for the Reevo dataset, but significantly lags behind for the Facebook data. It can then be thought of a tie between the two types of approaches. But, if we are to bring into the picture the coverage factor, then supervised approaches dominate because they are achieving a 100% coverage, whereas the lexicon based approaches are not (Table 3).

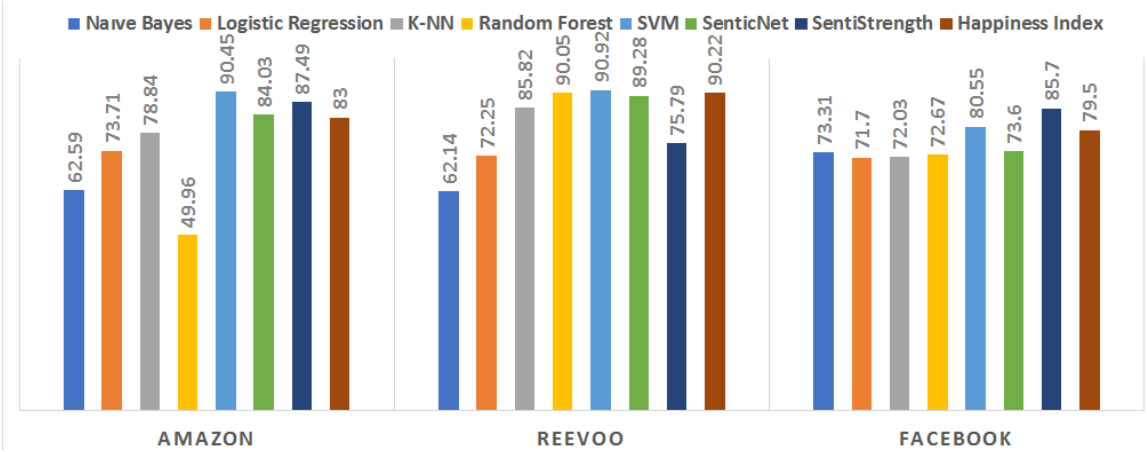

Fig. 4: Comparing Accuracy of Same Source Trained Supervised Classifiers with Lexicon Approaches.

Therefore, to conclude, if enough labelled data is available, then supervised learning approaches will outperform the lexicon based approaches. Whereas, if labelled data is not available, and coverage is important, then a labelled data from a different source should be used to train the model. If coverage is not important lexicon based approach may be used.

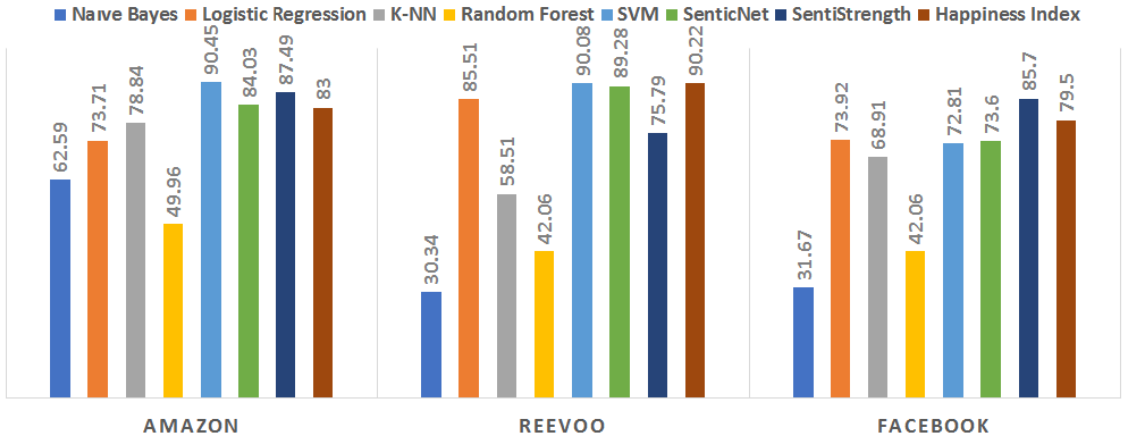

Fig. 5: Comparing Accuracy of Different Source Trained Supervised Classifiers with Lexicon Approaches.

5.0 CONCLUSION

Online shopping has made analyzing consumer reviews almost as important as any other strategic objective for any business. Together with creating a brand image, it gives the required feedback needed by the business to try and improve in areas they lack. It also gives an overall idea of the competitive market in which the business strives hard to expand. Two approaches exist to evaluate the online product reviews, namely, supervised ML approaches, and lexicon based approaches. ML approaches tend to give better results than the later and have 100% coverage, but require labelled training data to build the model, which is seldom available and is expensive to generate. We, therefore, in this study show that even if the labelled training data is missing, supervised learning approaches, specially support vector machines, learned on training data from some other source can be effectively used to classify unseen product reviews for the Reevo, and Facebook dataset. We observed this by comparing the results of five ML classifiers trained using labelled training data from the same and different source with three lexicon based approaches. Furthermore, we also found that unigram features combined with bigram features give the best result. The effect of varying the training data size on the performance of ML classifiers was in some cases

significant whereas in other cases it did not have any effect. According to our results, Amazon dataset was easiest to classify, followed by the Reevo dataset. The Facebook dataset was the most difficult to classify. Overall, we conclude that in case the labelled training data is unavailable, supervised ML classifier trained on data from a different source should be used if coverage is important, otherwise lexicon based approach may be used for similar performance but lesser coverage.

In future, we plan to model this problem as a class imbalance problem because the majority of the reviews in our datasets were positives. Secondly, we believe that the coverage issue for the lexicon approaches can be tackled by creating a dictionary based on the reviews that were left uncovered by them. This will not only help to improve the accuracy but is also likely to improve the coverage as well. Furthermore, the unlabelled data of Reevo and Facebook could be classified using semi-supervised approaches, this may help to improve the overall performance.

REFERENCES

[1] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proceedings of the first ACM conference on Online social networks*. ACM, 2013, pp. 27–38.

[2] 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*. ACM, 2004, pp. 168–177.

[3] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[4] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems (TOIS)*, vol. 21, no. 4, pp. 315–346, 2003.

[5] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 1275–1284.

[6] M. Annett and G. Kondrak, "A comparison of sentiment analysis techniques: Polarizing movie blogs," in *Advances in artificial intelligence*. Springer, 2008, pp. 25–35.

[7] M. Araújo, P. Gonçalves, M. Cha, and F. Benevenuto, "ifeel: A system that compares and combines sentiment analysis methods," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. International World Wide Web Conferences Steering Committee, 2014, pp. 75–78.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[9] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009.

[10] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 1320–1326.

[11] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003, pp. 519–528.

[12] H. Cui, V. Mittal, and M. Datar, "Comparative experiments on sentiment classification for online product reviews," in *AAAI*, vol. 6, 2006, pp. 1265–1270.

[13] S. B. Mane, Y. Sawant, S. Kazi, and V. Shinde, "Real time sentiment analysis of twitter data using hadoop," *International Journal of Computer Science and Information Technologies,(3098-3100)*, vol. 5, no. 3, 2014.

[14] M. Vasuki, J. Arthi, and K. Kayalvizhi, "Decision making using sentiment analysis from twitter," *International Journal of Innovative Research in*, 2014.

[15] D. D. M. K. Kurian, S. Vishnupriya, R. Ramesh, G. Divya, and D. Divya, "Big data sentiment analysis using hadoop," *International Journal for Innovative Research in Science and Technology*, vol. 1, no. 11, pp. 92–96, 2015.

[16] P. Gupta, P. Kumar, and G. Gopal, "Sentiment analysis on hadoop with hadoop streaming," *International Journal of Computer Applications*, vol. 121, no. 11, 2015.

[17] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysisa review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[18] A. Qazi, R. G. Raj, G. Hardaker, and C. Standing, "A systematic literature review on opinion types and sentiment analysis techniques: Tasks and challenges," *Internet Research*, vol. 27, no. 3, pp. 608–630, 2017.

[19] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto, "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ Data Science*, vol. 5, no. 1, p. 23, 2016.

[20] J. Messias, J. P. Diniz, E. Soares, M. Ferreira, M. Araújo, L. Bastos, M. Miranda, and F. Benevenuto, "An evaluation of sentiment analysis for mobile devices," *Social Network Analysis and Mining*, vol. 7, no. 1, p. 20, 2017.

[21] C. Pujari, N. P. Shetty *et al.*, "Comparison of classification techniques for feature oriented sentiment analysis of product review data," in *Data Engineering and Intelligent Computing*. Springer, 2018, pp. 149–158.

[22] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017.

[23] C. Ahuja and E. Sivasankar, "Cross-domain sentiment analysis employing different feature selection and classification techniques," in *Information and Communication Technology for Sustainable Development*. Springer, 2018, pp. 167–179.

[24] P. Chaovalit and L. Zhou, "Movie review mining: A comparison between supervised and unsupervised classification approaches," in *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. IEEE, 2005, pp. 112c–112c.

[25] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[26] M. Thelwall, "Heart and soul: Sentiment strength detection in the social web with sentistrength," *Proceedings of the CyberEmotions*, pp. 1–14, 2013.

[27] P. S. Dodds and C. M. Danforth, "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents," *Journal of Happiness Studies*, vol. 11, no. 4, pp. 441–456, 2010.

[28] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999.

[29] E. Cambria, C. Havasi, and A. Hussain, "Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis." in *FLAIRS conference*, 2012, pp. 202–207.

[30] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "Senticnet: A publicly available semantic resource for opinion mining." in *AAAI fall symposium: commonsense knowledge*, vol. 10, 2010, p. 02.

[31] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM New York, 2001, pp. 41–46.

[32] V. Vapnik, "The nature of statistical learning theory. 2000," *There is no corresponding record for this reference*.

[33] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved k-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012.

[34] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] A. Kumar and T. M. Sebastian, "Sentiment analysis on twitter," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 3, pp. 372–378, 2012.