A Prediction of South African Public Twitter Opinion using a Hybrid Sentiment Analysis Approach

Matthew Brett Shackleford[1], Timothy Temitope Adeliyi[2], Seena Joseph[3]*

Department of Information Technology, Durban University of Technology, Durban, South Africa[1, 3]

Engineering, Built Environment, and Information Technology, University of Pretoria, Pretoria, South Africa[2]

Abstract—Sentiment analysis, a subfield of Natural Language Processing, has garnered a great deal of attention within the research community. To date, numerous sentiment analysis approaches have been adopted and developed by researchers to suit a variety of application scenarios. This consistent adaptation has allowed for the optimal extraction of the authors emotional intent within text. A contributing factor to the growth in application scenarios is the mass adoption of social media platforms and the bondless topics of discussion they hold. For government, organizations and other miscellaneous parties, these opinions hold vital insight into public mindset, welfare, and intent. Successful utilization of these insights could lead to better methods of addressing said public, and in turn, could improve the overall state of public well-being. In this study, a framework using a hybrid sentiment analysis approach was developed. Various amalgamations were created – consisting of a simplified version of the Valence Aware Dictionary and sEntiment Reasoner (VADER) lexicon and multiple instances of classical machine learning algorithms. In this study, a total of 67,585 public opinion-oriented Tweets created in 2020 applicable to the South African (ZA) domain were analyzed. The developed hybrid sentiment analysis approaches were compared against one another using well known performance metrics. The results concluded that the hybrid approach of the simplified VADER lexicon and the Medium Gaussian Support Vector Machine (MGSVM) algorithm outperformed the other seven hybrid algorithms. The Twitter dataset utilized serves to demonstrate model capability, specifically within the ZA context.

Keywords—*Sentiment analysis; opinion mining; machine learning; government; public service delivery; twitter*

I. INTRODUCTION

Communication has become more dynamic due to the widespread adoption of both mobile devices and the Internet; this has allowed for people to express themselves regardless of location. In both the public and private sectors, organizations have either begun or currently leverage technical methods of analyzing public opinion [1-4]. Governments and organizations deemed the assessment necessary to gain a full understanding of the emotional state of a specific group in relation to that of their own performance [5].

Sentiment analysis or alternatively referred to as opinion mining, is a branch of Natural Language Processing that uses a variety of methods to determine an author's positive, negative, and neutral emotional stance from snippets of their writing [6]. Typically, analysis takes place at an aspect, sentence, or document level, however a combination of levels is an additional possibility and at times - can offer a more comprehensive perspective. Successful sentiment analysis would be the precise extraction of the emotions surrounding a viewpoint within the context in which correct interpretation was intended by the author [7].

A significant proportion of the sentiment analysis attempts in the South African (ZA) context have been geared toward specific events or themes, such as gauging perspectives on ethnic division [8], sexual violence [9], and the #FeesMustFall campaign [10]. This focus on judging instances ignores the importance of attaining a multidisciplinary model qualified to assess the emotional state of the ZA population, throughout a multitude of varying scenarios.

Sims determined in [11] that losing track of the public's opinion or neglecting a society's overall emotional state can lead to unfavorable public behavior, such as protests, which has a multifaceted and detrimental influence on the country. One can conclude that an accurate means of analyzing the Tweets of ZA public perception, would result in the emotional state of the public being handled in a more proactive and effective manner. Furthermore, such a model might be used to detect future communication patterns that may fluctuate the emotional state of the targeted demographic (positively or negatively). As more people around the world utilize social media to communicate their thoughts and feelings, the necessity to iteratively construct a better system of assessing an author's sentiment, using various sentiment analysis approaches that are suitable for the intended provided circumstances, becomes a challenge [12].

This study serves as an example that the created model may additionally work within the context of other countries, as the Valence Aware Dictionary and sEntiment Reasoner (VADER) lexicon is not specific to the ZA domain in which it succeeded - but rather likely the American domain from whence it came. The testing and creation of a novel sentiment analysis hybrid geared toward social media is highly beneficial, as it serves globally as an option for social media text analysis. Additionally, it was created with the hopes of effectively analyzing social media within the ZA context. As mentioned later in the related works section, hybrids have the tendency of outperforming standalone lexicon and machine learning models. As suggested, a social media hybrid model such as the hybrid in this study might be better suited to reanalyze the #FeesMustFall dataset [10].

The layout hereafter is as follows. A brief overview of the literature in Section II. Section III focuses on: the VADER and Twitter dataset; hardware and software used; study

architecture; planned approach; data preparation and processing; machine learning model parameters; evaluation criteria; as well as Twitter dataset preparation, processing, and evaluation. Section IV focuses on an analysis of the results. Section V consists of speculation over the results. Limitations of the study are highlighted in Section VI. The concluding remarks are found in Section VII. Future research is briefly discussed in Section VII.

II. RELATED WORK

There are four supervised classification algorithm types that contain models commonly used for sentiment analysis, according to Birjali et al., [13]. The first is a Linear approach, a statistical method for categorizing sentiment using hyperplane or linear decision boundaries. The linear approach ultimately outputs the most probable class (either negative or positive within the sentiment analysis domain) of a prescribed input. Secondly and in contrast to the linear technique, the probabilistic classifier – typically founded on Bayes' theorem – works by predicting a probability distribution over a set of classes. The third approach is rule-based classification, which refers to any classification scheme that uses IF-THEN rules to predict class membership. As a result, a set of rules ultimately guides the classifier in this technique to accomplish sentiment classification. The fourth option is the decision tree strategy, which decomposes the training data space hierarchically using an attribute value condition to classify input data into a limited quantity of predetermined classes. Additionally, the ensemble approach exists as a varied combination of the above approaches. The fundamental proposition behind this idea is to form an ensemble of different classifiers that outperform their standalone examples.

The authors in [14,15] attempted to apply and compare a variety of machine learning models, however, their results varied depending on the methodologies and datasets employed. This is mainly because a machine learning model that performs well in one trial does not necessarily guarantee equivalent performance in subsequent studies, due to varying circumstance. As results differ, models need to be tested separately under exact circumstances. Traditional machine learning approaches use a single learner method; however, an ensemble approach uses numerous learner methods to better leverage each learner's unique strengths while also covering any shortcomings that may exist, resulting in increased model precision and reliability [16,17]. When opposed to using a single classifier model, the most significant disadvantage of ensemble techniques is the additional processing time and power required [13].

Bagging, boosting, stacking and voting are the ensemble approaches leveraged to better the results of traditional standalone machine learning models [15, 18]. Bagging is a bootstrap aggregating prediction ensemble machine learning technique. Boosting ensemble is a strategy for training a set of weak classifiers that were previously poorly trained [19]. When presented with noisy data, such as unstructured Twitter text, boosting is more sensitive, while bagging is more resistant [20]. Despite this, boosting strategies have been utilized in several studies to increase model performance [21, 22]. Stacking is an ensemble model where various diverse methods

are devised using training data [18]. The use of the ensemble machine learning method to analyze sentiment is limited [23]. This additionally holds true with the analysis of ZA citizen text via digital platforms.

An ensemble machine learning strategy that focuses solely on learning classifier combinations is successful in multiple investigations [15, 24]. To obtain a final prediction, the voting ensemble approach uses pre-processing sentiment analysis, feeding several algorithms, and then integrating the results into a voting average method. Other models may use various components at different stages while still adhering to the same process. An ensemble method in the literature performed well throughout the evaluation when different classical machine learning algorithms were placed via a vote on average probability approach.

Pre-trained word embeddings are created using Bidirectional Encoder Representations from Transformers (BERT), an ensemble of binary classifiers [25]. The feature selection and feature extraction processes used in sentiment analysis pre-processing are not required because this approach takes tokenization as input [26]. As a result, researchers in [25] observed that a combination of BERT supplied pre-trained word embeddings along with the Random Forest classifier led to an accuracy of 94%, exceeding current models at the time of publication. The study in [26] achieved comparable success when creating a BERT - Support Vector Machine hybrid model, attaining 94% accuracy.

Hybrid sentiment analysis models typically exist as a combination of machine learning and lexical approaches. Hybrid sentiment analysis models are popular as an alternative for sentiment analysis as they combine the versatility of machine learning algorithms with the superior performance of lexicons [27, 28]. When compared to the state-of-the-art machine learning and lexicon approaches available, Abd El-Jawad et al [14] discovered that hybrid models tended to outperform the latter. As a result, it is concluded that hybrid sentiment analysis approaches would seem to offer top results.

One of the more widely used supervised machine learning techniques within the sentiment analysis domain is the Support Vector Machine (SVM) algorithm [29]. While the original SVM model uses a linear hyperplane, non-linear hyperplane models are often accommodated using a kernel method such as the Gaussian function. The Medium Gaussian Support Vector Machine (MGSVM) deals well with data of medium complexity and adopt the kernel scale $\sqrt{P}$ – where the value P refers to the dimension size or the number of features of the vector x_i [30, 31].

Many lexical approaches exist such as the manual, dictionary-based, corpus-based, and statistical approaches to name a few, and within these approaches lay a variety of lexicons [13]. One lexical approach that has proven to be a valuable outlier within the social media sentiment analysis domain is the VADER lexicon. The VADER lexicon is tuned towards assessing contents like that of a microblog and has particularly succeeded in the social media domain [32].

Based on the advantages of both the MGSVM and VADER approaches as well as the hybrid approaches tendency to

outperform standalone models, a gap in the literature was found where a hybrid of the two was yet to be created. This study sought to create a hybrid of the two as well as various other hybrid approaches, to compare and establish an effective means of assessing ZA public Twitter opinion.

III. Materials and Method

The VADER lexicon contains a lexical corpus that holds a variety of words and their associated negative or positive numerical weighting depending on the keyword's perception and has been empirically validated by various individuals [32]. The VADER lexical dataset was obtained and is freely accessible via GitHub [32]. Based on the keyword sentiment weighting, the July 2019 version of the VADER lexical dataset contains 7517 keywords, 3344 of which are classed as positive and 4173 as negative. The hybrid model was also evaluated on a dataset of ZA public Tweets from the public repository site Kaggle. The experiment was performed using MATLAB R2021b on a Windows 10 Professional machine.

The study architecture is depicted in Fig. 1. The hybrid model creation and selection process takes place in the first stage, whereas the capability of the chosen hybrid model is tested on a ZA public dataset in the second stage. In this study,

a hybrid strategy with two stages is developed to overcome the shortcomings of the separate techniques.

A. Stage 1 – Creating a Top-Performing Hybrid Model:

Five-fold cross-validation was used to train each model. The use of this validation method minimizes the effects of sampling bias. This is done through the random division of the lexical dataset – into five equal sections. Four of the five equal sections are used to train the algorithm, with the remaining section set aside to test the algorithm. This is repeatedly done until all combinations of the test/training set have been exhausted and, the resultant model produced is the mean of the training iterations [33]. Additionally, this procedure was opted for as it prevents both overfitting and underfitting to provide a more exact result.

Simplified VADER lexicon – The original VADER lexicon is acquired, and the keywords are assigned polarity, either positive or negative depending on their allocated numerical rating. The simplified VADER dataset is utilized to train various classical machine learning models. These hybridized models are evaluated utilizing performance evaluation measures at a later stage.
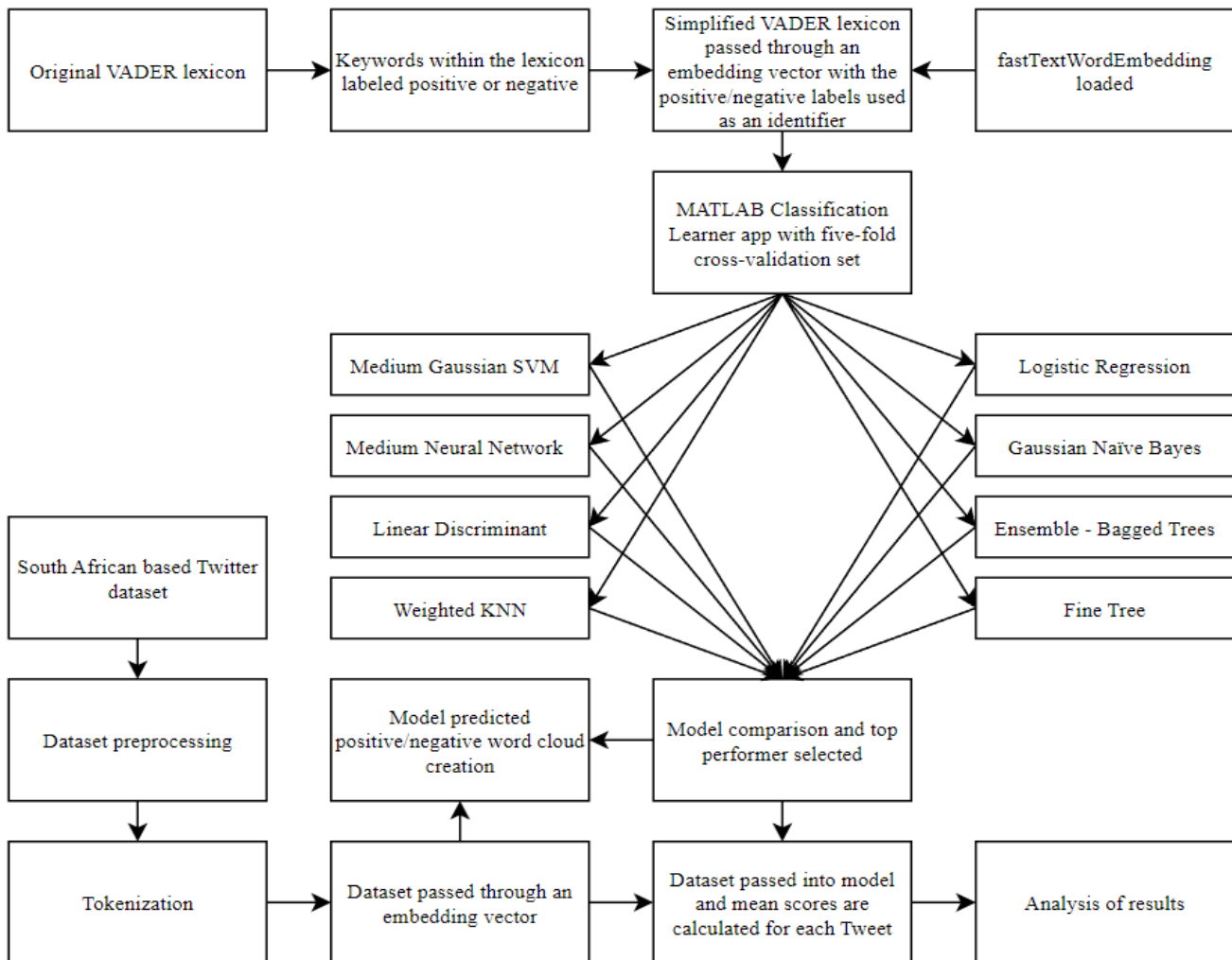

Fig. 1. The study architecture.

Vectorizing the simplified VADER Lexicon - A subfunction of the fastTextWordEmbedding function in MATLAB, word2vec is obtained via the "text analytics toolbox". The need to pass the simplified lexical corpus through word vectoring originates from distributed representation; the function assists in passing the words into numerical vectors [34], which assist the algorithm to better correlate words with their planned outcome.

Reconnecting the sentiment and the lexical vector - The linked sentiment is substituted as the vector's "predictor" after the lexical keywords have been vectorized.

Learner application for MATLAB classification - To check the generalization capacity of predictive models and avoid overfitting, the transformed lexicon is entered into the MATLAB classification learner program, and the validating method is set to five-fold cross-validation.

Training the machine learning models - The various hybrid models are created from a selection of classical models from the MATLAB classifier learner program. The chosen classical models are: Fine tree (FT), Medium Neural Network (MNN), Logistic Regression (LR), Gaussian Naive Bayes (GNB), MGSVM, Linear Discriminant (LD), Weighted K Nearest Neighbor (WKNN) and Ensemble-Bagged Trees (EBT). The models were programmed to make use of parallel processing. Table I highlights the parameters used to train each of the models. Additionally, hyperparameter options and principal component analysis was disabled for all the chosen models. The various models ran on default parameters, assigned to each model based on the way in which MATLAB interpreted the training dataset.

Having trained the various classical models on the VADER lexicon, the various hybrid models are compared using the following performance metrics: Validation confusion matrix (VCM), Accuracy, Recall, Precision, Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC), F1-score, training time, and prediction speed.

B. Stage 2 - An Overview of how the Hybrid Model was Utilized to Process the ZA Public Twitter Dataset:

Obtaining a Twitter based ZA public opinion dataset - The dataset of ZA public Twitter opinion is regarded as crucial for carrying out the experiment and creating a suitable environment. Because the presence of any manipulation could skew the results and compromise the authenticity of the simulated environment, the dataset must include the raw text from users' Tweets. This dataset might be retrieved through a selection of internet-based data repository sources, for instance "Kaggle," or it could be retrieved by leveraging the Twitter research Application programming interface (API) through an authorization process. As the researcher was unable to secure Twitter research API access, the dataset used in this article was obtained through Kaggle.

Data preparation - This part of the research entails processing the dataset into a suitable state for the trained hybrid model to process and output an authentic result.

TABLE I. MATLAB CLASSIFICATION LEARNER MODEL PARAMETERS

MATLAB Preset	Other
Medium Gaussian SVM (MGSVM)	Kernal function: Gaussian Kernal scale: 17 Box constraint level: 1 Multiclass method: One-vs-One Standardize data: true Cost matrix: default
Medium Neural Network	Number of fully connected layers: 1 First layer size: 25 Activation: ReLU Iteration limit: 1000 Regularization strength (λ): 0 Standardize data: yes
Linear Discriminant (LD)	Covariance structure: Full Cost matrix: default
Weighted KNN (WKNN)	Number of neighbors: 10 Distance metric: Euclidean Distance weight: squared inverse Standardize data: true Cost matrix: default
Logistic Regression (LG)	None
Naive Bayes (GNB)	Distribution name for numeric predictors: Gaussian Distribution name for categorical predictors: N/A Cost matrix: default
Bagged Trees (EBT)	Ensemble method: bag Learner type: Decision tree Maximum number of splits: 6092 Number of learners: 30 Cost matrix: default
Fine Tree (FT)	Maximum number of splits: 100 Split criterion: Gini's diversity index Surrogate decision splits: off Cost matrix: default

- Removing columns - The Kaggle sourced Twitter dataset holds columns such as - "Tweet author", "Tweet created at", "Tweet coord", "Tweet favorite count", "Tweet hashtag", "Tweet retweet count", "Tweet place", and "Unique ID" - that are deemed extraneous. All irrelevant columns are eliminated, leaving the column "tweet text," which contains the textual data required.

- Erasing hexadecimal Unicode - The Kaggle obtained dataset was originally extracted by the author (Mbuso Makitla) utilizing the Tweepy API. Due to decryption limitations and time constraints, the MATLAB function "regexprep(str,expression,replace)" is used to remove the Unicode, with the "expression" and "replace" values of '\\[a-z0-9]3}' and ' respectively'.

- URL removal - To exclude URLs from the Twitter text, the erase URL (str) function in MATLAB is used. These occurrences usually happened because of Tweet authors retweeting or linking to other websites within their Tweets.

- Stop word elimination – According to [5] common phrases like "an", "it", and "the" are removed. The function removeStopWords(documents) in MATLAB is used to accomplish this. This is necessary to mitigate the consequences of noisy data.

- Normalization - This phase involves removing punctuation using MATLAB's erasePunctuation(str)

function, furthermore lower(str) is used to transform all Tweet text to lower case. The goal of normalization in the preprocessing stage is to increase text homogeneity [35].

- Tokenization - This is used to break down Tweets into smaller text samples for sentiment analysis; word tokenization is leveraged as the text is evaluated on a per word basis.

Using word2vec to transform the preprocessed ZA Twitter dataset – To produce distributed word representations, and to match the database to the training dataset, word2vec was used on the preprocessed ZA Twitter dataset.

Word cloud evaluation - This visual tool is leveraged to roughly assess the classificational ability of the hybrid model. The word cloud outputs assumptions for those words on which the hybrid model was trained, and new words the hybrid model has yet to encounter. The compilation of the word cloud is based off words contained in the ZA Twitter dataset.

Hybrid model implementation - The hybrid classifier is used once again to run the preprocessed ZA Twitter dataset. The classifier calculates the sentiment of each word in the Tweet and the segment of code thereafter assigns a mean score to the entire text sample. The outcome is responsible for determining the polarity of the text, negative values equate to negative sentiment as does a positive value indicate positive sentiment. Naturally the closer the sample is to zero the more neutral the Tweet text.

Sentiment analysis system evaluation – A function in MATLAB is leveraged to record the duration required to run the preprocessed ZA Twitter dataset through the hybrid model. Additionally, the final estimated outcome of the hybrid classifier and original Tweet is to lightly undergo a comparative observational analysis. This comparative analysis helps to identify instances where the hybrid model may have succeeded, misjudged, or failed to appropriately determine the sentiment of the Twitter text.

IV. RESULT ANALYSIS

A. A Comparison of the Various Created Hybrid Models:

The True Positive (TP) value serves as an indicator of the successful positive predictions made by the various hybrid models, as indicated in Fig. 2. The leading hybrid model (MGSVM) surpasses the runner-up (MNN) by a margin of 64 predictions. The False Positive (FP) value denotes how many positive samples the various hybrid models predicted incorrectly, as indicated in Fig. 2. The leading hybrid model (MGSVM) predicts more accurately than the runner-up (EBT) by a margin of 58 less incorrect predictions. The False Negative (FN) value indicates how many negative samples the various hybrid models predicted incorrectly, as indicated in Fig. 2 The MGSVM hybrid model makes fewer prediction errors than the runner-up (MNN) by a margin of 64 less inaccurate predictions. The True Negative (TN) value serves as an indicator of the successful negative predictions made by the various hybrid models, as indicated in Fig. 2. Where the MGSVM hybrid model surpasses the runner-up (EBT) by a margin of 58 predictions. In terms of VCM performance, the

MGSVM hybrid model is undoubtedly the winning hybrid model – as it outshone the remaining hybrid models in all four VCM categories. However, it is important to note that in the runner-up category, MNN (second best for TP and FN) was bested by EBT in the FP and TN categories FN, however EBT almost placed last in the TP and FN categories - this suggests that the MNN hybrid model is the rightful runner-up to the MGSVM hybrid model. Subject to a larger training dataset, the EBT algorithm does have the possibility of excelling, but the likelihood is that the EBT algorithm would continue to call an excessive number of values - negative. The WKNN and EBT hybrid model share similar patterns despite the different degree of extremes, which means the WKNN hybrid model may also possibly benefit from a lager training dataset. A possible cause of the labelling bias may stem from the fact that the VADER dataset contains 829 more negative than positive samples, this translates to a skew of 55.51% negative to 44.49% positive.
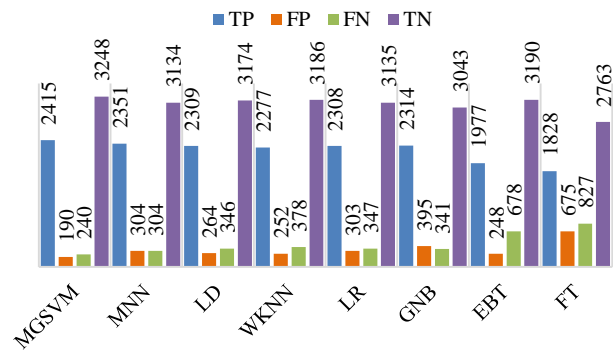

Fig. 2. A grouped bar chart comparing the VCM values of the eight hybrid models.

Table II holds performance metrics such as: accuracy, precision, recall and f1-score. These metrics are a resultant of calculations derived from VCM values. Table II additionally highlights the dominant stance the MGSVM hybrid model holds in every category in comparison to the several other hybrid models. The MGSVM hybrid model generates: an accuracy lead of 2.9% in comparison to the runner-up MNN; a precision of 92.7%, which in turn corresponds to a substantial contrast of 2, 7% in comparison to WKNN at 90%; a recall difference of 2.5% in comparison to the runner-up MNN at 88.5%; and finally, a specificity ratio at 0,945, in comparison to the runner-up EBT at 0,928.

TABLE II. THE PERFORMANCE EVALUATION METRICS FOR THE EIGHT HYBRID MODELS

Model	Accuracy (%)	Precision	Recall	Specificity	F1-Score
MGSVM	92,9	0,927	0,910	0,945	0,918
MNN	90,0	0,885	0,885	0,912	0,885
LD	90,0	0,897	0,870	0,923	0,883
WKNN	89,7	0,900	0,858	0,927	0,878
LR	89,3	0,884	0,869	0,912	0,877
GNB	87,9	0,854	0,872	0,885	0,863
EBT	84,8	0,889	0,745	0,928	0,810
FT	75,3	0,730	0,689	0,804	0,709

When compared to the other hybrid models, the MGSVM hybrid model is the overall best performer in the F1-score category – at 91.8% it secures a marginal triumph of greater than 3%. Following the MGSVM hybrid model in the F1 performance metric was MNN at 88.5% and LD at 88.3%. Comparatively the FT hybrid model performed terribly and only achieved an F1-score of 70.9%.

The ROC curve serves as a graphical representation of the relationship between the TP rate and FP rate indices of a model. It additionally serves as a visual means of assessing the model's diagnostic capacity. The MGSVM hybrid model steadily outperforms the other hybrid models, as seen in Fig. 3. MNN, LD, WKNN, and LR, the middle runners, are all close to each other. Furthermore, falling just short of the middle runners is the GNB hybrid model, and despite the smooth curve, the EBT hybrid model lags to an even greater extent. Finally, the FT hybrid model curve shows how the slow to train and erratically the hybrid model behaves as it places in the last position.

Fig. 4.   AUC, TP, and FP rate of all hybrid models.

TABLE III.   ADDITIONAL GENERAL HYBRID MODEL TRAINING RESULTS

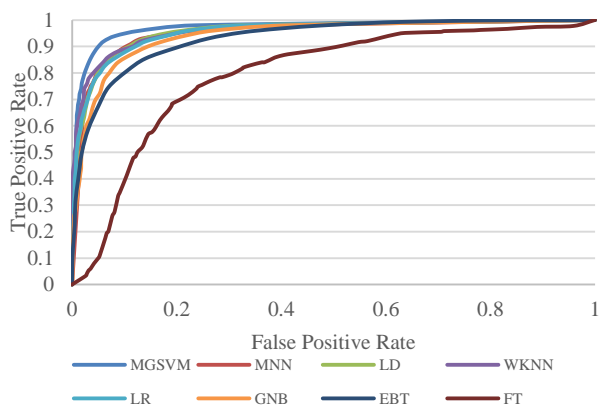| Created Hybrid Model	Training Results (General)	
	Total Misclassification Cost	*Prediction speed (objects/milliseconds)*
MGSVM	430	~7,5
MNN	N/A	~72
LD	610	~23
WKNN	630	~1
LR	N/A	~25
GNB	736	~45
EBT	926	~22
FT	1502	~43

Fig. 3.   The ROC curve of the eight hybrid models.

The area beneath the ROC is also of great importance as larger AUC values are often desired and this number falls between 1 and 0. Fig. 3 shows the superior area beneath the curve which the MGSVM hybrid model holds, and subsequently has a very good AUC score of 0.98 as per Fig. 4. The majority of the other hybrid models fall between 0.95 and 0.93, resulting in a minimum 0.02 deficit in favor of the MGSVM hybrid model. At 0.79, FT performs dismally once more. Fig. 4 highlights both the AUC, and the "co-ordinates" where the "current classifier" point exists on the ROC in Fig. 3, although not depicted; these co-ordinates may be found along the columns labelled "True Positive Rate" and "False Positive Rate".

As presented in Table III, each hybrid model furnished generic training results with metrics, such as 'total misclassification cost' and 'prediction speed'. The entire misclassification cost of the MGSVM hybrid model was 430, and the forecast speed was approximately 7, 5 objects per millisecond.
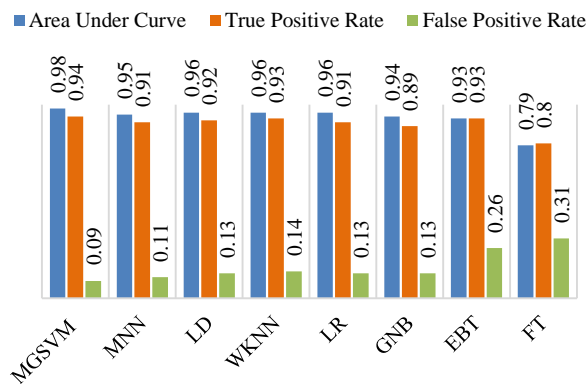
Training time – another important model value – is predominantly evaluated in two scenarios: firstly, a constant supply of emerging data which in turn leads to frequent retraining of the model; secondly, more complex or lager datasets can also lead to a dramatic increase in training times. Due to the use of the smaller simplified VADER corpus, training time is less of a concern in comparison to a larger lexicon, such as SentiWordNet, which includes around 117000 keywords.

Fig. 5 shows the LD hybrid model was trained in a mere 5, 18 seconds – making it undoubtedly the quickest to train in this instance. Second to this was GNB at 9,39 seconds followed by LR at 12,12 seconds. The favored MGSVM hybrid model achieved fifth position at 23,78 seconds. Following from here onward the training times of the other hybrid models, except for the FT hybrid model, rapidly deteriorated to approximately the 70-80 second domain. Keeping the context of the study in mind, training times would only be relevant should the dataset: grow in some way; be replaced by the production or disclosure of a larger more suitable lexical dataset; or require frequent retraining. In the context of training time, the LD hybrid model may only be worth taking into consideration as a backup option should it not require frequent retraining, as it is not susceptible (more so than other hybrid models) to dramatic increases in training time (relative to the size of the new dataset) or should the MGSVM hybrid model fail. Finally, one can concur that

the MGSVM is more than acceptable for this application, because of the one-time training requirement and a more than bearable training duration.

B. Hybrid Model Performance Regarding Twitter Dataset Sentiment Analysis:

Pre-processing the ZA public Twitter dataset was determined as a straightforward task that was successfully completed. Prior to allowing the MGSVM hybrid model to determine the sentiment of the pre-processed ZA public Twitter dataset, it was crucial to consider how the hybrid model will handle both in terms of data it had encountered and that in which it hadn't. Fig. 6 shows two-word clouds: the one on the left highlights terms that the hybrid model predicts will have a positive opinion, while the one on the right highlights phrases that the hybrid model predicts will have a negative emotion. This step is crucial as it allows for basic visual conformation that the hybrid model is behaving reasonably and within expected range when judging sentiment for both words it has yet to encounter, as well as those contained within the lexical corpus training dataset.

As per the positive word cloud in Fig. 6, it is observed that the hybrid model correctly classified most new terms supplied as positive. For instance, enriching, excellence, happy, inspirational, and wholesome have an undeniably positive connotation. However, there were several unusual predictions made by the hybrid model:

- Kelly - a proper noun used to provide someone/something a name. Due to the lack of emotion

behind the word it remains commonly regarded as neutral.

- Cuppa - a traditionally British colloquialism for "a cup of tea". It does not feature in the VADER lexicon; however, it tends to carry a positive sentiment.

- Polo - a term that refers to either an equestrian sport or a popular Volkswagen automobile model. Although traditionally the word is neutral, the positive sentiment may have incorrectly been derived from either apology or one of its derivatives that VADER deems a positive term - the derivative "apologizing" is the only derivative seen as negative.
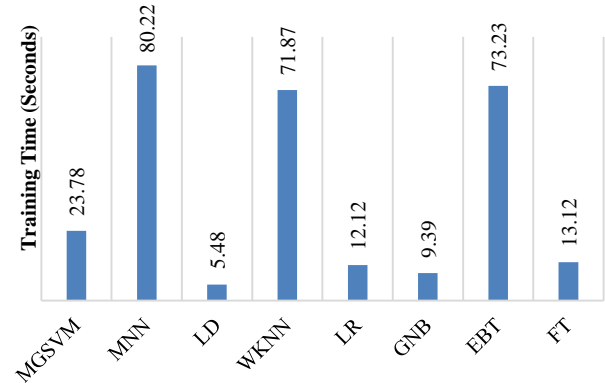

Fig. 5. A hybrid model training time comparison.

Fig. 6. A word cloud containing words of positive (left) and negative (right) sentiment that the hybrid model has labelled from the ZA Twitter public dataset.

The word cloud consisting of 'positive' words tended to contain terms of a predominantly positive nature - with only a few neutral exceptions. It did not contain any negative words.

Observably the negative word cloud in Fig. 6, highlights that the hybrid model correctly classified most of the new terms provided as negative. Corrupt, inhumane, misogynistic, plagued, and suffer are all words that have a negative connotation. Only words with a negative emotional context were found to be in the 'negative' word cloud, with no discernible words of a neutral or positive nature detected. That said, an argument could be made that the 'negative' word cloud contained a few words that held contextually varying sentiment. This can be demonstrated with the word "cut", which should appear as neutral in "my new knife has a brilliantly clean cut." Considering the word "cut" as negative in this context risks throwing the sentiment of a clearly positive string of text into an overall neutral result.

After successfully running the pre-processed ZA public Twitter dataset through the MGSVM hybrid model, the results were further processed utilizing an aggregate scoring technique to achieve document-level sentiment analysis – that output a resultant Tweet sentiment score. Overall, it took 45 minutes and 25 seconds to run the hybrid model on the entire pre-processed ZA public Twitter dataset (67585 total Tweets) to get the mean sentiment of each Tweet.

V. Discussion

The hybrid model identified the below five Tweets in the dataset as the most positively oriented in terms of prediction efforts (any duplicate Tweets by users are ruled out):

- "Neymar enjoying his @redbull https://t.co/rb1JxQiIct" – Perceivably positive. Neymar is pictured in the post as being in a celebratory mood posing with an energy drink.

- "Thank you @shotsbysbu this is so beautiful #bbnajia2020 #SSDiski #MotoGP https://t.co/NJ6kAClvdq" – Perceivably positive. The Twitter shortened link leads to the image of a wonderful art piece that the author refers to.

- "Together with her beauty #SkeemSaam https://t.co/I99y40BxUP" – Perceivably positive. The author further complements and connects with a Twitter shortened link, to which another Twitter user encourages the wedlock of a TV personality.

- "Children? Goodluck https://t.co/Y3FPLJkwmA" – Perceivably negative. The author is either implicating that complex matters may be difficult to educate children on, or alternatively believes the Twitter linked author has had a lapse of judgement for wanting to educate their children about the LGBTQI+ community. The hybrid model can only analyze the textual component and is unable to study the context required from the additional Twitter shortened link, resulting in the prediction error.

- "This is nice. #CouplesDay https://t.co/AeEaNPIGA4" – Perceivably positive. The author expresses enthusiasm towards couple's day. Although doubtful, the term of phrase "This is nice." could also be taken sarcastically. However, there are no grounds to support either assertion as the URL is no longer available.

The hybrid model identified the below five Tweets in the dataset as the most negatively oriented in further prediction efforts (once again, any duplicate Tweets by users are ruled out):

- "Neymar the Dangerous https://t.co/np6RzEMoxp" – Perceived as either negative or positive. The author is either a fan or an oppositional critic of Neymar, a professional football athlete. The link leads to an image showing Neymar in possession of the ball and closely surrounded by three oppositional players. The hybrid model is clearly focused on the negative connotation of the term dangerous. Due to the inability to further explore image sources the hybrid model is unable to gain greater contextual understanding.

- "YOU ARE CURSED https://t.co/ftP4t4eG0y" – Perceivably negative. The Tweet author is damning another user. The original Tweet author implies that the author in the link is possessed, as they would like to begin teaching children about homosexuality from the age of seven.

- "Mama boyza she lying jerrrrrrr #SkeemSaam" – Perceivably negative. The author is upset and frames a female character on the TV show "Skeem Saam" as dishonest.

- "Wow, Im sad #RIPNdlovu" – Perceivably Negative. The author is grieving the loss of a figure in their lives.

- "Zimbabwe. We are in trouble. #PutSouthAfricansFirst. #SAMediaMustFall https://t.co/EPo3JS4E4N" – Perceivably Negative. The author may share a negative sentiment with the link's author, who is concerned about the asylum issue and believes that South Africa suffers the brunt of the responsibility.

The below Tweets were deemed neutral by the hybrid model, despite the positive sentiment intended by the author:

- "Same. https://t.co/xesDjgfA4d" – Perceivably positive. The author and link author share the same mindset to teach their children about relationships and the LBGTQI+ community. Due to model limitations, the hybrid model has no further context and the Tweets main text "Same." ends up taking on a neutral sentiment.

- "Saving this https://t.co/354EU4auV1" – Perceivably positive. The author agreeingly feels that educating their children about homosexuality and relationships is essential, hence the desire to save the content for future reference. The term "saving this" is interpreted as neutral by the hybrid model as it is unable to further explore the contents behind the Twitter shortened link.

The below Tweets were deemed neutral despite the hybrid model having acquired clear direction in the training dataset:

- "Thank you @chrishemsworth https://t.co/VCFxtYuYLG #motivate #sundaymotivation #covid" – Perceivably positive. The author is thanking a famous film actor via their Twitter handle and using hashtags of a positive nature, except for "#covid" which has a negative stigma. Even though the hybrid model had previously identified, within the positive word cloud, the closed compound word "thankyou", the inability of the hybrid model to further perceive the open compound of "thank you" as positive is uncertain.

- "Wtf is this? https://t.co/hffp9iOFBU" – Perceived as either negative or positive. The author is confronted by the image of a French Poodle on a red sofa followed by the heading "nothing to see here", this method of notifying a user of a missing page is done by Twitter to try and lift the users' spirits. In this instance the author might use the phrase "Wtf is this?" as a positive means to show their followers the amusing content. Alternatively, the author may be frustrated by the missing content hence the use of "Wtf". Further analysis reveals that the acronym 'wtf' is classified as negative in the VADER lexicon, as to why the Tweet was deemed as neutral is a mystery.

VI. LIMITATIONS

Throughout the research several limitations were discovered, this is to be expected. Effort was made to overcome several of the limitations – however the list below highlights unresolved limitations faced:

- Twitter API access and the use of a publicly available dataset. Despite numerous attempts the researchers could not gain research access to the Twitter API, this meant that a publicly available dataset became the next best available option. However, the use of a publicly available dataset means that search selection criteria are lost, and one could argue that it leads to an imperfect artificial environment for model testing.

- Punctuation removal. Certain entries in the VADER lexicon are punctuation based emojis. Additionally, the general use of punctuation adds to sentiment. The removal of punctuation resultantly leads to a loss in sentiment.

- Twitter dataset Hexadecimal Unicode removal. Aside from punctuation based emojis which were ultimately removed, the VADER dataset does not contain any hexadecimal Unicode emojis. The removal of hexadecimal Unicode, however, ultimately results in a loss of sentiment when assessing the Twitter dataset.

- The removal of certain words that do not appear in the isVocabularyWord subfunction of the function fastTextWordEmbedding. To effectively use the word2vec subfunction in MATLAB textual data should pass through the isVocabularyWord subfunction. However, passing certain misspelled, unknown slang, or other forms of other unknown words – means that they are eradicated from both the training dataset and the tested Twitter dataset.

- Lack of contextual inclusion. Despite Twitter predominantly being a text-based platform, social media contains text, images, videos, and URLs. These other forms of medium are unreadable by a text-based sentiment analysis model.

VII. CONCLUSION

The research evaluated several hybrid sentiment analysis models, of which a state-of-the-art simplified VADER centered hybrid sentiment analysis model was produced. In accordance with the results, the MGSVM hybrid model outperformed the other hybrid models by a significant margin. Training time was identified as the key disadvantage of the MGSVM hybrid model creation process, demonstrated to be inconsequential in the future as instances of retraining would be rare and the use of a more expansive and/or intricate corpus was not a necessity. Additionally, the LD, LR, MNN, WKNN, and GNB hybrid models, albeit the latter to a lesser extent, were among the middle runners. By far the FT hybrid model was deemed as the poorest performer amongst the hybrid models. The chosen MGSVM-VADER hybrid proved to be highly effective when applied to a ZA Public Twitter opinion dataset. Ultimately, the research question "Will the MGSVM-VADER hybrid model approach produce accurate sentiment analysis of public Twitter opinion in South Africa?" was in turn effectively answered.

VIII. FUTURE WORK

The following work would build upon the MGSVM-VADER hybrid in this study. A specialty language/slang modified VADER lexicon poses as a research possibility for a particular future audience. The creation of a system that also considers emojis, hexadecimal Unicode, and punctuation, and then translates them into the textual equivalent to be used with purely text-based sentiment analysis models. Additionally, the creation of a contextually aware hybrid sentiment analysis model – that could possibly explore subject matter behind URL links, videos, and images – could be an interesting future prospect.

DATA AVAILABILITY

The full VADER Dataset is available online at: https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt.

The SA Twitter dataset is available online at: https://www.kaggle.com/mbusomakitla/south-african-twitter-dataset.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

REFERENCES

[1] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, "Sentiment analysis on social media," in Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, pp. 919-926. 26-29 August 2012.

[2] R.M. Alguliyev, R.M. Aliguliyev, and G.Y. Niftaliyeva, "Extracting social networks from e-government by sentiment analysis of users' comments," Electronic Government, an International Journal, 15(1), pp. 91-106. 2019.

[3] A, Corallo, L. Fortunato, M. Matera, M. Alessi, A. Camillò, V. Chetta, E. Giangreco, and D. Storelli, "Sentiment analysis for government: An optimized approach," in Proceedings of International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, pp. 98-112. 2015

[4] S. Verma, "Sentiment analysis of public services for smart society: Literature review and future research directions," Government Information Quarterly, 39(3), 101708. 2022.

[5] O. Alqaryouti, N. Siyam, A.A. Monem, and K. Shaalan, "Aspect-based sentiment analysis using smart government review data," Applied Computing and Informatics, 2020.

[6] F. Hemmatian and M.K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," Artificial intelligence review, 52(3), pp. 1495-1545. 2019.

[7] Y. Wang, Y. Rao, and L. Wu, "A review of sentiment semantic analysis technology and progress," in proceedings of 2017 13th International Conference on Computational Intelligence and Security (CIS). IEEE, 452-455. 15-18 December 2017.

[8] E. Kotzé and B. Senekal, "Employing sentiment analysis for gauging perceptions of minorities in multicultural societies: An analysis of Twitter feeds on the Afrikaner community of Orania in South Africa," TD: The Journal for Transdisciplinary Research in Southern Africa, 14 (1), pp. 1-11, 2018.

[9] J. Oyasor, M. Raborife, and P. Ranchod, "Sentiment analysis as an indicator to evaluate gender disparity on sexual violence tweets in South Africa," in proceedings of 2020 International SAUPEC/RobMech/PRASA Conference. IEEE, pp. 1-6, 29-31 January 2020.

[10] Y. Khan, "A longitudinal sentiment analysis of the# FeesMustFall campaign on Twitter," diss, 2019.

[11] H. Sims, "Public confidence in government, and government service delivery," Citeseer, 2001.

[12] E. Isin and E. Ruppert, "Being digital citizens," Rowman & Littlefield Publishers, 2020.

[13] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," Knowledge-Based Systems, 226, p.107134. 2021

[14] M.H. Abd El-Jawad, R. Hodhod, and Y.M. Omar, "Sentiment analysis of social media networks using machine learning," in proceedings of 2018 14th international computer engineering conference (ICENCO). IEEE, pp. 174-176. 29-30 December 2018.

[15] C.J. Varshney, A. Sharma, and D.P. Yadav, "Sentiment Analysis using Ensemble Classification Technique," in proceedings of 2020 IEEE Students Conference on Engineering & Systems (SCES). IEEE, pp. 1-6, 10-12 July 2020.

[16] O. Araque, I. Corcuera-Platas, J.F. Sánchez-Rada, and C.A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," Expert Systems with Applications, 77, pp. 236-246, 2017.

[17] D. Tiwari and B. Nagpal, "Ensemble Methodsof Sentiment Analysis: A Survey," in proceedings of 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, pp. 150-155, 12-14 March 2020.

[18] J. Dou, A.P. Yunus, D.T. Bui, A. Merghadi, M. Sahana, Z. Zhu, C.W Chen, Z. Han, and B.T. Pham, "Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan," Landslides, 17 (3), pp, 641-658, 2020.

[19] R. Kashef, "A boosted SVM classifier trained by incremental learning and decremental unlearning approach," Expert Systems with Applications, 167, p.114154, 2021.

[20] J.Y. Huang, W.P. Lee, and K.D. Lee, "Predicting Adverse Drug Reactions from Social Media Posts: Data Balance, Feature Selection and Deep Learning," Healthcare, 10(4), p. 618, MDPI. 2022.

[21] F. Alzamzami, M. Hoda, and A. El Saddik, "Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation," IEEE Access, 8, pp.101840-101858, 2020.

[22] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, and G.S. Choi, "GBSVM: Sentiment Classification from Unstructured Reviews Using Ensemble Classifier," Applied Sciences, 10(8), p. 2788, 2020.

[23] A. Alrehili and K. Albalawi, "Sentiment analysis of customer reviews using ensemble method," in 2019 International conference on computer and information sciences (ICCIS), pp. 1-6, IEEE. 03-04 April 2019.

[24] N. Sultana and M.M. Islam, "Meta classifier-based ensemble learning for sentiment classification," in proceedings of International Joint Conference on Computational Intelligence, pp. 73-84, Springer. 2020.

[25] B. Ray, A. Garain, and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," Applied Soft Computing, 98, p.106935, 2021.

[26] L. Zhao, L. Li, X. Zheng, and J. Zhang, "A BERT based sentiment analysis and key entity detection approach for online financial texts," in proceedings of 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 1233-1238, IEEE, 05-07 May 2021.

[27] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," Applied Intelligence, 48, pp. 1218-1232, 2018.

[28] S. Mendon, P. Dutta, A. Behl, and S. Lessmann, "A hybrid approach of machine learning and lexicons to sentiment analysis: Enhanced insights from twitter data of natural disasters," Information Systems Frontiers, 23, pp. 1145-1168, Springer, 2021

[29] Y. Al Amrani, M. Lazaar, and K.E. El Kadiri, "A Novel Hybrid Classification Approach for Sentiment Analysis of Text Document," International Journal of Electrical & Computer Engineering, 8(6), pp. 4554-4567, 2018.

[30] X. Jiang, "Isolated Chinese sign language recognition using gray-level co-occurrence matrix and parameter-optimized medium gaussian support vector machine," In Frontiers in Intelligent Computing: Theory and Applications, pp. 182-193. Springer, 2020

[31] SL. Lin, "Application of Machine Learning to a Medium Gaussian Support Vector Machine in the Diagnosis of Motor Bearing Faults," Electronics, 10(18), p. 2266, MDPI. 2021.

[32] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in proceedings of the International AAAI Conference on Web and Social Media, 8(1), 2014.

[33] C.A. Ramezan, T.A. Warner, and A.E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification," Remote Sensing, 11(2), p.185, MDPI. 2019.

[34] T. Mikolov, K.Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[35] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," Procedia computer science, 152, pp. 341-348, 2019.