The University of
Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Hamad, Rebeen Ali and Alqahtani, Saeed M. and Torres Torres, Mercedes (2017) Emotion and polarity prediction from Twitter. In: Computing Conference 2017, 18-20 July 2017, London, U.K..

Access from the University of Nottingham repository:
http://eprints.nottingham.ac.uk/41167/1/SAI_EmotionPolarity.pdf

Emotion and Polarity Prediction from Twitter

Rebeen Ali Hamad

School of Computer Science

University of Nottingham,

Nottingham, UK,

psxrah@nottingham.ac.uk

Saeed M. Alqahtani

School of Computer Science,

University of Nottingham,

Nottingham, UK,

saeed.alqahtani@nottingham.ac.uk

Mercedes Torres Torres

School of Computer Science,

University of Nottingham,

Nottingham, UK,

mercedes.torrestorres@nottingham.ac.uk

Abstract—Classification of public information from microblogging and social networking services could yield interesting outcomes and insights into the social and public opinions towards different services, products, and events. Microblogging and social networking data are one of the most helpful and proper indicators of public opinion. The aim of this paper is to classify tweets to their classes using cross validation and partitioning the data across cities using supervised machine learning algorithms. Such an approach was used to collect real time Twitter microblogging data tweets towards mentioning iPad and iPhone in different locations in order to analyse and classify data in terms of polarity: positive or negative, and emotion: anger, joy, sadness, disgust, fear, and surprise. We have collected over eighty thousand tweets that have been pre-processed to generate document level ground truth and labelled according to Emotion and Polarity. We also compared some approaches in order to measures the performance of K-NN, Nave Bayes, and SVM classifiers. We found that the K-NN, Nave Bayes, SVM, and ZeroR have a reasonable accuracy rates, however, the K-NN has outperformed the Nave Bayes, SVM, and ZeroR based on the achieved accuracy rates and trained model time. The K-NN has achieved the highest accuracy rates 96.58% and 99.94% for the iPad and iPhone emotion data sets using cross validation technique respectively. Regarding partitioning the data per city, the K-NN has achieved the highest accuracy rates 98.8% and 99.95% for the iPad and iPhone emotion data sets respectively. Regarding the polarity data sets using both cross validation and partitioning data per city, the K-NN achieved 100% for the all polarity datasets.

Index Terms—Data Mining, Sentiment Analysis , Sentiment Classification, Emotion , Polarity, Machine Learning, Twitter Data

I. Introduction

Social network services have increasingly been being used by people across numerous platforms to communicate among themselves, developing modern opportunities to monitor opinions that are expressed by people and their individual emotion. Social network platforms such as Twitter and Facebook are have revolutionised communications compared to era before the 2000s, and each presently has hundreds of millions of active monthly users worldwide [1]. Social network services are distinguished by offering the opportunity to openly express and publish opinions, which affects various aspects of information such as services, socialising, products and events discussed and shared by active users. This opportunity has enabled massive volumes of textual data to be generated on social networking services every day, particularly about products and events. Therefore, several multiple companies and organisations are attempting to utilise this deluge of data in order to extract users opinions about their products [2]. Twitter is one of the most popular online social networking and microblogging services. It was created in 2006 and has continually increased its monthly active users, with an average of 313 million [3]. Twitter allows users who have registered to post and read messages, known as tweets. Tweets are limited to a relatively short 140 characters, including spaces and symbols. Tweets containing users opinions about sentiments towards various topics such as events, news, products, films and famous individuals. Tweets of active users resulting in sufficient data are provided to generate indicative data for almost any major trend. Twitter data can be analysed and evaluated in order to extract useful information towards any requested topic by researchers. Accordingly, sentiment analysis and classification about different topic matters from online social networking is a highly innovative method, especially for market analysis. In order to make it more specific, sentiment analysis can be performed on online social networking data from explicit areas. The aim of this paper is to evaluate, analyse and classify the opinions on behalf of users tweets towards smart devices according to emotion (anger, joy, sadness, disgust, fear, and surprise) and polarity (positive and negative). We decided to focus on two of the most popular smart devices, the iPhone and iPad, in order to perform sentiment analysis and classification for these two brands from various locations [4].

In this paper, we firstly present the related work on twitter data and some reviews data mainly to identify sentiment. Then we show the investigative approach explaining the required data needed for conducting the classification procedures , building data sets, and tweets. After that, we present the achieved results that were compared and discussed in experimental results and discussion. Finally, we summarise the findings, and limitations of the research followed by future work in conclusion and future work.

II. Related Work

Sentiment analysis can be performed for various purposes, including for user reviews of products in marketing and predicting elections in political analysis. Sentiment orientation is computed from a different level of the overall sentiment analysis. Sentiment analysis is divided into three levels based on the granularity of text, namely document level, sentence level, and entity and aspect level, in order to perform different detection tasks at various text granularity levels [5].

A. *Sentiment Analysis Document Level*

The document level involves considering the whole document and whether it expresses a positive or negative sentiment. It is usually performed by averaging the sentiment orientation of whole sentences [5]. In this level of analysis, a single review about a single entity is considered [2], [6]. Thus, this document level is applicable to statements when dealing with comparative or evaluative multiple entities without their attributes. However, this becomes a challenge because it might be that all sentences in any reviews in the document may not be related or express any positive or negative opinions. Nevertheless, it is certainly useful for analysing and evaluating users opinions about multiple entities, such as iPad and iPhone, to determine whether the opinions towards entities are positive or negative [7]. A valid example of a positive document would be: I brought a new camera five days back. It is a good camera. I essentially love this camera. Accordingly, in this paper the document level will be used for comparing iPhone and iPad on behalf of users tweets in terms of polarity and emotion, as the document level is suitable for unstructured data such as tweets because of their generally poor grammar and their sentences not being separated to be able to use other levels [7], [8].

B. *Sentiment Analysis Sentence Level*

Sentence level sentiment analysis is used to evaluate the sentiment of each subjective sentence after subjective and objective sentences are separated. This classification level is particularly useful for simple sentences that have a single opinion about any entity, because it is nearly relevant to subjectivity classification. The purpose of sentiment analysis sentence level is to classify opinions expressed in each sentence. Therefore, in this level we may have positive and negative sentences in one document. For instance, iPhone is good. The wireless of iPhone is really bad. However, generally sentences are short documents so that there is no fundamental difference between the sentence and document levels [9]. Thus, this analysis has a limitation which is not applicable for complex sentences. In 2013, different methods were performed in sentiment classification for sentence-level sentiment classification. The sentences were assigned as positive, neutral and negative. Text classification by using document level and sentence level did no provide the details about all aspects of the entity [8]. Hence to obtain these details we may need to use another aspect which is aspect level.

C. *Entity and Aspect Level*

This is a phrase level sentiment analysis. Both previous levels work properly either when each single sentence or the whole document indicates a single entity. However, in many cases, individuals express their opinions about entities that have various aspects, and they have a diverse opinion about each of these aspects [9]. This level of analysis is useful especially in reviews of products, because it can indicate different opinions about various aspects of an entity. For instance, the voice quality of iPad is not good, but touch screen is good.

This level of sentiment analysis is a more accurate method for opinion mining because it directly identifies the entities and their aspects. Therefore, this level can be useless approach for comparing entities without comparing their aspects or attributes [10].

III. REVIEW OF SENTIMENT ANALYSIS TECHNIQUES

The classification techniques of sentiment analysis can be divided into two main approaches: machine learning-based approach (MLA), lexicon-based approach (LBA) [11].

A. *Machine Learning-Based Approach (MLA)*

Machine learning is a part of artificial intelligence (AI), using various algorithms that help devices such as a computer to learn. An algorithm is given a data set whereby it concludes information about the properties of the data to produce a prediction about the new data that it may meet in the future [12]. The ability machine learning algorithms to create predictions about unobserved data is attainable because of the fact that non-random data usually holds patterns that permit machines to generalise. In order to generalise patterns, the machine trains a model with what it determines to be the essential parts of the data. Fundamentally, the ability of algorithms are usually different, to generalise over large sets of patterns, because of the fact that a pattern may differ after being initially seen by the algorithm earlier, therefore it is likely to misinterpret without new analysis. Linguistically, frequently appearing patterns are uncommon, and exceptionally occurring patterns are predominant. This helps MLA to generalise patterns based on the data observed previously. Researchers utilise several techniques and features in the learning process. Selecting an appropriate set of features is one of the main tasks in sentiment classification. Sentiment mining uses MLAs that comprise two stages. First, feature extraction from the training set and transforming them to feature vectors. Second, training the classifier on the feature vectors, not training. Converting a text into a feature vector is a significant section of MLAs in sentiment classification. The relevant information can be achieved from features that were obtained from the text. This helps classifiers be more efficient in the learning process by identifying relevant features to be considered, and reducing the amount of data to be analysed. In sentiment classification, methods of MLA are explained as described below [13].

1) Unsupervised learning: In sentiment classification, it is challenging to build labelled training documents; however it is not difficult to collect such documents using unsupervised learning technique, as this method overcomes these difficulties by grouping data into relates sets. In 2002, Turney used unsupervised algorithm for sentiment analysis to classify user reviews regarding polarity [6]. They utilized phrases in order to classify reviews rather than words. The algorithm was implemented based mainly on Pointwise Mutual Information and Information Retrieval to figure out semantic orientation of the given review. In this study three stages were performed: (1) extracting phrasal lexicon from users review; (2) identifying each phrases polarity; and (3) determining whether reviews

were negative or positive based on average phrase polarity. In 2008, Zegibalov et al. described and evaluated a new technique of automatic seed word selection for unsupervised sentiment classification of product reviews in Mandarin Chinese [14]. The technique did not require any annotated training data, it only needed information about usual appearing negations and adverbials in order to iteratively find sentiment-bearing items. The results from this technique were close to those of the supervised classifiers. In 2010, Rothfels and Tibshirani performed unsupervised sentiment classification using an automatic selection of negative and positive sentiment items of English movie reviews, achieving 65.5% accuracy [15]. In 2011, Zhai et al. proposed an unsupervised method for sentiment classification in order to identify the evaluative sentences in the online review, achieving an F-score of 76% [16]. Singh in 2012 used unsupervised learning on collected blog data concerning socio-political, cross-cultural psychological and sociological topics, converting the collected blog data into vector space representation, and then performing sentiment classification; however, the accuracy of this study was not reported [17]. In 2015, Hridoy et al. pointed out a straightforward technique to perform sentiment classification based on an unsupervised linguistic method in order to calculate the overall sentiment scores of each sentence [18]. The reviews were classified in terms of polarity: positive and negative. SentiWordNet was used, which is a lexicon resource for sentiment classification. Divya proposed a technique that focuses on unsupervised learning approach using real-time data from Twitter, performed for classifying the review to determine negative, positive and neutral sentiments [19].

The limitation of such approaches is that they typically require a large volume of data to be trained accurately. Unsupervised learning is a method in which the algorithm uses only the independent attribute values, with no response attribute values, and the learning task is to gain some understanding of relevant structure patterns in the data. In 2015, Madhoushi et al. claimed that entirely unsupervised methods usually produce incoherent features, because the objective functions of the feature methods do not regularly correlate well people judgements [20]. However, knowledge can be achieved using unsupervised learning approach about the data without any annotation.

2) Supervised learning: Classification algorithms are used by MLA to classify text. This mainly consists of two sets of data: training and a test set. The training set is utilised by an algorithm classifier to learn the differentiating characteristics of texts and then a test set that is utilised to validate the performance of the algorithm classifier. The algorithms are most likely suitable for sentiment analysis, especially supervised classification. In the text categorisation, a successful process can be achieved using machine learning techniques, the most suitable of which in the natural language processing domain are, according to [?], Nave Bayes (NB), Support Vector Machine (SVM), and K-Nearest Neighbour (K-NN). However, their performance differs notably, depending on which datasets and features are used. These are broadly used

for document classification because they are effective classification algorithms with relatively good predictive performance [2]. In 2004, Pang and Lee classified documents in their study on movie reviews by negative or positive polarities using NB, Maximum Entropy (MaxEnt) and SVM to compare and specify which algorithm is more suitable regarding accuracy [21]. They concluded that machine learning techniques are considerably domain dependent due to words implying different emotions from one domain to another. They found that SVM algorithm was the best among the three classifiers tested. Gamon in 2004 conducted sentiment classification for a very noisy domain of customer review data from a global support service [22]. Proper classification accuracy was achieved by utilising large feature vectors in combination with feature reduction trained using SVM. Kenndey and Inkpen studied sentiment classification for movie reviews using SVM in order to classify reviews in terms of polarity [23]. It was indicated that a suitable accuracy was attained, and it was inferred that SVM is more suitable for sentiment analysis. In 2006, Chen et al. conducted a study in sentiment analysis for book reviews in 2006 performed different MLAs, namely decision tree, NB, and SVM, and they also concluded that SVM outperformed the others MLAs [24]. In 2011, Xia et al. conducted a comparative study of ensemble method was made for sentiment classification [25]. Three common text classification algorithms were employed, namely SVM, MaxEnt, and NB as a base classifier for each feature sets. They performed these algorithms on movie review documents that were already labelled in terms of polarity (positive and negative). It was concluded that SVM outperformed the others. Dayalani and Patil in 2014 also conducted a study on sentiment classification on Twitter data into three classes of polarity: positive, negative or neutral by using SVM algorithm [26]. These researches achieved reasonable results in their task for sentiment analysis. However, the limitation of their work is that they only classified tweets in terms of polarity. In 2015, Zhang et al. conducted a study using SVM for Mandarin comments on clothing products; they achieved superior performance in the experimental results of their sentiment classification [27]. Although SVM in most cases outperformed NB algorithm classifier, a study conducted in 2015 used NB for small a set of training data that outperformed SVM algorithm [13]. The limitation of this work is that a small data set was used, which cannot be relied upon for generalisation, while sentiment analysis works well with large amounts of data. In 2016, Tripathy and Agrawal used four different machine learning algorithms: NB, MaxEnt (ME), Stochastic Gradient Descent (SGD), and SVM to classify the tweets in terms of polarity for human sentiments [28].

Table I shows the summary of several types of research conducted using different machine learning techniques with their accuracies and data sources. In most comparative researches, SVM performed better than other machine learning techniques for polarity in sentiment classification. Correspondingly, another commonly used classification algorithm, NB, is a part of probabilistic classifiers based on applying Bayes theorem

with substantial assumptions between features [29]. In 2009, Melville et al. used NB for sentiment classification as it is considered one of the most common approaches in sentiment classification, as well as natural language processing [30]. The reason behind using NB broadly is that its assumption of word independence and the simplicity of this assumption causes the computation of NB to be much more efficient. Hence, NB is broadly used for document classification [31]. In 2014, Hennessy classified Twitter data using NB to determine whether tweets content reflects a negative or positive opinion on behalf of users [32]. The limitations of this work included that only one algorithm was used, and its accuracy was not reported. Secondly, a small data set of 1500 tweets was used, so it can be said that we cannot rely on this amount of data in order to generalise. Furthermore, MaxEnt is an MLA that generalises logistic regression to multiclass problems. The concept behind MaxEnt models is that one should prefer the most consistent model that satisfies a given constraint [33]. Bang et al. stated that from past studies, MaxEnt performed better than NB occasionally, but not continually, because the features of MaxEnt (unlike NB) make no independent assumptions [2]. This reveals that features such as bigram and phrases can be added to MaxEnt without features overlapping. Although in practice NB can perform better than MaxEnt on a variety of problems, MaxEnt theoretically outperforms NB because it is superior handling features for overlap.

B. Lexicon-Based Approach (LBA)

In 2004, Kamps et al. used LBA to perform sentiment classification in order to measure the semantic orientation of the text [34]. Zhu et al. later proposed a solution for sentences that are segmented, with every segment being assigned to one of the aspects found in the sentence [35]. Taboada et al. in 2011 proposed a method of sentiment classification using LBA whereby the polarity of each segment was decided whereby an aspect-polarity pair was created in order to reflect the complete polarity for this aspect within a particular review by using a sentiment LBA, using dictionaries of negative and positive polarised words related to their tasks [36]. These dictionaries build semantic orientation calculators by negation words and incorporating intensifiers. The overall average performance of 78.37% for reviews from various domains like camera and movie was achieved. Similarly, Ohana and Tierney performed LBA on the task of document-level sentiment classification utilising the polarity data set of film reviews [37]. The LBA in their paper was applied by counting negative and positive terms found in the document, then sentiment orientation was determined based on which class received the highest score. Although the major disadvantage of the dictionary-based LBA method is its incapability to find opinion words with domain and context particular orientations [38], Palanisamy in 2013 used such a method of LBA to extract sentiments from Twitter data [39]. Some recent studies have claimed that LBA is the most useful method to perform sentiment classification, because it calculates the sum of the number of polarity sentiment words appearing in the data in order to determine the

sentiment of the data [4]. Additionally, LBA has two methods, dictionary based approach (DBA) and corpus-based approach (CBA), both of which rely on sentiment lexicon to analyse the text [40], [41]. Firstly, the DBA relies on finding opinion seed words, and then searches the dictionary of their antonyms and synonyms [42]. Such methods usually use prebuilt dictionaries such as SentiWordNet, the standard dictionary for such purposes that defines the semantic orientation of words [43]. Secondly, this problem was solved by the CBA by starting with a seed list of opinion words [9], then moving to a larger corpus to find other opinion words to help in understanding context-specific orientations [44]. This method could be conducted by using semantic or statistical techniques [45]. The LBA has two major approaches:

1) Statistical Approach: The aim of statistical approach is to use statistical techniques by finding the co-occurrence words in order to achieve sentiment classification [46].

2) Semantic Approach: A semantic space is provided by the semantic approach to represent the terms for finding different type of semantic relationship between terms. Semantic spaces represent their words as high dimensional vectors achieved by statistical properties of word context [47].

IV. AN INVESTIGATIVE APPROACH

A. Collecting Data

Due to the increasing popularity of Apple products, especially the iPhone and iPad, they generate vast volumes of approval or criticism in users tweets. Tweets about these products cause large amounts of data to pass through Twitters data storages [18]. Accordingly, real time data was collected from Twitter on behalf of user tweets towards the iPhone and iPad from four cities, San Francisco, Los Angeles, New York and Chicago (using their longitude and latitude) for 24 hours (contemporaneously) during one week. The location was focused on the city where the tweets were released as a filtering mechanism to explore differences in users points of view about the iPhone and iPad relative to their location (i.e. city) in terms of their polarity and emotion that subsequently will be analysed and classified. The Twitter API was utilised to retrieve the data that needs authentication by Twitter. The Twitter API only strictly retrieves public tweets. Therefore, we avoided downloading any personal information such as date of birth and gender that users did not make explicitly public. Importantly, the retrieved data will be fully anonymised using cleaning processes. The researcher in charge of the data and the study also he will removed everything related to self-referencing by individuals, such as names and account identification from tweets captured based on location. The data has been being stored securely in the researchers encrypted hard drive according to the Universitys ethical regulations, subject to the provisions of the UK Data Protection Act (1998), and the researcher was the only person with access to the retrieved data.in this paper, tweets were collected only from public users of Twitter using streaming APIs [3]. The data collected from Twitter included the following attributes:

1) Text of the tweet.

TABLE I: Historical Review of Sentiment Classification

No	Authors	Algorithms	Contributions	Accuracy %
1	[22]	SVM,NB	Movie reviews (IMDb), 1000 (+) and 1000 (-) reviews	86.4-87.2
2	[23]	SVM	Customer reviews (feedback)	69.5-77.5
3	[24]	SVM	Movie reviews (IMDb), 1000 (+) and 1000 (-) reviews	80–85.9
4	[25]	Decision Trees C4.5, SVM, NB	Book reviews (Amazon), 3,168 reviews	84.59 (SVM has highest value)
5	[27]	SVM	Public opinion (Twitter)	Not reported
6	[28]	SVM	Mandarin comments on clothing products	90.30
7	[29]	SVM, NB, ME, Stochastic Gradient Descent (SGD)	Public opinion (Twitter)	Not reported
8	[30]	NB, ME, SVM	Movie reviews (IMDB), product reviews (Amazon)- 1000 (+) and 1000 (-) reviews	88.65
9	[32]	SVM, NB	Public opinion, certain keywords (Twitter)	Not reported

2) Favourite (true, false).

3) Favourite count (the number of times the tweet is favourite).

4) Created (date).

5) Twitter IDs.

6) Status source (URL of tweet).

7) Screen Name.

8) Was Retweeted (true, false).

9) Retweet count (the number of times the tweet is retweeted).

10) Location based on longitude and latitude .

The retrieved data based on the aforementioned locations was unstructured and all its attributes except the text of tweets was useless, because they will not be utilised in this paper. Therefore, data needed to be pre-processed in preparation for analysis and classification.

B. *Pre-Processing of Data*

The retrieved data contains a large amount of irrelevant and redundant information. Hence, several pre-processing steps will be performed in order to optimise the data for further experimentations. The pre-processing will be applied on the collected tweets from each city separately to analyse users behaviour in different locations. The aim of the data pre-processing was mainly to extract only the relevant content from the tweets while removing noise, meaningless symbols and all useless attributes in the data, The steps below were followed to clean data.

1) Removing retweets

2) Removing screen name

3) Removing (@) and (#) and nay irrelevant information such as True or False.

4) Removing punctuation

5) Removing numbers such as ID and Date.

6) Removing URLs (http).

7) Removing tabs and Blank space at the beginning and end of tweets.

8) Removing duplicate tweets.

After undertaking these steps, the tweets were cleaned and transformed to lower case to make uniform pattern. The cleaned tweets will be passed through a sequence of pre-processes that help in the conversion of the tweet strings into the feature vector of the Term Document Matrix (TDM). The TDM assists in gaining a better understanding of users textual sentiment and then generating ground truth for tweets in terms of polarity (positive and negative) and emotion (anger, disgust, fear, happiness, sadness and surprise). The pre-processing steps implemented are clarified below.

1) **Tokenization:** Tokenization describes the general process of splitting the text of a document into a series of tokens in order to identify all words in a given document for further processing, especially to create term document matrix. Table II shows that how a document was divided into individual words.

TABLE II: Tokenization of Text

a runner likes running			
a	runner	likes	running

2) **Stop words removal:** This step was performed to remove words that were occurring frequently and which may not contribute to the sentiment, such as prepositions, conjunctions and articles, resulting in the remaining data shown in table III.

TABLE III: Removing Stop Words

runner	likes	running	runs	bridge
runner	Like	run	run	bridge

3) **Stemming:** In this step, the tokens or words were reduced to their root form, as shown in Table IV.

TABLE IV: Stemming Words in The Text

runner	likes	running	runs	bridge
runner	Like	run	run	bridge

4) **Normalization:** Data is normalized for classification in order for it to be rescaled to the unit interval. Normalisation is important because without it the measure will be dominated by the largest scale variable. Table V shows the pros and cons of normalization.

TABLE V: : Pros and Cons of Normalization

Pros	Cons
Helps optimize machine learning algorithms	Results depend on chosen data normalization procedure
Normalization data is more accessible	It is not fit for data scales which have significance
Normalized data is less resource consuming to operate with. It improves data storing capacities and is not affected by delete or update oddities	Restrains when working with data validation

C. Generating Ground-Truth (Label)

The data collected from four cites concerning users opinions of the iPhone and iPad was pre-processed to generate ground truth in order to analyse users behaviour about these popular smart products based on their locations. Sentiment package was used after the data was successfully prepared in order to build data sets that include tweets with their features and ground-truth. This package contains two stages. First, it uses its lexicon dictionary to identify sentiment from tweets in terms of polarity and emotion. Regarding polarity classes, both positive and negative features are created, and also based on emotion classes, all six features are created for tweets. For instance, dictionary is used to identify all positive words in a given document, and then increment positive count is applied for each of them. Similarly, dictionary is used to detect all negative words in the tweets in order to increment negative count for each word. These processes enabled identification of both features (positive and negative) for polarity class. Using the same process for emotion, we produced six features: anger, disgust, fear, happiness, sadness and surprise. In this stage we could select features for data sets of both products, namely the emotion dataset, which has tweets with six features, and polarity dataset, which has tweets with two features. Accordingly, each product has a polarity dataset and an emotion dataset collected on behalf of users points of view from different locations. Second, NB and SVM were used within Sentiment package in order to generate ground truth based on the features selected for tweets. Both NB and SVM were labelled tweets in terms of polarity and emotion, and only these tweets were chosen to build the data sets when both algorithms were agreed about the label of tweets. For instance, tweets labelled positive by both NB and SVM contributed in the dates. In contrast, the tweets that attained different labels by NB and SVM were not contributing to the data sets, as can be seen in figure 1. In the second stage, tweets retrieved from various cities that reflect users assessment towards iPhone and iPad achieved ground truth and then 10,509 tweets were randomly selected from each city in order to have uniform data sets. Accordingly, data sets were fully prepared for analysis and then classification, to build data sets that comprise tweets with their features and ground truth.

D. Tweets Classification

Various well-known algorithms will be applied to obtain reasonable results for sentiment classification, including the supervised algorithms ZeroR, NB, SVM and K-NN, then the
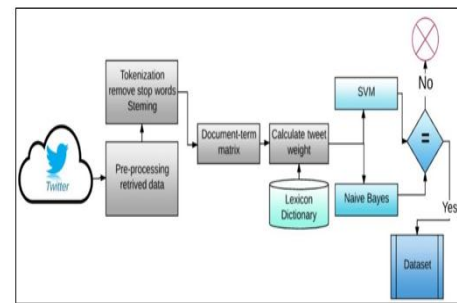

Fig. 1: Pros and Cons of Normalization

sentiment classification performance metrics accuracy, error rate, precision, sensitivity, and F-measure will be explained. The purpose of each of these algorithms is to build a model that accurately predicts the class of the unseen data, as shown in figure 2. Data classification normally contains two stages. First, the training set that is used to create the model includes a set of data that has pre-classified output and input variables. Second, the testing set that is used to evaluate how well the model performs with unseen data, which also comprises the pre-classified results data, which will not be used during running the test set data through the model until the end.

The pre-classified data will be compared with the results of model, then the model is adjusted in order to reduce error on the test set. As far as partitioning data is concerned, the techniques explained below will be used in order to attain classifier accuracy for be comparison and evaluation.
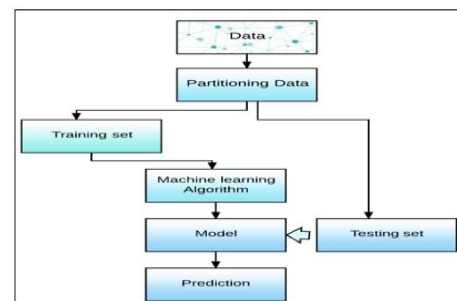

Fig. 2: Model for Prediction of New Instances

1) Separate Data by City: In this technique, data will be prepared for training and testing sets with the aim of the entire retrieved data to be participating in both training and testing sets. Therefore, the data collected from four cities will be separated based on the city for training and testing sets. Properly, the training set will be contained the data collected from three out of the four cities to build a model, whereas the last part of the data collected from a different city will be used for testing.

This process will be circulated until all the cities data will be involved in the building model and testing set, as shown in Figure 3. Consequently, the performance results that will be obtained based on applying model on the test set of each

city will be compared and evaluated. The purpose of this exercise is to identify the training data sets that could be employed to build a model capable of providing generalised explanations. This will enable us to make reliable predictions on the unobservant future data using the model.

Fig. 3: Separating Data Per City

2) Cross Validation: Cross validation will be used to develop and fine-tuning the classifier. It is a statistical technique for evaluating and comparing learning classification algorithms by estimating the performance of a classifier. The essential form of cross validation is known as k-fold cross validation. Data randomly will be divided into k-folds of the same size, as presented in Figure 4.

Segments of the available data as a training set will be used by K-fold cross validation to fit model and a different segment will be used as a testing set to test the model. This process will be circulated so that each point of the data has an opportunity of being validated. in this paper four-fold cross validation method will be used for experimentations, whereby the model will be trained using three folds, and tested using one fold. This process will repeated four times, thus all folds will be utilised for testing to determine performance computations and training.

Fig. 4: Partitioning Data Through Cross Validation

V. Experimental Results and Discussion

The ZeroR, NB, K-NN and Linear SVM using the cross validation technique have been used to classify the iPhone and iPad data sets according to the polarity and emotion. The used algorithms have been compared based on their trained time and achieved accuracy rates on the testing sets, as shown

in Table VI. It can be observed that the ZeroR has obtained the lowest accuracy rates 59.45% and 56.6% for the iPad and iPhone emotion data sets respectively. Besides, the lowest accuracy rates 58.53% and 67% has also attained by the ZeroR for the iPad and iPhone polarity data sets respectively. The reason behind this is that the ZeroR can only predict the majority class. Therefor only the prediction for the Joy class has been correctly conducted by the ZeroR because the majority instances in the emotion data set are labelled by the Joy class. Additionally the majority instances in the polarity data sets are Positive class so that only this class has properly predicted. On the other hand, the achieved accuracy rates for the emotion and polarity data sets using the NB are higher than the achieved accuracy rates using ZeroR because the achieved accuracy using the NB has considerably increased compared to the ZeroR. For example the accuracy using the ZeroR for the iPad emotion data set is 59.45%, whereas, the accuracy for the same data set using the NB is 93.73%. However the K-NN has achieved the higher accuracy rates 99.94%, 96.58%, 100% and 100% than the NB and ZeroR for the iPhone emotion data set, iPad emotion data set, iPhone polarity data and iPad polarity data sets respectively. The highest K-NN achieved accuracy rates could be related to the fewer features of the data sets because this algorithm is more suited for the lower dimensional data. Moreover, the K-NN achieved accuracy rate for the iPad emotion data set is the highest accuracy rate over the other algorithms accuracy rates. In contrast, it can be indicated that the highest accuracy rate 99.96% for the iPhone emotion data sets has been obtained by the SVM. Although the SVM accuracy rate (99.96%) slightly higher (by 0.02) than the K-NN accuracy rate (99.94%) for the iPhone emotion data sets, the SVM trained time (112.24 seconds) is extremely higher than the K-NN trained time (0.01 seconds) because the kernel parameter can make the SVM computationally expensive. Thus it could be said that the predicted iPhone emotion data set by the K-NN is practicable and its accuracy is reasonable. Regarding the polarity data sets, the same accuracy rates have been attained by the NB, K-NN and SVM. Therefor the highest accuracy rate has been determined by the minimum trained model time. This reason has led to specify the K-NN as a suitable algorithm for the polarity data set. With respect to the trained time, the NB method required more time than the ZeroR and K-NN methods. In addition, in spite of the fact that the linear SVM has a proper generalisation performance according to the data sets, the trained models using the linear SVM have required more time than the trained models by using the NB and K-NN. For instance, the trained time using the SVM for the iPhone polarity needed 70.6 seconds, whereas the trained time for the same data set using the K-NN needed 0.02 seconds. The reason behind this could be related to the fact the K-NN has no explicit training phase or it is quite minimal. Hence, the training phase is considerably swift because the K-NN needs the training data instances during the testing phase.Notably, using the cross validation technique, it could be said that the K-NN is the appropriate and reasonable algorithm with respect

to the data sets. Consequently, because of having the proper generalisation performance, the models could be used for the future data of a different products in terms of the polarity and emotion. Regarding the generalizing models across cites, the ZeroR, NB, K-NN, and SVM have been used to classify the data sets based on the partitioning data per city in terms of the polarity and emotion. Regarding the polarity, the all algorithm models have achieved the highest accuracy rates when they have been applied on the San Francisco testing sets compared to the all other cities testing sets for the polarity data sets. After that, the accuracy rates of the algorithms have been compared to specify which algorithm has achieved the reasonable accuracy rate. In order to identify the suitable algorithm the trained time has also been considered as shown in the Table VII. In addition, this table also shows that the training data, testing data, and parameters of each algorithm. It can be seen that the ZeroR has achieved the lowest accuracy rates 59.72% and 70.63% for the iPad and iPhone polarity data sets respectively. Since the ZeroR can predict the majority category properly, only the major Positive class has correctly classified. In contrast, the highest accuracy rate (100%) has been achieved by the NB, K-NN, and SVM for the iPad and iPhone polarity data sets. Since the NB, K-NN, and SVM have the same accuracy rate, the minimum built model time have been considered in order to identify the most proper algorithm. It can be seen that the trained model time in SVM is considerably higher than the trained model time in the NB and K-NN. The reason behind this is due to the fact that the kernels make SVM computationally expensive. Moreover, The Nave Bayes model time is slightly longer than the trained time using the K-NN. Thus it is appeared that the trained models time in the K-NN has required shorter time because the generalization will not be performed on the training data. As a result, the all training data will be used in the testing phase. In addition, the K-NN using Manhattan distance has been specified as a proper algorithm for polarity data sets when the k value is equal to 10 for the iPad data set and the k value is equal to 9 for the iPhone data set. These are signifying the New York, Los Angeles, and Chicago data that are used for the training data are proper examples data for creating the models in order to be generalized. In addition, since the reasonable accuracy rates have been achieved by the applied models on the San Francisco unseen data, the model can be applied on the other cities data for different products.

Regarding the emotion data sets, table VIII shows that the ZeroR and NB have achieved the highest accuracy when they have been applied on the San Francisco testing data compared to the other cities testing sets. Moreover when K-NN has been applied on the all cites testing sets, the highest accuracy (98.8%) has been obtained for the iPad emotion data set when the San Francisco data was used as a testing set. In addition, the highest accuracy (99.95%) has been obtained for the iPhone emotion data set when the Chicago data was used as a testing set. Furthermore, when the Linear SVM has been applied on the all cites testing sets, the highest accuracy (96.79%) has been obtained for the iPad emotion

data set when the Los Angeles data was used as a testing set. Additionally, the highest accuracy (99.94%) has been obtained for the iPhone emotion data set when the New York data was used as a testing set. Table VIII also shows the built model time and parameters of each algorithm. It can be seen that the K-NN requires the minimum built model time and it also achieved the highest accuracy rates 98.8% and 99.95% for the iPad and iPhone emotion data sets respectively. Table VIII also shows the K-NN using Manhattan distance has achieved the highest accuracy rate when the k value is equal to 24 for the iPad emotion data set and when the k value is equal to 23 for the iPhone emotion data sets. Consequently, it can be said that the achieved models via the K-NN have a proper generalisation performance and they could be used for the other cities data towards different products in terms of the polarity and emotion.

VI. CONCLUSION AND FUTURE WORK

Sentiment classification and analysis is a developing field which has various applications such as classifying movie reviews and user comment towards different products, events, and honoured individuals. Despite of the fact that the sentiment classification and analysis task is challenging because it is a part of natural language processing, sentiment classification has been quickly progressed in the last few years because it has highly been demanded. Recently not only the companies are trying to realize how their services or new products are perceived by users or customers, but also users or customers are attempting to understand the opinions of the other users before deciding to purchase an item. The increasing requirement for new products insight, the technical challenges and development various social networking services currently facing the field which will retain sentiment classification and analysis relevant for predictable future. The future generation sentiment classification and analysis systems require a deeper bind between reasoning approaches inspired by scientist people thought and psychology with all knowledge bases. This will help us to gain better comprehension of the natural language opinions and will more properly bridge the gap between unstructured data that expresses of people opinions and structured data that can be processed, analysed, and evaluated by a machine. In this paper, the data was retrieved from Twitter about iPhone and iPad in certain locations: Chicago, Los Angeles, San Francisco and New York. The data then was passed through a sequence of pre-processing and cleaning in order to extract relevant data which was the cleaned tweets. The prepared tweets were passed to further process using domain relevant packages in order to generate ground truth for tweets. After the tweets were classed in terms of polarity (positive and negative) and emotion (anger, joy, sadness, disgust, fear, and surprise) the data sets were built which are: iPad emotion and iPhone emotion data sets, iPad polarity, and iPhone polarity data sets. The different locations were focused to collect the data in order to generalise the models which will be conducted across cities. The built data sets were experimented using the machine learning algorithms

TABLE VI: Accuracy and Training Time of The Algorithms

Algorithms and Training time (second)	Data sets			
	Emotion data sets		Polarity data sets	
	iPad	iPhone	iPad	iPhone
ZeroR	59.45 %	56.6 %	58.53%	67 %
Training time	0.02	0.01	0	0
Naïve Bayes (NB)	93.73 %	99%	99.9 %	99.97 %
Training time	0.03	0.05	0.03	0.03
K-NN with Manhattan	96.58 %	99.94 %	100%	100%
Training time	0.01	0.01	0.03	0.02
Linear SVM	96.37%	99.96%	100%	100%
Training time	254.62	112.24	219.93	70.62

TABLE VII: Accuracy, Training Built Model, Time and Parameter of Algorithms for The Emotion Dataset

Algorithms	Training Dataset for Building Model	Parameter	Time	Testing Dataset	iPad Polarity	iPhone polarity
ZeroR	New York, Los Angeles, Chicago	-	0	San Francisco	59.72%	70.63%
Naïve Bayes	New York, Los Angeles, Chicago	-	0.02	San Francisco	100%	100%
KNN	New York, Chicago, San Francisco	Manhattan, K(10)	0.001	San Francisco	100%	
KNN	Los Angeles, Chicago, San Francisco	Manhattan, K(9)	0.001	San Francisco		100%
Linear SVM	New York, Los Angeles, Chicago	C(0.001)	35	San Francisco	100%	-
Linear SVM	Los Angeles, Chicago, San Francisco	C (0.001)	36	San Francisco	-	100

TABLE VIII: Emotion Accuracy and Training Built Model

Algorithm	Training set data for building model	parameter	Time	Testing set data	iPad Emotion	iPhone Emotion
ZeroR	New York, Los Angeles, Chicago	-	0	San Francisco	61.62%	56.71 %
Naïve Bayes	New York, Los Angeles, Chicago	-	0.06	San Francisco	97.75%	-
Naïve Bayes	New York, Los Angeles, Chicago data	-	0.05	San Francisco	-	99.77 %
KNN	New York, Chicago, San Francisco	Manhattan, K (24)	0.01	San Francisco	98.8%	
KNN	Los Angeles, Chicago, San Francisco	Manhattan, K (23)	0.01	Chicago		99.95%
SVM	New York, Los Angeles, Chicago	Linear , C(15)	39.1	Los Angeles	96.79%	
SVM	Los Angeles, Chicago, San Francisco	Linear , C(0.001)	46	New York		99.94

with cross validation technique and partitioning the data across city in order to predict the future tweets to obtain users opinion. Regarding randomly splitting data, four folds cross validation used for the all data sets. It was found that the models have a proper generalisation performance which can lead to be performed on the different products. Specifically, it was found that the K-NN using Manhattan distance has achieved the highest accuracy rates 96.58 % and 99.94 % for iPad and iPhone emotion data sets by using the cross validation technique. Regarding partitioning the data per city, it was found that the NB, K-NN, and SVM have achieved reasonable accuracy mostly when their models applied on the San Francisco testing set data. This signifies that the used data for training models are proper example for building models in order the models to be generalised. In addition, the KNN was also outperformed the NB and SVM because of two main reasons, firstly, the KNN has minimum trained time to build model. Secondly, with using KNN, the highest accuracy rates 98.8% and 99.95% were achieved for the iPad and iPhone emotion data sets respectively. Regarding polarity data sets using both cross validation and partitioning data per city the K-NN achieved 100% for all the polarity data sets.The retrieved data from Twitter was nearly enough to demonstrate a proper generalisation performance. However the lower dimensional data compared to the number of the observations is the main limitation of this research. This limitation was led to achieve a quite high accuracy particularly for the polarity data sets. In future, several works are required to improve the evaluation measures on Twitter or reviews data such as implementing fuzzy string matching technique to remove similar documents or tweets in a percentage after removing duplicate tweets. Removing similar tweets in a percentage will help researcher to build data sets more properly. Moreover, adding more features for the data sets especially for the polarity data set such as neutral very positive, very negative. In addition, sentiment classification and analysis could be performed for new applications. In spite of the fact that the techniques and machine learning algorithms are used for Sentiment classification and analysis are advancing proper, however, there are many challenges or problems in this field of research remain unsolved. The main problem and challenge aspects remain ongoing for the sentiment classification and analysis of other languages, handling of implicit attributes of products, building summary of opinions based on attributes of products and dealing with negation opinion expressions. Further research to these challenges could be conducted.

REFERENCES

[1] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower

graph," in *Proceedings of the First workshop on Unsupervised Learning in NLP*. Association for Computational Linguistics, 2011, pp. 53–63.

[2] 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.

[3] D. Reagan, *Twitter application development for dummies*. John Wiley and Sons, 2010.

[4] Y. Wan and Q. Gao, "An ensemble sentiment classification system of twitter data for airline services analysis," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 1318–1325.

[5] V. Pandey and C. Iyer, "Sentiment analysis of microblogs," *CS 229: Machine learning final projects*, 2009.

[6] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 417–424.

[7] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[8] R. Varghese and M. Jayasree, "A survey on sentiment analysis and opinion mining," *IJRET: International Journal of Research in Engineering and Technology eISSN*, vol. 23191163, 2013.

[9] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[10] Z. C. A. M. B. Liu, "Aspect extraction with automated prior knowledge learning," 2014.

[11] D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," in *Extended Semantic Web Conference*. Springer, 2011, pp. 88–99.

[12] T. Segaran, *Programming collective intelligence: building smart web 2.0 applications*. " O'Reilly Media, Inc.", 2007.

[13] S. K. Yadav, "Sentiment analysis and classification: a survey," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 3, no. 3, pp. 113–121, 2015.

[14] T. Zagibalov and J. Carroll, "Automatic seed word selection for unsupervised sentiment classification of chinese text," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008, pp. 1073–1080.

[15] J. Rothfels and J. Tibshirani, "Unsupervised sentiment classification of english movie reviews using automatic selection of positive and negative sentiment items," *CS224N-Final Project*, 2010.

[16] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Clustering product features for opinion mining," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 347–354.

[17] V. K. Singh, M. Mukherjee, G. K. Mehta, S. Garg, and N. Tiwari, "Opinion mining from weblogs and its relevance for socio-political research," in *International Conference on Computer Science and Information Technology*. Springer, 2012, pp. 134–145.

[18] S. A. A. Hridoy, M. T. Ekram, M. S. Islam, F. Ahmed, and R. M. Rahman, "Localized twitter opinion mining using sentiment analysis," *Decision Analytics*, vol. 2, no. 1, p. 1, 2015.

[19] E. Divya, "Real time sentiment classification using unsupervised reviews," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2.

[20] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, "Sentiment analysis techniques in recent works," in *Science and Information Conference (SAI), 2015*. IEEE, 2015, pp. 288–291.

[21] 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*. Association for Computational Linguistics, 2004, p. 271.

[22] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004, p. 841.

[23] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational intelligence*, vol. 22, no. 2, pp. 110–125, 2006.

[24] C. Chen, F. Ibekwe-SanJuan, E. SanJuan, and C. Weaver, "Visual analysis of conflicting opinions," in *2006 IEEE Symposium On Visual Analytics Science And Technology*. IEEE, 2006, pp. 59–66.

[25] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138–1152, 2011.

[26] G. G. Dayalani and B. Patil, "Emoticon-based unsupervised sentiment classifier for polarity analysis in tweets," *International Journal of Engineering Research and General Science*, vol. 2, no. 6, 2014.

[27] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and svm perf," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

[28] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117–126, 2016.

[29] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," *International Journal*, vol. 2, no. 6, 2012.

[30] 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.

[31] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2622–2629, 2008.

[32] A. Hennessy, "Sentiment analysis of twitter: Using knowledge based and machine learning techniques," Ph.D. dissertation, Dublin, National College of Ireland, 2014.

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

[34] J. Kamps, M. Marx, R. J. Mokken, M. d. Rijke *et al.*, "Using wordnet to measure semantic orientations of adjectives," 2004.

[35] J. Zhu, H. Wang, B. K. Tsou, and M. Zhu, "Multi-aspect opinion polling from textual reviews," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 1799–1802.

[36] 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.

[37] B. Ohana and B. Tierney, "Sentiment classification of reviews using sentiwordnet," in *9th. IT & T Conference*, 2009, p. 13.

[38] G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu, and C. Chen, "Dasa: dissatisfaction-oriented advertising based on sentiment analysis," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6182–6191, 2010.

[39] P. Palanisamy, V. Yadav, and H. Elchuri, "Serendio: Simple and practical lexicon based approach to sentiment analysis," in *proceedings of Second Joint Conference on Lexical and Computational Semantics*. Citeseer, 2013, pp. 543–548.

[40] S. Gao, J. Hao, and Y. Fu, "The application and comparison of web services for sentiment analysis in tourism," in *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*. IEEE, 2015, pp. 1–6.

[41] M. Kanakaraj and R. M. R. Guddeti, "Nlp based sentiment analysis on twitter data using ensemble classifiers," in *Signal Processing, Communication and Networking (ICSCN), 2015 3rd International Conference on*. IEEE, 2015, pp. 1–5.

[42] S. Kalarikkal and P. Remya, "Sentiment analysis and dataset collection: A comparitive study," in *Advance Computing Conference (IACC), 2015 IEEE International*. IEEE, 2015, pp. 519–524.

[43] S. Y. Ganeshbhai and B. K. Shah, "Feature based opinion mining: A survey," in *Advance Computing Conference (IACC), 2015 IEEE International*. IEEE, 2015, pp. 919–923.

[44] U. C. A. KUMAR, P. JAISWAL, "Comparative study on sentiment analysis and opinion mining." IEEE.

[45] A. Balahur, R. Mihalcea, and A. Montoyo, "Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications," *Computer Speech & Language*, vol. 28, no. 1, pp. 1–6, 2014.

[46] P. Stalidis, M. Giatsoglou, K. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "Machine learning sentiment prediction based on hybrid document representation," *arXiv preprint arXiv:1511.09107*, 2015.

[47] M. W. Berry, A. H. Mohamed, and B. W. Yap, *Soft Computing in Data Science: First International Conference, SCDS 2015, Putrajaya, Malaysia, September 2-3, 2015, Proceedings*. Springer, 2015, vol. 545.