EDAS: Embedded Distribution Assessment Scale to Choose the feature n-gramsfor Sentiment Classification Over Twitter Streams.

T. Jhansi Rani
Research scholar, Dept., of CSE
JNTUH, Hyderabad
Telangana, India
jhansi.rani.t@gmail.com
K. Anuradha
Professor, Depat., of CSE
GRIET, Hyderabad
Telangana, India

Abstract:

In order to determine the target user's perception of the source environment, opinion mining on twitter streams is essential for current market surveys, product identity evaluations, and more. There are several alternatives for choosing characteristics because of the tweets' high cohesiveness toward NLP. Using supervised learning techniques, modern opinion mining and sentiment analysis determine opinion polarity using sentiment lexicons. Positive and negative sentiment polarities were observed in the majority of recent submissions. Only a few recent contributions employed supervised learning to distinguish between neutral, positive, and negative viewpoints. The Embedded Distribution Assessment Scale (EDAS), a tool for choosing the best n-grams for Twitter sentiment categorization, was introduced in this work. The proposed approach uses supervised learning to forecast the polarity of sentiment that is positive, negative, neutral, and unreviewed. In contrast to current models, this paper does supervised learning using four sentiment polarity labels and consecutive n-grams. By comparing cross validation metrics from the suggested model with other current models, the performance of the proposed model was examined, highlighting its performance advantage.

Keywords:*NLP, Likert Scale, Support Feature vector Machine, Bernoulli NB, Multinomial Naïve Bayes, MLPE, Cuckoo Search.*

DOI Number: :10.14704/nq.2022.20.8.NQ44180 **NeuroQuantology 2022 ;20(8):1682-1694**

1 Introduction

The web serves as a conduit in this movement, enabling user interaction and information sharing [1]. A web-based method called social media [2] makes user communication interactive. Users' views, feelings, and opinions are expressed on microblogging and social media sites like YouTube, Facebook, and Twitter [3].

A microblogging site called Twitter [3] has 326 million members who send 500 million tweets daily and 200 billion tweets annually.

Users of Twitter have access to tweets regarding a variety of topics, including goods, movies, events, organisations, political problems, and more [4].

Large-scale data extraction and processing can assist in providing answers to technological and societal concerns [5]. Twitter views, on the other hand, are seen as necessary data that reflect the wants and sentiment of the general population. With the use of data mining and sentiment analysis, these posts are utilised to gauge user sentiment.

The difficulty of sentiment analysis is in categorising the opinions or feelings conveyed in texts by different persons. By identifying people's negative or positive emotions, sentiment analysis may determine their level of interest in a subject [6]. Finding every tweet with more than 140 characters utilising acronyms, URLs, emotions, abbreviations, hashtags, and slang is a big and challenging challenge in Twitter sentiment analysis [7]. The most popular sentiment analysis models, according to [8], may be divided into two groups and subclasses.

Sentiment analysis requires a non-linguistic or linguistic feature set for ML-based models with supervised and unsupervised learning.

current models consist of: In contribution [7], the TSA model is put out; researchers used unigrams to extract characteristics and treated tweets like word clouds. An information retrieval approach is used to create a fuzzy lexicon that uses sentiment-similarity to represent every attribute. Three well-known classifier algorithms, SVM, MNB, and BNB, were trained using these characteristics (Bernoulli NB).

Each technique has drawbacks and advantages that might influence the effectiveness and accuracy of the classifier. Overall matrices would combine the advantages of all four approaches and address their drawbacks. The goal of this book is to create a hybrid representation matrix using four different matrices. 4 matrices are combined via a fusion operator. It wouldn't be ideal to take into account a comparable contribution from each model when creating the projected matrix. The weight of each model should be stated.

2 Related work

Experts have tried a variety of research tactics for recovering prediction from such assets due to the enormous amount of information available on a variety of interpersonal connection sites where individuals express their thoughts, feelings, and conclusions [9]. Tweets and other informal, concise information have also been helpful. Researchers developed a predictive highlight generator using a hybrid approach. The developed approach is robust against maximum entropy, NV, and SVM. Based on the findings, sentiment characteristics may offer greater accuracy than conventional text analysis [10]. Multi-class classifiers have also been utilised for outcomes with more than two class labels, and performance has been evaluated using various technique selection and assessment procedures [11]. Sarcastically, "Bootstrapped learning of negative and positive attitudes".

Other recent works on sentiment analysis include "Solving the Twitter Sentiment Analysis Problem" [12] and "Machine Learning for Product Evaluation" [13]. Both analyse Twitter sentiment using machine learning. When choosing the best features with a genetic algorithm, STSAP employed feature weights as fitness measurements. From tweets, MLPE [13] retrieved product reviews. These contributions make use of n-grams and sentiment lexicons. In order to generate optimum features, the STSAP [12] use evolutionary algorithms, a nondeterministic method that increases complexity and lacks precision. The other addition, which is not original, employs regular n-gram patterns of sentiment lexicons as features. The technique described in this publication uses n-gram sequences as features, integrates diversity assessment metrics to choose the best n-gram sequences as features, and employs a heuristic multi-label classification model to train and classify with the least amount of false alerts.

3 Methods and Materials

3.1 The Data

The dataset that considered for experimental study is Twitter US Airline Sentiment [14] 2364 positives, 3100 neutrals, 9166 negatives. However, the label statistics are not balanced. Hence, the data has been synthesized to balance the records count. In addition, set of tweets have been synthesized, which have been labeled as “not a review”. The final dataset contains twitter tweets of equal size (records of size 9166 for sentiment polarity labels positive, negative, neutral, and “not a review”). Hence, the total records of the data set are 36664.

3.2 Preprocessing

For each record of the training corpus, the preprocessing phase removes stop words (not special characters), further stems “ing” and “ed” forms from the resultant record of the stop words removal phase. Later on, the resultant bag of tokens, which includes words, special characters like emoticons. Each result record will have their original sentiment polarity label.

3.3 Model Description

The best word patterns are chosen from diversity assessment measures using supervised learning in the proposed approach for sentiment polarity recognition from Twitter tweets. Cuckoo search, a heuristic search method, makes hierarchical predictions on sentiment polarity. A technique for weighting word patterns in relation to sentiment polarity was created under the influence of the Likert Scale [15]. Methodology: Each record includes word patterns based on Twitter tweets that fall into several sentiment polarities. The final column of each record displays sentiment polarity. Furthermore, it divides the provided labelled records of various sentiment polarities into a number of groups, each of which includes records of a certain sentiment polarity.

For every record encoding a sentiment polarity, it also finds sequential n-grams, including emoticons.

Applying diversity evaluation metrics to each size of the n-grams of one group's records with the corresponding column of the other group is necessary to determine the best n-grams for each sentiment polarity. The optimality of a column is determined by the observed diversity. In the last stage, successive n-grams are derived from the best word patterns for each sentiment polarity in order to train a classifier [16] that was created using "cuckoo search" [17]. The last stage of the concept forecasts the sentiment polarity of Twitter tweets.

The proposal adopted the MWU-Test, the dual-tailed t-test, and the KS-Test for the fusion of diversity measures.

3.4 Mann-Whitney U Test

Mann-Whitney U Test (MWU-Test) [19] is among the multiple diversity assessment methods, which does not include centric to the distribution format that deserves in most of the datasets having recorded with diversified labels. The description of the MWU-Test implementation process is as follows:

The notations v_1, v_2 denote the feature vector distributions used as input to the method MWU-Test to conclude the scope of diversity between corresponding feature vectors, which is as follows.

$RST_1 = RS_1 - \frac{|v_1| \times (|v_1|+1)}{2}$ ...(Eq 1)// the notation $|v_1|$ denotes the size of the feature vector v_1.

Similarly, the rank-sum threshold RST_2 of the feature vector v_2 will be determined

Then the rank-sum threshold RST of the feature vectors' entries v_1, v_2 is the sum of rank-sum thresholds RST_1, RST_2 of the feature vectors v_1, v_2 that are followed in (Eq 3).

$$RST = RST_1 + RST_2 \text{ ...(Eq 2)}$$

In order to find the z-score [21], Initially, find the mean m_{RST} and standard deviation d_{RST} as follows in (Eq 4), (Eq 5):

$$m_{RST} = \frac{RST}{2} \text{ ...(Eq 3),}$$

$$d_{RST} = \sqrt{\frac{|v_1|*|v_2|*(|v|+1)}{|v|}} - \sqrt{\frac{|v_1|*|v_2|}{|v|}\left((|v|+1) - \sum_{i=1}^{k}\frac{t_i^3 - t_i}{|v|*(|v|-1)}\right)} \text{ ...(Eq 4)}$$

Here in (Eq 4), (Eq 5), the notation k denotes the number of distinct ranks, t_i denotes the number of entries sharing the same rank i

Further, the z-core assesses as $z = \frac{RST - m_{RST}}{d_{RST}}$ **...(Eq 5)**

Then find the p-value of the depicted z score in z-table [22]. If the p-value is found to be greater than the given probability threshold (usually 0.01, 0.05, or 0.1), then the distribution of the feature vectors v_1, v_2 is found to be diversified. Else the distribution is similar.

3.5 KS-test

Kolmogorov-Smirnov test (KS-test) [18] is a distribution diversity assessment measure, which has been used to assess the diversity between the values projected for each feature attribute of given two datasets.The significance of the ks-test is that it can apply to assess the diversity between two feature vectors of variable size $(|fv_a| \neq |fv_b|)$. The diversity assessment of two feature vectors by KS-Test is as follows:

The given two feature vectors fv_a, fv_b, representing the feature values of feature attribute in distinct datasets. The KS-Test will implement in concern to evaluate the distributions of 2 feature vectors are similar or divergent as follows:

$$\begin{array}{l} cr = 0 \\ \overset{|fv_j|}{\underset{i=1}{\forall}} \left\{ \left(cr = \frac{el_i}{Ag(fv_j)} + cr \right) \exists el_i \in fv_j \right\} \\ CR_{fv_j} \leftarrow cr \end{array} \quad \text{...(Eq 6)}$$

The notations used in Eq 7 are,

The notation el_i denotes the feature value representing the corresponding feature attribute.

The notation $Ag(fv_j)$ denotes the aggregate of the feature values listed in feature vector fv_j

The notation CR_{fv_j} denotes the set of cumulative ratio of the elements $\{el \exists el \in fv_j\}$ exists in given feature vector fv_j.

Concerning to the aforesaid ks-test process, the cumulative ratios CR_{fv_a}, CR_{fv_b} of the values representing the feature vectors fv_a, fv_b in respective order.

Further discovers the absolute Difference as a set $AD_{\{CR_{fv_a} \leftrightarrow CR_{fv_b}\}}$ of thecumulative ratios of the features listed in sets CR_{fv_a}, CR_{fv_b} in respective order, which is as follows.

$\overset{\max(|CR_{fv_a}|,|CR_{fv_b}|)}{\underset{i=1}{\forall}} \{cr_i(fv_a), cr_i(fv_b) \exists cr_i(fv_a) \in CR_{fv_a} \wedge cr_i(fv_b) \in CR_{fv_b}\}$ Begin // for allcumulative ratios exists in sets CR_{fv_a}, CR_{fv_b}

$AD_{CR_{fv_a} \leftrightarrow CR_{fv_b}} \leftarrow abs(cr_i(fv_a) - cr_i(fv_b))$ //discovering the absolute Difference $abs(cr_i(fv_a) - cr_i(fv_b))$ as a set $AD_{\{CR_{fv_a} \leftrightarrow CR_{fv_b}\}}$ of the cumulative ratios of the features listed in sets CR_{fv_a}, CR_{fv_b} in respective order

End

The maximum value of the set $AD_{CR_{fv_a} \leftrightarrow CRf_{v_b}}$ denotes further as d-stat helps to find diversity scope. If the d-stat is greater than the d-centric, then confirms the diversity of the given feature vectors is poor, else the diversity of the both vectors is significant. The aforesaid d-centric is the degree of probability threshold of the sets $Ag(fv_a), Ag(fv_b)$ that tracked from the KS-table [23].

3.6 Dual Tailed Variance Test

The distribution diversity assessment scale Dual tailed diversity test that renowned as a t-test has been used in the proposed fusion technique. This method enables us to determine the diversity between n-gram values projected for a n-gram of the records having different labels. The outcomes of the dual tailed t-test refined in contemporary contributions [20], [24] have motivated to adopt this method to use in the suggested model.

The t-score (diversity) between given two feature vectors $\{r_i, r_j \exists i \neq j\}$ distributions shall estimate as follows in (Eq 8):

$$f\left(r_i, r_j\right) = \left(\mu(r_i) - \mu\left(r_j\right)\right) * \left(\sqrt{\sigma_{(r_i)} + \sigma_{\left(r_j\right)}}\right)^{-1} \text{ ... (Eq 7)}$$

The function $f\left(r_i, r_j\right)$ stated in equation 10 determines the t-score between the feature vector distributions $\{r_i, r_j \exists i \neq j\}$. Notations $\mu(r_i), \mu(r_j)$ reflect the mean values of the corresponding feature vector distributions, and the notations $\sigma_{(r_i)}, \sigma_{(r_j)}$ state the standard deviation of the respective feature vector distributions. The t-score of the two feature vector distributions denotes the proportionate value of the difference between the corresponding feature vectors' mean values against the square root of the sum of their standard deviations.

The t-score that was found to be lesser than the given probability value (p-value) [22], [21] denoting the diversity between the given feature vector distributions, is significant.

3.7 The classifier

Many modern meta-heuristic algorithms stimulated by nature were emerging & becoming popular in solving many engineering problems. In this contribution, the cuckoo search (CS) technique is used. These cuckoos lay the eggs in nests, even though they might eradicate others' eggs for enhancing the hatching possibility of their eggs. The relative amount of species has engaged to brood parasitism obligation through laying eggs in another host-birds' nests. There were three fundamental brood parasitism types: 1. Cooperative breeding, 2. Nest takeover & 3. Brood parasitism of intra-specific. When discovering the eggs as not their own by host bird, they might either throw away alien eggs or abandon their nest simply and construct a novel nest.

3.8 Optimal N-gram Selection

1686

For each set D_j of the records representing j^{th} the label, find the optimal n-grams compared to the counterpart set $\{D_k \exists k \neq j\}$. For each set D_j, a n-gram x_i is optimal if the values projected to the i^{th} set's n-gram D_j are having distribution diversity with the values projected for the same n-gram x_i in other sets $\{D_k \exists k \neq j\}$. For each n-gram x_i of the set D_j, the process shall estimate the diversity weight towards each of the other sets $\{D_k \exists j \neq k\}$, which is the absolute difference between maximum similarity one and probable similarity observed $(0 \leq p-value \leq 1)$. The mathematical model of identifying optimal n-grams from each pair of sets portrayed in the following description.

$\overset{|X|}{\underset{i=1}{\forall}} \{x_i \exists x_i \in X\}$ Begin // for all the n-grams

$\overset{(n-1)}{\underset{j=1}{\forall}} \left\{\left[x_i^j\right] \exists \left[x_i^j\right] \in D_j\right\}$ // Begin

$dw_{x_i \Rightarrow D_j} = 1$ // the overall diversity of the n-gram x_i concerning the set D_j (label)

$\overset{(n)}{\underset{k=1}{\forall}} \left\{\left[x_i^k\right] \exists \left[x_i^k\right] \in D_k, j \neq k\right\}$ // Begin

$$\left.\begin{array}{l} p_{ks} = KS-test\left(\left[x_i^j\right],\left[x_i^k\right]\right); p_{mwu} = MWU-test\left(\left[x_i^j\right],\left[x_i^k\right]\right); \\ p_{dt} = DT-test\left(\left[x_i^j\right],\left[x_i^k\right]\right) \end{array}\right\}$$

// performing the fusion of diversity estimation of the n-gram x_i between the sets D_j, D_k

$d(x_i)_{D_j \Leftrightarrow D_k} = d\tau$ // the diversity $d(x_i)_{D_j \Leftrightarrow D_k}$ of the n-gram x_i between sets $D_j \Leftrightarrow D_k$ has initialized to distance threshold $d\tau$

$if\left((p_{ks} < p\tau) \vee (p_{ks} < p\tau) \vee (p_{ks} < p\tau)\right)$ Begin// either of the probable similarity value $(p_{ks}, p_{mwu}, p_{dt})$ observed for the n-gram x_i between the sets D_j, D_k has found to be greater than the given probability threshold $p\tau$

$d(x_i)_{D_j \Leftrightarrow D_k} = (p_{ks} \otimes p_{mwu} \otimes p_{dt})$ // the diversity $d(x_i)_{D_j \Leftrightarrow D_k}$ of the n-gram x_i between sets $D_j \Leftrightarrow D_k$ has been discovered from the fusion of the diversity estimation measures.

End //of the condition

$dw_{x_i \Rightarrow D_j} = dw_{x_i \Rightarrow D_j} \otimes d(x_i)_{D_j \Leftrightarrow D_k}$

End // of the iterations

$if\left(dw_{x_i \Rightarrow D_j} \geq d\tau\right)$ Begin // if the diversity weight $dw_{x_i \Rightarrow D_j}$ of the n-gram x_i towards the set D_j (label) is greater than or equal to the given diversity threshold $d\tau$

$fD_j \leftarrow x_i$ //then consider the n-gram x_i is optimal to the set D_j and move that to the optimal n-grams set fD_j

End

End// of the iterations

End // of iterations

3.9 Preprocessing the datasets of diversified labels

1687

$\overset{lc}{\underset{j=1}{\forall}} \overset{|X|}{\underset{i=1}{\forall}} \left\{x_i \exists x_i \in X \wedge x_i \notin fD_j\right\}$ Begin // for each n-gram x_i that selected as an optimal n-gram of the set D_j of the label j

$\left\{D_j\right\} \setminus [x_i]$ // discarding the n-gram x_i and values projected to the corresponding n-gram from the set D_j

End

3.10 Selecting possible sequential n-grams for each set of sentiment polarity

$\overset{lc}{\underset{j=1}{\forall}} \overset{|D_j|}{\underset{i=1}{\forall}} \left\{r_i \exists r_i \in D_j\right\}$ Begin// for each record r_i of each set D_j

$nG_j \leftarrow \left(nG_j \cup nGrams(r_i)\right)$ // discovering all possible sequential n-grams of size 1 to n and moving resultant unique sequential n-grams to the set nG_j, which does not exist in the set nG_j

End

Further phase determines the set (label) level sentiment polarity label probability, purity, and decision support of sequential n-grams occurrence in the respective set

$\overset{lc}{\underset{j=1}{\forall}} \overset{|nG_j|}{\underset{i=1}{\forall}} \left\{ng_i \exists ng_i \in nG_i\right\}$ Begin // for each n-gram $\left\{ng_i \exists ng_i \in nG_j\right\}$ of the set D_j

$ng^{+}_{[j,i]} = \left(\sum_{k=1}^{|D_j|} \left\{1 \exists ng_i \subseteq r_k \wedge r_k \in D_j\right\}\right) * \left(\left|D_j\right|\right)^{-1}$ // the positive probability ng_j^{+} of the n-gram ng_i concerning the set D_j

$ng^{-}_{[j,i]} = \left(|D_j| - \left(ng^{+}_{[j,i]} * |D_j|\right)\right) * |D_j|^{-1}$ //the negative probability ng^{+}_{j} of the n-gram ng_i concerning the set D_j

$ng^{w}_{[j,i]} = 1 - \begin{pmatrix} ng^{+}_{[j,i]} * \left(1 - ng^{+}_{[j,i]}\right) + \\ ng^{-}_{[j,i]} * \left(1 - ng^{-}_{[j,i]}\right) \end{pmatrix}$ The purity of the n-gram ng_i towards the set D_j, which has derived by Gini impurity estimation [25]

End //of the iterations indexed by i

End //of the iterations indexed by j

$\overset{lc}{\underset{j=1}{\forall}} \overset{|nG_j|}{\underset{i=1}{\forall}} \{ng_i \exists ng_i \in nG_i\}$ Begin // for each n-gram $\{ng_i \exists ng_i \in nG_j\}$ of the set D_j

$ip_{[j,i]} = 1$ // the overall impurity $ip_{[j,i]}$ of the n-gram $ng_{[j,i]}$ is initialized to maximum, which is 1

$\overset{lc}{\underset{k=1}{\forall}} \{ng_i \exists ng_i \in nG_k \wedge j \neq k\}$ Begin // for each n-gram $\{ng_i \exists ng_i \in nG_k\}$ of the set D_j

$ip_{[j,i]} = ip_{[j,i]} * ng^{w}_{[k,i]}$ // updating the impurity $ip_{[j,i]}$ of the n-gram $ng_{[j,i]}$

End

$ng^{dw}_{[j,i]} = ng^{w}_{[j,i]} - \left(1 - ip_{[j,i]}\right)$ // The decision weight $ng^{dw}_{[j,i]}$ of the n-gram $ng_{[j,i]}$ towards the set (sentiment polarity) D_j

$if\left(ng^{dw}_{[j,i]} < d\tau\right)$ Begin // If the decision weight $ng^{dw}_{[j,i]}$ of the n-gram $ng_{[j,i]}$ towards the set (sentiment polarity) D_j is less than the decision weight threshold

$nG_j \setminus ng_i$ // discarding the n-gram ng_i of the set D_j

End

End

3.11 Build the Classifier

Concerning performing classification using the recommended classifier built on cuckoo search, the classifier's learning phase portrays the perch hierarchy for each sentiment polarity of the training corpus, such that hierarchy built-in multiple levels, and each level contains one or more perches. Each perch of the hierarchy represents size of the n-gram in descending order of the n-gram sizes, and perches of each level present the n-grams of the same size representing the corresponding perch. The perches listed in a level represents the n-grams, which are having size lesser than the size of n-grams represented by the perches in the predecessor level and greater than the size of n-grams represented by the perches in the successor level of the hierarchy. The further task of the learning phase performs placing unique n-grams in perches represented by the size of the corresponding n-grams (as eggs) of each sentiment polarity.

3.11.1 Classification process

The predictive analysis task that predicts the test record sentiment polarity in the testing phase performs as follows.

Let the notation tr represents the given test record

Let the notation nG_{tr} represents the set that contains all possible unique n-grams of size one to the size $|tr|$ of the test record tr

$\overset{lc}{\underset{j=1}{\forall}} \{j \exists 1 \leq j \leq lc\}$ // Begin // for each sentiment polarity j

$jF_{tr} = 0$ // Represents fitness of the test record tr

$jF_{tr} = \sum_{i=1}^{|tnG_{tr}|} \sum_{idx=1}^{|H_j|} \sum_{k=1}^{|l_{idx}|} \left\{ \begin{matrix} tng_{[j,i]}^{dw} \exists \\ tng_{[j,i]} \in p_k \wedge \\ p_k \in l_{idx} \wedge \\ l_{idx} \in H_j \end{matrix} \right\}$ // Aggregating the decision weights of all sequential n-grams of the test record tr towards the hierarchy H_j of the sentiment polarity j

$jep_{tr}^{dw} = jF_{tr} \times \left(\sum_{idx=1}^{|H_j|} |l_{idx}| \right)^{-1}$ // Finding the empirical probability jep_{tr}^{dw} of the decision weights of sequential n-grams of the test record tr towards the hierarchy H_j of the sentiment polarity j

$$jd_{tr}^{dw} = \left(\sum_{i=1}^{|tnG_{tr}|} \sum_{idx=1}^{|H_j|} \sum_{k=1}^{|l_{idx}|} \left\{ \begin{matrix} \sqrt{\left(jep_{tr}^{dw} - tng_{[j,i]}^{dw} \right)^2} \exists \\ tng_{[j,i]} \in p_k \wedge \\ p_k \in l_{idx} \wedge \\ l_{idx} \in H_j \end{matrix} \right\} \right) \times \left(\sum_{idx=1}^{|H_j|} |l_{idx}| \right)^{-1}$$

// Finding deviation jd_{tr}^{dw} of the decision weights of sequential n-grams of the test record tr towards the hierarchy H_j of the sentiment polarity j

End

As stated in the description above, for each sentiment polarity j, the empirical probability jep_{tr}^{dw} of the decision weights observed for all possible sequential n-grams of the given test record tr and respective deviation jd_{tr}^{dw} shall assess.

3.11.2 Estimating the sentiment polarity labeling scope

The decision about the recommended sentiment polarity of the given test record tr shall perform as follows

$\underset{k=1}{\overset{|lc|}{\forall}} \{k \exists 1 \le k \le lc\}$ // Begin //for each sentiment polarity k

$ip_k = 1 - \prod_{j=1}^{|lc|} \left\{ \left(jep_{tr}^{dw} - jd_{tr}^{dw} \right) \exists j \ne k \right\}$ //impurity ip_k of the sentiment polarity k, which is the absolute difference between maximum impurity (refers as 1) and product of the absolute difference between empirical probability and respective deviation of all sentiment polarities other than the sentiment polarity

// fitness of the given record tr towards the sentiment polarity

End

Finally, it predicts the sentiment polarity of the given test records follows, concerning each sentiment polarity j, prioritizes the given test record by the empirical probability jep_{tr}^{dw} of decision weights, respective deviation jd_{tr}^{dw}, and the fitness f_j^{tr} towards the corresponding sentiment polarity. This approach has taken from the concept of the Likert scale [26].

Since the fitness of the test record towards the corresponding sentiment polarity is highly prioritized that followed by the empirical probability and deviation are in sequence of the respective order. According to the Likert Scale, the parameter with the lowest priority shall index by 1, and the index of the other parameters increment by 1 in ascending order of their priority. In order to conclude the correlation between the test record and the corresponding sentiment polarity, the values of the parameters shall multiply with their normalized value of the index as follows.

$$\overset{|lc|}{\underset{j=1}{\forall}}\{j\exists 1\leq j\leq lc\} \quad \text{// for each sentiment polarity } j$$

$$cr_{tr}^{j}=1-\left(\left(jd_{tr}^{dw}*(1-1^{-1})\right)*\left(jep_{tr}^{dw}*(1-2^{-1})\right)*\left(f_j^{tr}*(1-3^{-1})\right)\right)$$

Further concludes that the test record fall under the sentiment polarity j, which is having the highest correlation cr_{tr}^{j} towards the sentiment polarity j

Since three parameters have been considered for correlation assessment, the index range will be 1, 2, and 3, reflecting deviation, empirical probability, and fitness in respective order. Henceforth, the correlation of the test record tr towards the sentiment polarity j shall estimate as

4 Experimental Study

In the experimental study, the performance of suggested EDAS and contemporary models STSAP [12] and MLPE [13] was measured over four folds using precision, sensitivity, F-measure, and accuracy. The proposed model is compared to modern ones.

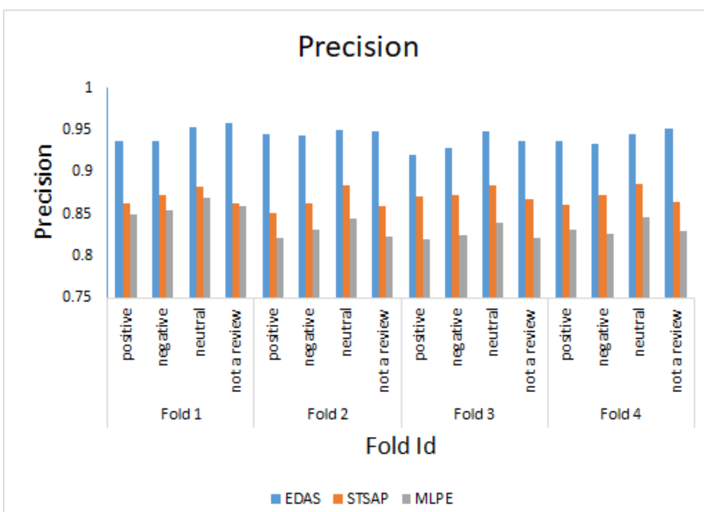

Figure 1: Metric precision comparison of EDAS, STSAP, and MLPE over four folds.

Figure 1 shows a graph with precision on the y-axis and 4 folds on the x-axis, labelled positive, negative, neutral, and not a review. EDAS, STSAP, and MLPE compare these labels. The standard deviation of the positive label for EDAS, STSAP, and MLPE are 0.9346±0.0090, 0.8610±0.0065, and 0.8307±0.01162, respectively. The negative label standard deviation for EDAS, STSAP, and MLPE is 0.9357±0.0052, 0.8705±0.0043, and 0.8341±0.0115. The neutral label standard deviation for EDAS, STSAP, and MLPE is 0.9485±0.0027, 0.8838±0.0013, and 0.8502±0.0112. The "Not a review" label's standard deviation is 0.9481±0.0076, 0.8634±0.0029, and 0.8333±0.0155 for EDAS, STSAP, and MLPE, respectively. Thus, the proposed model outperforms contemporary models in all precision labels.

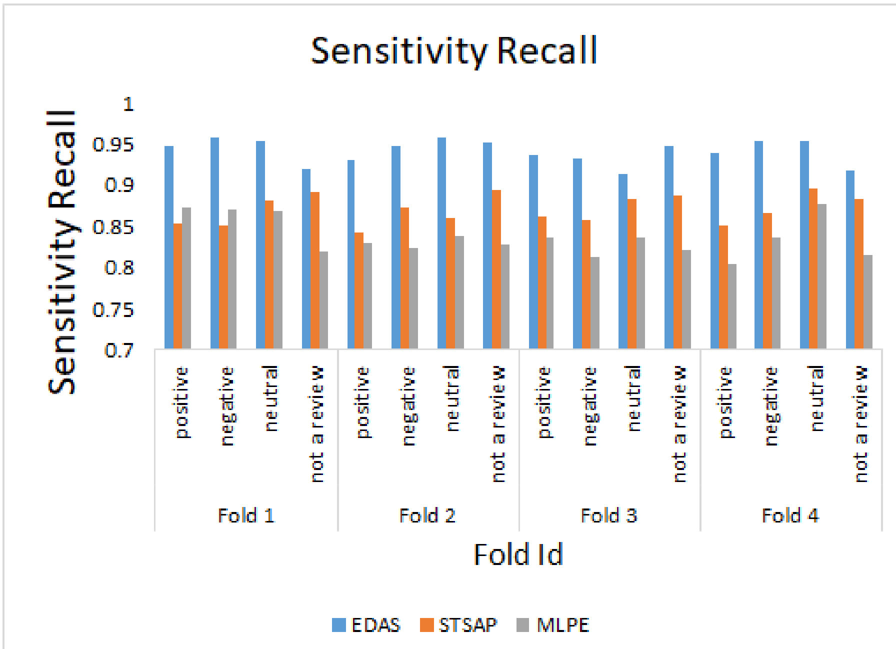

Figure 2: Comparison of EDAS, STSAP, and MLPE metric sensitivity over four folds.

Sensitivity is a ratio of true positives to overall positives called recall. This recall metric was used to compare the performance of EDAS, STSAP, and MLPE over 4 folds (Figure 2). A graph compares three models with positive, negative, neutral, and no review labels. The standard deviation of positive labels for EDAS, STSAP, and MLPE are 0.93890.0065, 0.85310.0065, and 0.83570.0243, respectively. The negative label standard deviation for EDAS, STSAP, and MLPE is 0.94780.0094, 0.86160.0080, and 0.83600.0218. Neutral label standard deviations are 0.94540.0184, 0.880290.0132, and 0.8557±0.0184 for EDAS, STSAP, and MLPE, respectively. The standard deviations of the Not a review label for EDAS, STSAP, and MLPE are 0.9343±0.0156, 0.8900 0.0044, and 0.8209±0.0047, respectively. Thus, the proposed model outperforms contemporary models in all sensitivity labels.

Figure 4 shows the accuracy metric used to compare the performance of EDAS, STSAP, and MLPE over 4 folds. A graph compares three models with positive, negative, neutral, and no review labels. The standard deviation of the positive label for EDAS, STSAP, and MLPE is 0.9367±0.0048, 0.85310.0065, and 0.8357±0.0243, respectively.

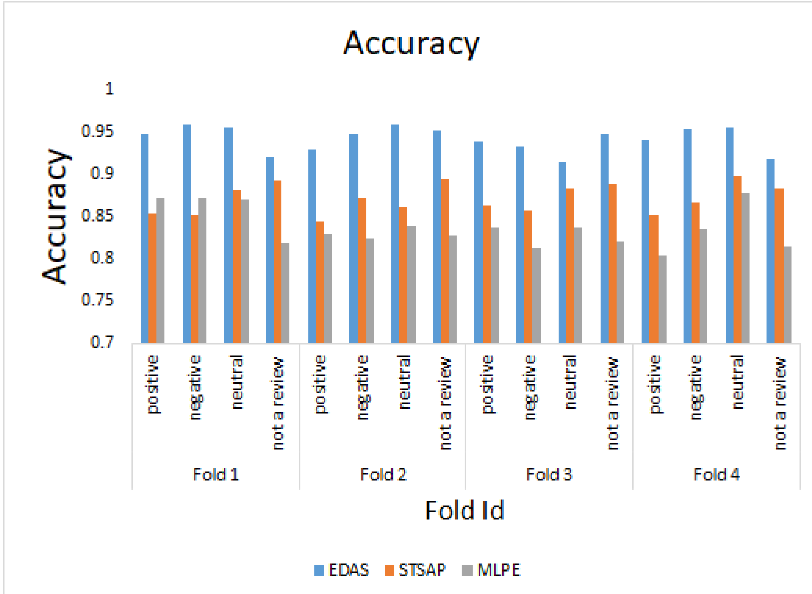

Figure 4: Metric accuracy comparison of EDAS, STSAP, and MLPE over four folds.

Negative label standard deviations for EDAS, STSAP, and MLPE are 0.9417±0.0062, 0.8616±0.0080, and 0.8360±0.0218. Neutral label standard deviations for EDAS, STSAP, and MLPE are 0.9469±0.0097, 0.88029±0.0132, and 0.8557±0.0184. The standard deviation of the Not a review label for EDAS, STSAP, and MLPE is 0.9410±0.0059, 0.8900±0.0044, and 0.8209±0.0047, respectively. The proposed model performs better in all accuracy labels than contemporary models.

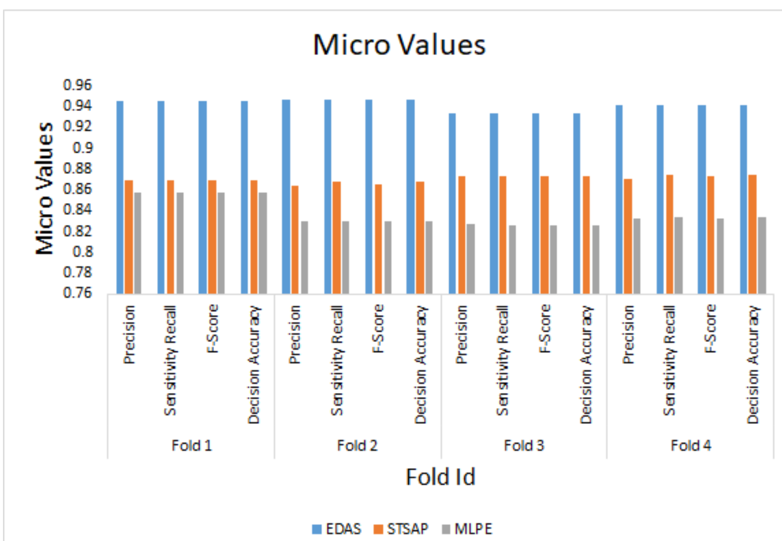

Figure 5: Comparison of EDAS, STSAP, and MLPE metrics.

Figure 5 plots micro values and metrics like precision, sensitivity, F-score, and decision accuracy over four folds for the proposed model EDAS and contemporary models STSAP and MLPE. The proposed model out-performs contemporary models in all metrics, as shown in the figure.

5 Conclusion

The objective of this manuscript is centric to portray a novel supervised learning approach that performs learning on n-grams to predict the sentiment polarity of four labels listed as positive, negative, neutral, and not a review. To identify the optimal features for learning process, fusion of diversity assessment measures has been used. A novel classification process has been tailored using heuristic search technique called cuckoo search. The experimental study has been carried on synthesized dataset by using the benchmark dataset called US Airline Sentiment. The 4-fold cross validation has been carried on proposed method and the other two contemporary models. The Performance results of the four-fold cross validation portraying that the proposed model is outperforming the contemporary methods. The future research shall focus to handle the curse of dimensionality in training corpus to achieve considerable ratio of label prediction accuracy with minimal false alarming.

References

[1] Supriya, B. N., Kallimani, V., Prakash, S., & Akki, C. B. (2016, March). Twitter senti-ment analysis using binary classification technique. In International Conference on Nature of Computation and Communication (pp. 391-396). Springer, Cham.

[2] Haque, M. (2014). Sentiment analysis by using fuzzy logic. arXiv preprint arXiv:1403.3185, vol. 4, no. 1, 33-48.

[3] Shirdastian, H., Laroche, M., & Richard, M. O. (2019). Us-ing big data analytics to study brand authenticity senti-ments: The case of Starbucks on Twitter. International Journal of Information Management, 48, 291-307.

[4] Mansour, R., Hady, M. F. A., Hosam, E., Amr, H., & Ash-our, A. (2015, April). Feature selection for twitter senti-ment analysis: An experimental study. In International Confer-ence on Intelligent Text Processing and Computa-tional Linguistics (pp. 92-103). Springer, Cham, vol. 9042.

[5] Bao, Y., Quan, C., Wang, L., & Ren, F. (2014, August). The role of pre-processing in twitter sentiment analysis. In In-ternational conference on intelligent computing (pp. 615-624). Springer, Cham.

[6] Keshavarz, H., & Abadeh, M. S. (2017). ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. Knowledge-Based Systems, 122, 1-16.

[7] Ismail, H. M., Belkhouche, B., & Zaki, N. (2018). Semantic Twitter sentiment analysis based on a fuzzy thesaurus. Soft Computing, 22(18), 6011-6024.

[8] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and appli-cations: A survey. Ain Shams engineering journal, 5(4), 1093-1113.

[9] Ramadhani, A. M., & Goo, H. S. (2017, August). Twitter sentiment analysis using deep learning methods. In 2017 7th International annual engineering seminar (InAES) (pp. 1-4). IEEE.

[10] Abd El-Jawad, M. H., Hodhod, R., & Omar, Y. M. (2018, December). Sentiment analy-sis of social media networks using machine learning. In 2018 14th international com-puter engineering conference (ICENCO) (pp. 174-176). IEEE.

[11] Shitole, A., & Devare, M. (2018). Optimization of person prediction using sensor data analysis of IoT enabled phys-ical location monitoring. Journal of Advanced Research in Dynamical and Control Systems, 10(9), 2800-2812.

[12] Kermani, F. Z., Sadeghi, F., & Eslami, E. (2020). Solving the twitter sentiment analysis problem based on a machine learning-based approach. Evolutionary Intelligence, 13(3), 381-398.

[13] Yadav, N., Kudale, O., Gupta, S., Rao, A., & Shitole, A. (2020, February). Twitter sen-timent analysis using ma-chine learning for product evaluation. In 2020 Internation-al Con-ference on Inventive Computation Technologies (ICICT) (pp. 181-185). IEEE.

[14] https://www.kaggle.com/crowdflower/twitter-airline-sentiment

[15] Mu, E., & Pereyra-Rojas, M. (2016). Practical decision making: an introduction to the Analytic Hierarchy Process (AHP) using super decisions V2. Springer.

[16] Abdualrhman, Mohammed Ahmed Ali, and M. C. Padma. (2019). “CS-IBC: Cuckoo search based incremental binary classifier for data streams.” Journal of King Saud Univer-sity-Computer and Information Sciences 31.3, (2019): 367-377.

[17] Yang, X. S., & Deb, S. (2014). Cuckoo search: recent ad-vances and applications. Neural Computing and Applica-tions, 24(1), 169-174.

[18] Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. Interna-tional journal of endocrinology and metabolism, 10(2), 486.

[19] McKnight PE, N. J. (2010, Jan 30). Mann-Whitney U Test. The Corsiniencyclopedia of psychology., 1-1.

[20] Budak H, T. S. (2016, Dec 20). A modified t-score for fea-ture selection. Anadolu Ün-iversitesi Bilim Ve Teknoloji Dergisi A-Uygulamalı Bilimler ve Mühendislik.,17(5), 845-52.

[21] Sahoo, P., & Riedel, T. (1998). Mean value theorems and functional equations. World Scientific.

[22] Z-table. (n.d.). Retrieved from http://www.sjsu.edu/faculty/gerstman/StatPrimer/z-table.PDF

[23] https://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/].

[24] Savoy. (2012, Jan). Feature selection in sentiment analysis. Proc. CORIA, no., 273-84.

[25] Rutkowski L, J. M. (2020). Misclassification Error Impuri-ty Measure. InStream Data Mining: Algorithms and Their Probabilistic Properties, 63-82.

[26] Xue Y, H. M. (2017). Active learning of classification models with likert-scale feedback. In Proceedings of the 2017 SIAM International Conference on Data Mining, (pp. 28-35).