I.J. Information Technology and Computer Science, **2024, 5, 23-35**
Published Online on October 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2024.05.02

A Machine Learning Approach for Sentiment Analysis Using Social Media Posts

Ritushree Narayan*
Magsdh University, Bodh Gaya, India
E-mail: 702004@gmail.com
ORCID iD: https://orcid.org/0000-0002-8204-8467
*Corresponding Author

Pintu Samanta
Mathematics department, Magadh University, Bodh Gaya, India
E-mail: pintu.math28@gmail.com

Received: 09 October 2023; Revised: 31 January 2024; Accepted: 11 July 2024; Published: 08 October 2024

Abstract: Sentiment analysis on Twitter provides organizations and persons with quick and effective instrument to observe the public's perceptions of them and their competition. A modest number of assessment datasets have been produced in recent years to check the efficiency of sentiment analysis algorithms on Twitter. Researchers offer a review of eight publicly accessible as well as manually annotated assessment datasets for analyzing Twitter sentiment in this research. As a result of this evaluation, we demonstrate that is a widespread weakness of many when using these datasets performing at sentiment analysis the objective (entity) level is indeed the absence of different sentiment classifications across tweets as well as the objects contained in them.[1], As an example all of that "I love my iPhone but I despise my iPad." Could be marked with a made-by-mixing classify however the object iPhone contained within this Twitter post should be annotated with just a label with an optimism. To get around this restriction and enhance existing assessment We have datasets that provide STS-Gold a novel assessment of datasets in which tweets or objects (entities) remain tagged separately hence might show alternative opinion labels. Though research furthermore compares the various datasets on multiple characteristics such as an entire quantity of posts as well as vocabulary size and sparsity.[2] In addition, look at pair by pair relationships between these variables and how they relate to sentiment classifier performance on various data. In this study we used five different classifiers and compared them and, in our experiment, we found that the bagging ensemble classifier performed best among them and have an accuracy level of 94.2% for the GASP dataset and 91.3% for the STS-Gold dataset.

Index Terms: Sentiment Analysis Twitter Datasets Bagging Ensemble Classifier Machine Learning.

1. Introduction

Creating evaluation sets of data that can be used to measure the performance of sentiment analysis techniques is required for their development. Numerous evaluation statistics for Twitter sentiment classification have been made public in recent years. The dataset for general evaluation is made up of a collection of tweets each analyzed with a sentimental tag [2]. Positive negative and neutral sentiment labels are the most commonly used however the evaluation of datasets takes extra sentiment labels including diverse other or immaterial. [3].

Several datasets provide arithmetic a sentiment strength of -5 to 5 ranging polarity shift from negative to positive as an alternative to the ultimate sentiment labels associated with the tweets. Evaluation datasets focus on providing sentiment labels associated with targets (entities) inside tweets in addition to sentiment labels connected with tweets. However, this dataset does not differentiate between both the sentimental label of somewhat such as the tweet and the sentiment labels things contained inside it [4]. The message tweeted "the iPhone is amazing but I am unable to upgrade:(" illustrates an unpleasant feeling. The entity "iPhone" on the other hand should be considered favorably. To address these limitations, we present STS-Gold a Twitter sentiment analysis evaluation dataset that concentrates on sentiment explanation on the tweet and entity levels. The explanation procedure enables again for labeling of tweets with differing polarities as well as the things included inside them. A sample of tweets from either the Stanford Twitter Sentiment Corpus [5] was selected for this dataset an additional entity extraction technique was employed to extract entities from this selection of tweets. Three distinct human evaluators manually annotated tweets and entities. The evaluation dataset is made up of 2206 twitter

posts and 58 items with sentiment labels.[6] The dataset's objective is to enhance present state data by providing entity sentiment labels and facilitating such assessment of sentiment classification algorithms there at entity and tweet levels.

This report summarizes eight freely available as well as manually analysed assessment datasets using Twitter sentiment analysis as well as the STS-Gold set of data characterization. The objective of all this study aims to give the reader an overview of current assessment datasets and their properties.[7] For that purpose, we examined these datasets along with a variety of variables including the average number of tweetslex is volume and data scarcity in general. [8]. We also look at the pair-wise correlations between these parameters and how they relate in relation towards the sentiment classification model across every set of data. Our findings show that the relationship between sparseness and categorization performance is intrinsic which means it may reside within the dataset itself. This isn't always the case across datasets. We also show that the connections exist. All of the relationships among sparseness lexical density and overall tweet number are robust. A large quantity of tweets is contained inside a dataset but on the other side is not always indicative of such a large dataset. A large vocabulary or even a top level of brevity. [9]. The remainder of paper is arranged as follows: Section 2 summarises the assessment data accessible for the Twitter sentiment study. Section 3 delves into STS-Gold our preferred assessment dataset. Section 4 compares and contrasts the two. In general, examine the assessment datasets.

2. Methodology

2.1. Datasets for Twitter Sentiment Analysis

This section covers eight separate datasets that are often used in the research on Twitter sentiment analysis. We selected datasets that are accessible online to the scientific research community (ii) meticulously annotate offering trustworthy judgments above tweets as well as (iii) use to evaluate several sentiment analysis approaches. [10].Tweets have also been decided to enter into such datasets are classified as either Negative Neutral Positive Mixed Other or Irrelevant. The overall proportion for tweets inside the eight datasets depending on all these sentiment categories is shown in Table 1.

The variations within assessment datasets are due to changes within sentiment analysis tasks. Twitter sentiment analysis difficulties include recognizing polarization (affirmative vs. pessimistic) subjectivity (polar vs. impartial) and overall sentimental zeal. These responsibilities can also be carried out at either the tweets as well as target (entity) level. [11]. The following sections provide an overview of accessible datasets for evaluation as well as the numerous sentiment actions through which they are employed.

Table 1. The total amount of messages and the distribution of tweet sentiment across all datasets

Dataset	No. of Tweets	#Negative	#Neutral	#Positive	#Mixed	#Other	#Irrelevant
STS-Test	499	175	138	183		0	0
RHC	2515	1380	471	540	46	0	79
OMD	3236	1195	0	711	244	1085	0
SS-Twitter	4243	1036	1952	1251	0	0	0
Sanders	5511	655	2504	571	0	0	1786
GASP	12770	5235	6268	1050	0	218	0
WAB	13340	2580	3707	2915	0	420	3718
SemEval	13975	2186	6440	5349	0	0	0

2.2. Set of Stanford Twitter Sentiment Tests (STS-Test)

Go et al. proposed their Stanford Twitter sentiment corpus which is divided into two groups: training and assessment. The preparation process included 1.7 million Twitter postings that were automatically classified as favorable or negative in nature on their emotional content. As an example, Twitter post is tagged optimistic if it contains :):-):D or =) or bad whether it contains :(:-(or: (.

Though automatic sentiment annotations of tweeting with emoticons are rapid their precision is essential and debatable as well as emotions may not adequately capture the overall Twitter sentiment. This inquiry focuses on meticulously annotated datasets. As a result, while we recognize the STS-trained dataset's potential for constructing sentiment analysis methods we omit that after the remainder of the study. This trial set (STS-Test) remains manual interpretation and consists of 177 negative 182 positive and 139 neutral tweeting's. These tweets are acquired by querying the Tweet API for precise terms such as product business and person names. The STS-Test dataset although its tiny size has been frequently utilized inside the literature for numerous evaluation purposes. Go et al. [12], Saif et al. Sperio su et al. and Bakliwal et.al. utilize this one to assess designs for divergence grouping (positive vs. negative). Marquez et al. utilize this dataset to assess subjectivity categorization through addition to classification and prediction (neutral vs. polar).

2.3. Reforming Health Care (RHC)

In March 2010 a Reforming Health Care (RHC) dataset is collected by crawling tweeter posts with the hashtag

"#RHC" (Reforming Health Care) [13]. The authors manually labeled a portion of this corpus using 5 tags (positive negative neutral irrelevant unsure(other)) and then divided this into education (839 tweets) development (838tweets) and test (837tweets) sets. In addition, the scientists assigned attitude labels to eight distinct goals drawn from three distinct datasets (Reforming Healthiness Care Obama Democrats Republicans Tea Party Traditional lists Liberals and Stupak). Though as observed in the printed form of such datasets (https://bitbucket.org/speriosu/updown) both the tweet and the targets inside it were labeled with the same emotion. We analyze these three subcategories (training development and testing) as either a single dataset in this work. The resulting dataset for example shown in Table 1 comprises 2517 Twitter posts including 1382 negative 470 neutrals and 540 positive Twitter posts. This RHC data has already been used to test polarities classification [14], or because it contains neutral Twitter posts it may also be used to check subjectivity classification.

2.4. Obama-McCain Debate (OMD)

An Obama-McCain Debate (OMD) data is created by scanning 3238 Twitter posts through the initial political TV discussion in the United States in September 2008 [15]. Amazon Mechanical Turk was used to collect sentiment labels for such tweets with each tweet classified either positive negative mixed or other through at least three annotators. [16], reported a 0.655 inter-annotator agreement showing relatively good agreement amongst annotators. The dataset as well as the votes of the annotators on each tweet are available at https://bitbucket.org/speriosu/updown. As the ultimate labels for the tweets, we picked personal sentiment labels and two-thirds of the voter agreed on that. As a result, there were 1,196 negative tweets and 710 positive Twitter posts with 245 mixed tweets.

2.5. Sentiment Strength Twitter Dataset (SS-Tweet)

These collection of data comprises 4242 Twitter posts that have been thoroughly tagged using positive and negative sentiment ratings. In other terms negative strength ranges from -1 (not negative) to -5. (very negative). Consequently, the optimistic strength is found between 0 and 1 (negative) to +5 (positive) (very positive). [17]. generated this dataset to assess SentiStrength (http://sentistrength.wlv.ac.uk/) the lexicon-based approach aimed at the strength of sentiment recognition. In the present work we proposed re-annotating Twitter posts in the dataset using sentiment labels (negative positive neutral) rather than sentiment strengths allowing this dataset to be utilized for both subjectivity categorization as well as sentiment strength recognition. For this specific purpose we allocate a single sentiment label for every post focused on two guidelines drawn from SentiStrength: (i)Twitter post remains measured unbiased in some way the absolute amount of its own negative to positive strength ratio corresponds to one; (ii) Twitter post seems to be positive if his positive sentiment strength stands 1.5 times greater than just his negative sentiment strength and negative otherwise. As shown in Table 1 the final dataset comprises 1,037 negative 1,953 neutrals and 1,251 positive Twitter posts. The unique data can be found at http://sentistrength.wlv.ac.uk/documentation/ also through 5 more datasets from MySpace Digg BBC forum Runners World forum and YouTube.

2.6. Sanders Twitter Dataset

Sanders' dataset takes 5510 tweets divided into four groups (Apple Google Microsoft and Twitter). Every post is carefully classified as positive unfavorable neutral or irrelevant to the problem by one annotator. The annotation technique produced 654 irrelevant negatives 2,503 neutrals 570 favorable also 1786 positive tweets. The dataset was used in [18], to categorize a collection of tweets based on their polarity and subjectivity. The Sanders dataset may be seen at http://www.sananalytics.com/lab/.

2.7. Twitter Dialogue Earth Corpus

Conversation of Earth Twitter dataset is organized into three categories. The first set is WA and the second set is WB have 4,491 and 8849 posts concerning the weather correspondingly whereas the third set (GASP) has 12,770 posts about gas prices. These datasets are developed as a part of the Dialogue Earth Project (www.dialogueearth.org) and are manually categorized having five tags: positive negative moderate unrelated and can't tell (other). In Section 4 we integrate two pairs of weather information into a single dataset (WAB). That resulted in 13340 posts among which 2580 are negative 3,707 are neutral and 2,915 are also positive. The GASP dataset comprises 5,235 negative Twitter posts 6,268 neutral twitter posts also 1,050 positive Twitter posts. WAB as well as GASP datasets are utilized to assess polarity classification machine learning algorithms (e.g. Naive Bayes SVM KNN) tweets [19].

2.8. SemEval-2013 Dataset (SemEval)

These datasets are produced for the job of Twitter sentiment analysis (Task 2) of a Semantic Assessment of Systems challenging tasks [20]. (SemEval-2013). SemEval's original dataset is made up of 20K tweets that are separated into three categories: training development and testing. Five Amazon Mechanical Turk employees meticulously labeled all the tweets as negative favorable or neutral. Turks were also questioned if Twitter phrases are subjective or objective. Utilizinga record of tweet identifiers from [21], we can obtain 13,975 Twitter posts encompassing 2186 negatives and 6440 neutrals with 5,349 positives. SemEval-2013 Task 2 participants This dataset was used to assess their systems Interpretation subjectivity identifying [22], and tweet-level sentiment polarity recognition. [23].

Summary: According to our findings there are two fundamental faults in such datasets when it's used to performance evaluation using Twitter sentiment models and algorithms. The first flaw is a lack of specifications for annotation technique used to tag twitter posts with sentiment using specific datasets (e.g. STS-Test RHC Sanders). [24], for example does not specify the number of annotators. Similarly, in [25], need not track annotation agreements between annotators. The second constraint is a majority of datasets are intended to assess the efficacy of sentiment analysis algorithms somewhere at the tweet level instead of the attribute level (i.e. Human segmentations on posts are included but not entities.). Only when the annotation process includes targeting entities such as with the RHC dataset are such entities assigned sentiment labels that seem to be comparable to the label of a tweet because they are relevant for. Interpretation of entity sentiment on the other hand is a critical topic meanwhile it is inextricably linked to the task of extracting the status of individuals besides corporations on Twitter.

2.9. STS-Gold Data

The suggested dataset STS-Gold is covered in the following sections. By offering a new sample in which tweets and entities were individually analyzed this suggested dataset seeks to enhance the existing Twitter sentiment classification assessment dataset. Permitting diverse sentiment tags to also be assigned to the tweets and items included inside it. The purpose is to help only with the performance assessment of entity-based sentiments modeling algorithms which would at present be underserved by the datasets available (see Section 2).

3. Data Acquisition

In the direction of create the dataset we initially take out all named items as of a sample of 180K tweets chosen at random from the Stanford Twitter corpus (see Section 2). We utilized Alchemy API an online tool that facilitates the extraction of things from the text as well as their associated semantic idea class to do this (e.g. Person Company City). Then comes we found the top two most common semantic concepts and for each of them the top two most common and two less common entities for example researchers can choose the two most commonly used entities for such semantic notion of Person. (Taylor Swift and Obama) as well as two mid-frequent entities (Oprah and Lebron). As illustrated in Table 2 this resulted in 28 distinct entities and their seven linked ideas.

Table 2. STS-Gold was built using 28 Entities and their semantic concepts

Concept	Top 2 Entities	Mid 2 Entities
Person	Taylor Swift Obama	Oprah Lebron
Company	Facebook Youtube	Starbucks Mc-Donald's
City	London Vegas	Sydney Seattle
Country	England US	Brazil Scotland
Organization	Lakers Cavs	Nasa UN
Technology	iPhone iPod	Xbox PSP
health condition	Headache Flu	Cancer Fever

The final step is to create in addition to construct a group containing tweets for sentiment explanation confirming that each post inside this collection had unique or either of this same 28 entities in Table 2. To do so researchers randomly pick 101 twitter posts starting residue left STS corpus for each of the 29 entities for the sum of 2,800 tweets. We then add 200 more tweets which made no definite reference to every entity for just a sum of 3,000 tweets. Then on the 3,000 Twitter posts we picked we utilized the Alchemy API. Sidewise the first 28 items the extraction tool produced 119 more for a total of 147 entities for 3,000 tweets.

Data Annotation

Three graduate students have been assigned the task of systematically classifying every 3000 tweets into five different categories: (Negative Positive Neutral Mixed and Other). Tweets expressing conflicting emotions earned the "Mixed" label whereas those who couldn't decide to receive the "Other" label. The students were also expected to characterize all things in a tweet using five sentiment classifications. The students were passed a leaflet outlining their Twitter posts and entity-level annotating responsibilities. As indicated in the document's appendix the brochure also offers a list of crucial instructions. It's worth noting that the analysis was completed with the help of Tweenator an online tool as well as researchers earlier generated compiled Twitter posts [26]. For such tweet-level annotation job researchers appreciated an inter-annotation agreement of 0.765 using Krippendorff's alpha measure [27]. We only got e = 0.416 when we analyzed the overall sentiment of the entity for each unique tweet for an entity-level annotation work which is rather insufficient for annotated data to be utilized. When we looked at aggregated emotion for an object, we discovered an exceptionally high inter-annotator agreement of e = 0.955.

We picked Twitter posts and things about which all three annotators agreed on sentiment classifications to generate the final STS-Gold dataset removing potentially erroneous information from the dataset. The STS-Gold dataset as shown in Table 3 includes 14 negative 26 positive and 17 neutral entities derived from 1401 negative 631 positive as well as 78

neutral Twitter posts (tweets). This same STS-Gold dataset consists sentiment labels for both twitter posts and entities making it possible to compare twitter post versus entity-based Tweet sentiment analysis algorithms.

Table 3. Shows how many twitter posts as well as entities of every type there are

Class	Negative	Positive	Neutral	Mixed	Other
No. of Entities	14	26	17	0	0
No. of Tweets	1401	631	78	91	5

4. A Comparison of The Twitter Sentiment Analysis Dataset

In this section we compare the dataset presented above on three diverse dimensions: vocabulary size the overall amount of twitter posts and data sparsity. We also explored inherent pair-wise relationship among such features and how this corresponds to sentiment classification model (correlations are computed by utilizing the Pearson correlation coefficient). For this purpose, we use an Optimum Entropy classifiers to undertake binary classification method (positive vs. negative) throughout all datasets (MaxEnt). It is worth noting that no data separation or filtration was carried out because the comparison's goal isn't to create strong classifiers. In place of this researcher seek to highlight distinct properties of the dataset and that variables may influence sentiment classifier performance.

4.1. Vocabulary Size

The number of different word unigrams in a dataset is used to calculate its vocabulary size. To count the number of unigrams we utilize a Tweet NLP tokenizer [28], which has been created specifically for usage with Twitter data. We take into account all tokens in tweets such as phrases integers URLs emojis or even special typescripts (e.g. question marks intensifiers hashtags etc). Figure 1 depicts the relationship between the size of the vocabulary and the total number of twitter posts in the dataset. Even though the relation between the two variables appears positive (= 0.95) boosting the number of Twitter posts (does not always) increase vocabulary size. An OMD dataset for example has much more Twitter posts than the RHC dataset while having a smaller vocabulary size.

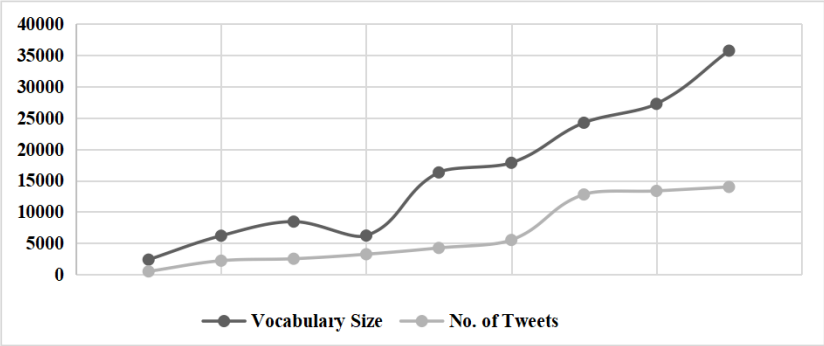

Fig.1. Showing the relationship between vocabulary size and No. of tweets

4.2. Sparsity of Data

Dataset sparsity has a significant impact on the overall performance of common machine learning classifiers. Tweet data as per Saif et al. are sparser than other forms of data (e.g. movie review data) because of the significant number of uncommon terms in tweets. In this part we will compare the sparsity of the datasets supplied. We use the following formula from to compute the sparsity degree of a given dataset:

$$S_d = 1 - \frac{\sum_i^n Ni}{n \times |V|} \quad (1)$$

where Ni represents the number of different words in the tweet I n represents the number of twitter posts inside the dataset as well as |V | denotes the dataset size. of the vocabulary. It is also worth noting that the sparsity degree has a substantial association with the entire number of the Twitter post contained inside datasets ($\rho = 0.71$) and an even stronger link with the dataset's vocabulary size ($\rho = 0.77$).

5. Classification

5.1. Maximum-entropy Learning

A fundamental concept in Bayesian statistics the principle of Maximum-Entropy was put forth by Jaynes. It states that given scientifically verifiable information the probability distribution with the highest entropy best captures the state

of knowledge (such as accuracy).

Numerous scientific fields have investigated this concept including statistical mechanics Bayesian inference unsupervised learning and reinforcement learning.

The regularization technique that implements a penalty for minimum entropy predictions has been explored in semi-supervised learning and deterministic entropy annealing for vector quantization. The regularization of the entropy of classifier weights has been used empirically and theoretically in machine learning. The random variable $X=\{x_0\ x_1\ \ldots x_n\}$ and the entropy for the random variable is given by:

$$\text{Entropy}\ \ H(X) = -\sum_{i=1}^{n} P(xi) \log 2\big(P(xi)\big) \tag{2}$$

The equation for Maximum Entropy is given by:

$$\pi^{*} = \text{argmax}\pi\theta E\pi\theta\ [\ \sum_{t=0}^{T-1} \gamma R(s_t\ a_t) + \alpha H(\pi(.|S_t))] \tag{3}$$

5.2. K-nearest Neighbor (KNN)

A prevalent distance-based algorithm for solving classification and regression issues is KNN which is non-parametric. A new instance is classified by contrasting it with the instances in the training set that are the most similar as KNN is also an instance-based learning algorithm. Distance measurements can estimate how similar the instances are to one another. The K value or the number of nearest neighbors is the main deciding factor in this classifier. If K is 1 the new instance is simply assigned to the class of its closest neighbor. In KNN several different distance measures including the Manhattan Distance Minkowski Distance Hamming Distance Euclidean Distance etc. can be applied. In this study the Minkowski Distance function was used.

5.3. Adaptive Boosting (AdaBoost)

A technique for ensemble learning is an adaptive boosting classifier also referred to as the AdaBoost classifier. It uses several weak classifiers to build a strong classifier. In this method a weak classifier picks up information from the mistakes of its stronger counterpart. Think of a sampled dataset with *n* samples. Each sample is initially given a weight of *(1/n).* A weak classifier is created using this dataset. This classifier's overall error is calculated. And using this total error the following equation calculates the effectiveness of this classifier in categorizing the data samples:

$$\alpha = \frac{1}{2} ln\left(\frac{1-\epsilon}{\epsilon}\right) \tag{4}$$

Here α is used to alter the dataset's samples' weights producing a new dataset. Another weak classifier is later built using this new dataset. Finally, the class label for a sample is decided by the weak classifiers who received the most votes.

5.4. Gradient Boosting (GB)

The Gradient Boosting (GB) classifier successively builds new models out of a sequence of weak models. Every new model makes an effort to reduce the loss function. GB calculates the loss function using the gradient descent technique. Boosting must behalted promptly with discontinuing criteria to prevent overfitting issues. A cap on the total number of models that can be developed or a cutoff point for predicted accuracy can serve as the stopping point.

Let us consider a function f: $R^n \rightarrow R$ of a linear combination of weak models g= $R^n \rightarrow R$ and the equation becomes:

$$f(x) = \sum_{i=1}^{M} .w_i.g_i(x;\ \theta i) \tag{5}$$

In this case $x \in R^n$ stands for the input vector. $M.w_i \in R^n$ indicates the weight which represents the number of weak models. The function is created by iteratively choosing a weak learner's parameter θi and weight w_i to minimize the loss function.

5.5. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting enhances Gradient Boosting Decision Tree's speed and scalability (GBDT) [29]. The regularization and loss functions are combined to create the XGBoost goal function. The following equation represents the objective function of XGBoost:

$$I(\theta) = L(\theta) + \Omega(\theta) where L(\theta) = l(zy)\ \ \Omega(\theta) = \alpha T + \left(\frac{1}{2}\right)\lambda||\omega||2 \tag{6}$$

In the above equation $L(\theta)$ represents the loss function and the regularisation parameters $\Omega(\theta)$ learned from the supplied data. T and respectively stand for learning frequency the number of leaves the weightiness of the leaves and the regularized parameter. If the expected output is the actual output, then either the mean square loss function or the logistic loss function can be used to calculate L. The following equation can be used to express L using the mean square loss

function:

5.6. Ensemble Bagging Classifier

The concept of bagging is based on the ideas of bootstrapping and aggregating. Bootstrap datasets are created from the training dataset for the Bagging classifier. Then various classifiers are trained using each of these bootstrap datasets. The final prediction is obtained by combining the output of several classifiers. The bootstrap dataset frequently avoids using misleading training items. The flow chart of the proposed system is shown in figure 2.

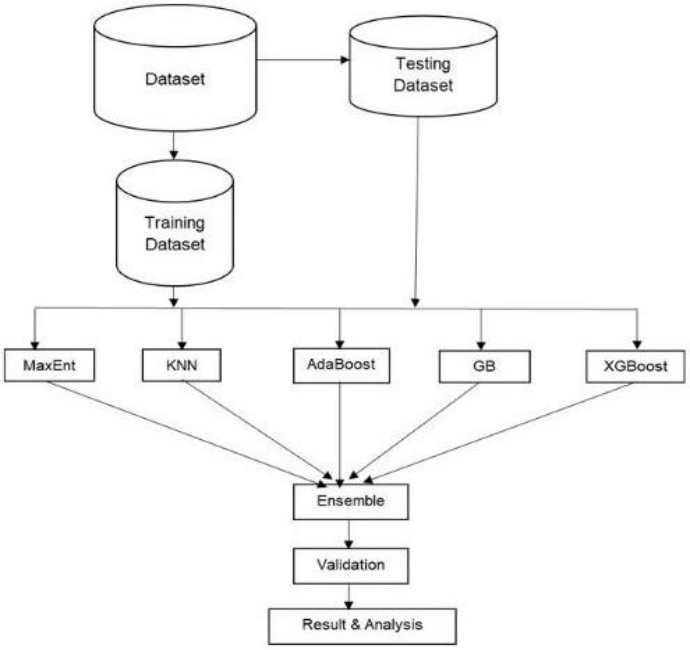

Fig.2. Showing flow chart of the proposed method

Additionally grouping the classifiers generally results in performance gains over using a single classifier. The Bagging classifier frequently performs better than the other classifiers since it combines these properties. Following are the steps that bagging takes:

Step 1: Bootstrapping: for n= 1 2 … N:
a. Bootstrapped samples X^n created from the practice set X.
b. Using X^n construct a model $B^n(x)$.
Step 2: Aggregation:
Aggregating the results model $B^n(x)$ where n= 1 2 … N by using the majority voting method.

6. Result And Analysis

The entire study was carried out on a system with a Windows™ 10 Pro operating system Intel® Core™ i7-8550U CPU with RAM storage of 16 GB. The coding part of the research work was carried out in Python language. We implemented the complete study in this study we implemented and analyzed the performance of each of the six classifiers using the test dataset and compared them using the five metrics provided in terms of accuracy sensitivity specificity Precision and F-Score.

Here in our research, we implemented all five classifiers and an ensemble classifier on all nine datasets and as a result we found that the bagging ensemble classifier has an accuracy rate of 94.2% for GASP dataset which is the maximum we achieved. The accuracy can be calculated by using the formula given:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \text{x } 100 \quad (7)$$

For the STS Gold dataset bagging ensemble classifier has 91.3%followed by AdaBoost with 88.6%whereas KNN has the lowest accuracy of 71.6% only as shown in table 4 and figure 4.

Table 4. Showing accuracy

Classifiers	DATA SET								
	STS GOLD	STS TEST	RHC	OMD	SS Twitter	Sanders	GASP	WAP	SemEval
KNN	71.6	69.8	68.7	72.1	70.9	78.9	83.7	81.1	79.4
MaxEnt	85.7	80.1	78.7	82.8	73.9	84.3	90.9	85.2	83.6
AdaBoost	88.6	86.9	85.4	87.3	87.2	87.9	91.1	88.1	87.9
GB	86.9	86.2	85.2	86.7	85.4	87.2	90.8	87.2	87.1
XGBoost	86.1	84.8	83.9	86.7	85.3	87.2	90.1	87.1	86.9
Ensemble	91.3	88.9	88.7	89.1	88.6	89.4	94.2	90.4	89.7

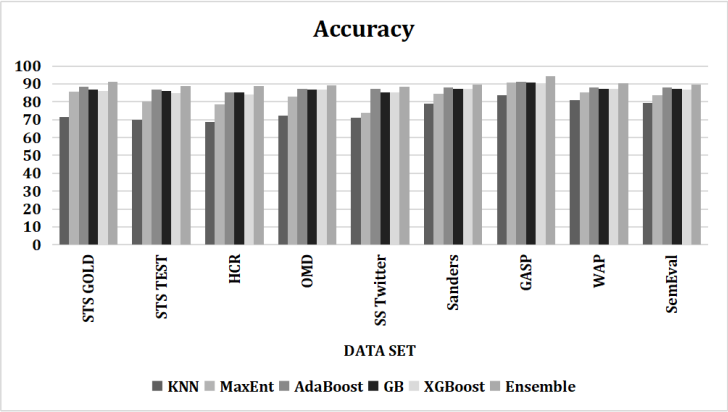

Fig.3. Showing accuracy

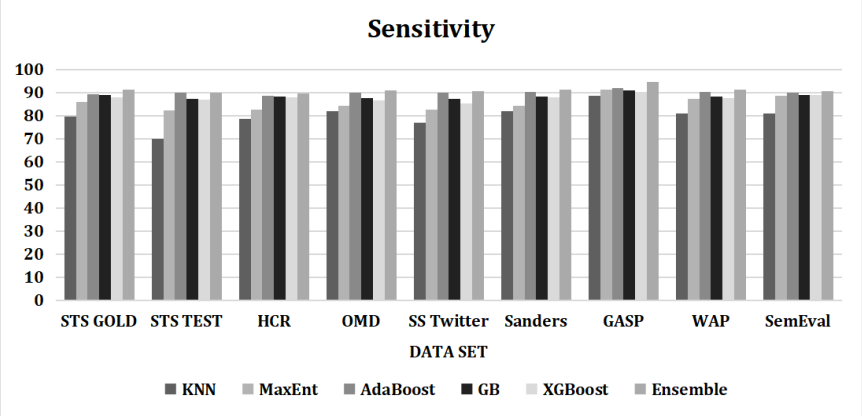

Fig.4. Showing sensitivity

Sensitivity can be explained as the ability of a classifier to identify the occurrences correctly and it can be calculated by using equation 8.

$$Sensitivity = \frac{TP}{TP+FN} \text{ x } 100 \tag{8}$$

All six classifiers are fed with the different datasets and we observed that all the classifiers worked well for the GASP dataset and the bagging ensemble classifier showed a sensitivity level of up to 94.7%. The bagging ensemble classifier's performance is approximately similar for datasets STS Gold OMD Sanders WAP and RHC around 91%. The performance of GB and XGBoost classifiers are nearly the same 88.9% and 88.1% for the STS Gold dataset whereas the performance of GB and Adabost are roughly the same 88.2% and 88.7% for the RHC dataset as shown in figure 5 and table 5.

Specificity can be explained as the ability of a classifier to correctly identify tweets with negative emotion which is shown in equation 9.

$$Specificity = \frac{TN}{TN+FP} \text{ x } 100 \tag{9}$$

Table 5. Showing sensitivity

Classifiers	DATA SET								
	STS GOLD	STS TEST	RHC	OMD	SS Twitter	Sanders	GASP	WAP	SemEval
KNN	79.6	69.8	78.7	82.1	76.9	81.9	88.7	81.1	81.1
MaxEnt	85.9	82.4	82.7	84.3	82.7	84.3	91.2	87.3	88.6
AdaBoost	89.4	89.9	88.7	89.9	90.1	90.3	92.1	90.3	90.1
GB	88.9	87.2	88.2	87.7	87.4	88.2	90.8	88.2	89.1
XGBoost	88.1	86.8	87.9	86.7	85.3	87.8	90.1	87.7	88.9
Ensemble	91.3	89.9	89.7	91.1	90.6	91.4	94.7	91.2	90.7

The performance of the XGBoost classifier for the computation of sensitivity seems to be very stable for all the datasets and performed best for GASP and SemEval datasets with 89.7% and 87.9% respectively. Figure 6 and table 6 show the performance of each classifier

Table 6. Showing specificity

Classifiers	DATA SET								
	STS GOLD	STS TEST	RHC	OMD	SS Twitter	Sanders	GASP	WAP	SemEval
KNN	80.6	79.8	78.7	82.3	78.9	81.2	88.7	81.2	81.1
MaxEnt	86.1	83.4	83.2	84.3	82.7	85.1	90.6	87.8	88.2
AdaBoost	91.4	90.9	89.3	89.9	90.1	90.8	93.1	90.9	90.7
GB	89.9	87.9	88.2	87.9	87.4	88.6	91.1	90.2	89.9
XGBoost	88.7	87.2	87.9	88.7	87.3	87.8	89.7	87.7	87.9
Ensemble	92.8	91.5	90.7	91.1	90.6	91.4	94.7	93.2	91.7

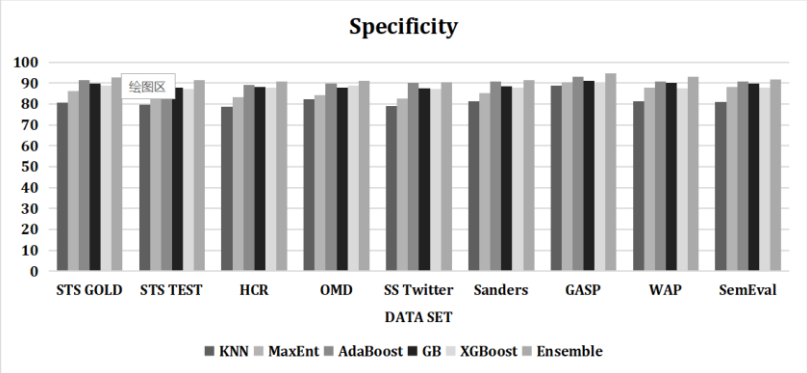

Fig.5. Showing specificity

Table 7. Showing precision

Classifiers	DATA SET								
	STS GOLD	STS TEST	RHC	OMD	SS Twitter	Sanders	GASP	WAP	SemEval
KNN	80.9	79.9	79.1	82.3	79.1	81.3	89.1	82.2	81.4
MaxEnt	86.4	83.6	84.8	84.9	83.9	85.2	90.5	88.5	88.3
AdaBoost	91.5	91.2	89.5	89.9	90.3	90.9	93.7	91.2	90.8
GB	89.9	88.2	88.2	87.8	87.9	88.3	91.6	90.3	89.9
XGBoost	88.9	87.5	87.9	89.7	88.3	87.6	89.9	88.7	89.5
Ensemble	92.9	91.6	91.4	91.1	90.8	91.5	94.7	93.3	91.9

The number of positive class forecasts that fall within the positive class is measured by precision which is shown in equation 10.

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

Figure 7 clearly shows that the bagging ensemble method outperformed other classifiers and it has above 90% for all the nine datasets which is appreciable. Interestingly the performance of XGBoost and AdaBoost classifier is almost

similar in the case of OMD dataset whereas the difference in performance is noticeable specially in the case of STS Test and GASP datasets.

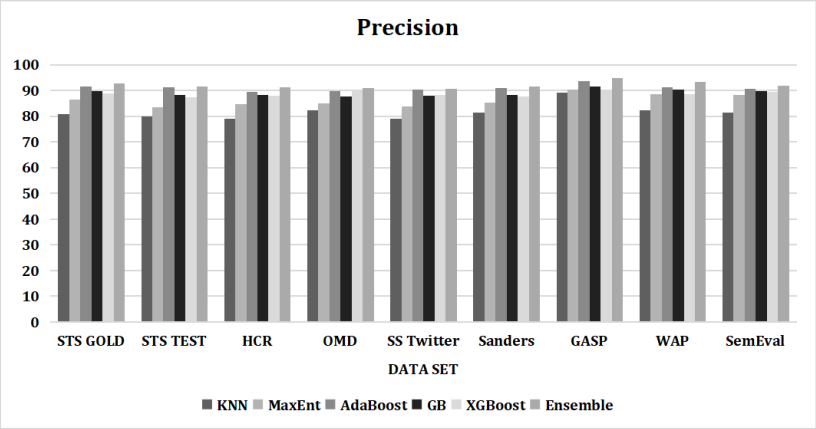

Fig.6. Showing precision

F1 Score takes into account recall and precision. It represents the harmonic mean (average) of recall and precision as shown in equation 11.

$$F1\ Measure = 2 \text{ x } \frac{(Precision \text{ x } Sensitivity)}{(Precision + Sensitivity)} \tag{11}$$

In this study we found that the F-Score of the bagging ensemble classifier outperformed the other classifiers with a maximum F-score of 94.3% for GASP dataset followed by AdaBoostwith 93.6% as shown in figure 8 and table 8.

Table 8. Showing F score

Classifiers	DATA SET								
	STS GOLD	STS TEST	RHC	OMD	SS Twitter	Sanders	GASP	WAP	SemEval
KNN	81.1	79.2	79.4	81.3	79.1	80.3	81.1	79.8	79.4
MaxEnt	86.5	83.6	84.4	84.6	83.8	85.2	89.5	88.1	88.1
AdaBoost	91.2	91.2	89.6	89.7	90.2	90.3	93.6	91.1	90.6
GB	89.7	88.1	88.2	87.7	87.9	88.1	91.4	90.2	89.3
XGBoost	88.3	87.4	87.7	89.4	88.3	87.4	89.2	88.4	89.5
Ensemble	92.8	91.5	91.3	91.5	90.4	91.6	94.3	93.1	91.7

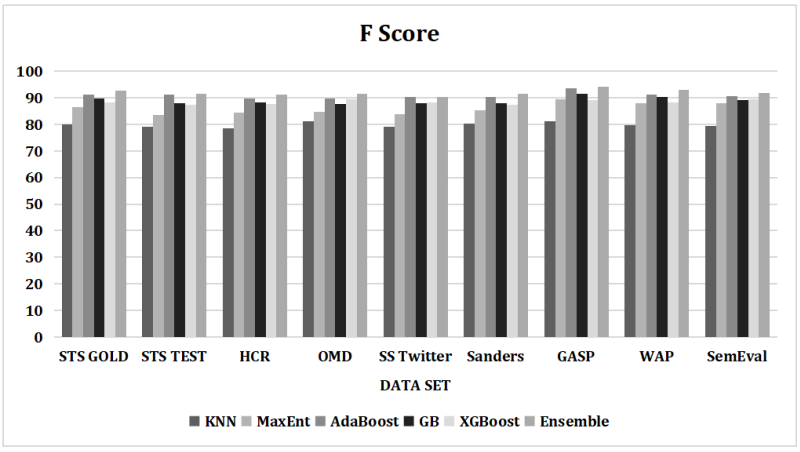

Fig.7. Showing F score

The MaxEnt classifier for the GASP dataset obtains the highest average F-measure 84.621% whereas WAB dataset has the best accuracy of 90.897%. It is also worth mentioning that the dataset's distribution of positive and negative tweets has a considerable influence on per-class performance. F-measure recognizing positive tweets (F-positive) bigger than F-measure detecting negative tweets (F-negative) with the positive dataset (i.e. datasets using more positive tweets

than negative tweets) like STS-Test SS-Twitter WAB and SemEval (F-negative). Similarly for negative datasets the F-negative score is greater than the F-positive score (those with more negative tweets than positive tweets). Moreover, negative datasets have an average accuracy of 84.53 % whereas positive Twitter posts have an accuracy of 80.37% detecting positive tweets is more difficult than detecting negative tweets. Makrehchi and Kamel show that the sparseness degree of the dataset may be utilized to measure text classifier performance trends. They observed that lowering the overall sparsity of a particularly given dataset increases SVM classifier performance. Their conclusions are focused on altering sparsity degree of a similar dataset by eliminating/retaining certain expressions. On the other hand, the dataset sparsity is consistent across all datasets. As the results show there is almost There is no link between classification outcomes with sparsity degree across datasets (acc = 0.06 f1 = 0.23). This sparsity-performance relationship is intrinsic which means it happens inside the dataset rather than across datasets. This is not surprising given that as noted previously in this section other dataset features such as the polarity classification process may affect overall performance as well as data sparsity.

7. Conclusions

In this research paper we offered an evaluation of eight openly available and properly annotated evaluation dataset samples of Twitter sentiment analysis. Unlike at the tweet level we discovered that comparatively little annotation efforts were dedicated toward building datasets for testing sentiment classification at the attribute level. As a result, we developed STS-Gold a novel testing dataset that allows us to evaluate sentiment classification algorithms on both the entity and tweet labels. Our dataset unlike others differentiates between both the sentiment of Twitter posts and the sentiment of items mentioned in them.

We similarly conducted a comparative study across all published datasets in terms of vocabulary size the overall number of posts and degree of sparsity. We looked at multiple pair-wise relationships between factors and the relationship between data sparsity with sentiment classification performance across datasets. Although there is a strong correlation between these two characteristics our investigation discovered that a large number of tweets within the dataset is not always predictive of both large vocabulary sizes. We also show that the sparsity-performance relationship is intrinsic which means it exists inside the dataset but is not necessary across datasets.

Conflict of Interest

The authors have no conflict of interest to declare.

Funding Declaration

No particular grant was given to this research by any funding organization in the public, private, or non-profit spheres.

Acknowledgement

The authors would like to thank the authors of previous studies for making datasets publicly available for the current effort.

References

[1] Asiaee T A. Tepper M. Banerjee A. Sapiro G.: 2012If you are happy and you know it... Tweet. In: (2012) Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 1602–1606. ACM.

[2] Bakliwal A. Arora P. Madhappan S. Kapre N. Singh M. Varma V.: 2012Mining sentiments from tweets. Proceedings of the WASSA 12.

[3] Bravo-Marquez F. Mendoza M. Poblete B.: 2012 Combining strengths emotions and polarities for boosting Twitter sentiment analysis. In: 2013 Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM 2013.

[4] Chalothorn T. Ellman J.:2013Top: Using Twitter to analyze the polarity of contexts. In: In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013) Atlanta Georgia USA June 2013.

[5] Deitrick W. Hu W.: 2013 Mutually enhancing community detection and sentiment analysis on Twitter networks. Journal of Data Analysis and Information Processing 1 19–29 2013.

[6] Diakopoulos N. Shamma D.: 2010 Characterizing debate performance via aggregated Twitter sentiment. In: 2010 Proceedings of the 28th international conference on Human factors in computing systems. ACM.

[7] Gimpel K. Schneider N. O'Connor B. Das D. Mills D. Eisenstein J. Heilmanet.al. :2010Part-of-speech tagging for Twitter: Annotation features and experiments. Tech. rep. DTIC Document.

[8] Go A. Bhayani R. Huang L.: 2009Twitter sentiment classification using distant supervision.CS224N Project Report Stanford.

[9] Hu X. Tang J. Gao H. Liu H.: Unsupervised sentiment analysis with emotional signals. In: (2013) Proceedings of the 22nd international conference on the World Wide Web. pp. 607–618. International World Wide Web Conferences Steering Committee.

[10] Hu X. Tang L. Tang J. Liu H.: Exploiting social relations for sentiment analysis in microblogging.In: 2013 Proceedings of the sixth ACM international conference on Web search and data mining. pp. 537–546. ACM.

[11] Krippendorff K.: 1980 Content analysis: anintroduction to its methodology.
[12] Liu K.L. Li W.J. Guo M.:2012 Emoticon smoothed language models for Twitter sentiment analysis. In: AAAI.
[13] Makrehchi M. Kamel M.S.: 2008 Automatic extraction of domain-specific stopwords from labeled documents. In: Advances in information retrieval pp. 222–233. Springer (2008).
[14] Martınez-C´amara E. Montejo-R´aez A. Martın-Valdivia M. Urena-L´opez L.: Sinai 2013 Machine learning and emotion of the crowd for sentiment analysis in microblogs (2013).
[15] Mohammad S.M. Kiritchenko S. Zhu X.: Nrc-canada: Building the state-of-theart in sentiment analysis of tweets. In: 2013In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013) Atlanta Georgia USA June 2013. (2013).
[16] Nakov P. Rosenthal S. Kozareva Z. Stoyanov V. Ritter A. Wilson T.: 2013 Semeval2013 task 2: Sentiment analysis in Twitter. In2013 In Proceedings of the 7th International Workshop on Semantic Evaluation. Association for Computational Linguistics. (2013).
[17] P. Vasilić S. Vujnović N. Popović A. Marjanović and Ž. Đurović2018 "Adaboost algorithm in the frame of predictive maintenance tasks" 2018 23rd International Scientific-Professional Conference on Information Technology (IT) 2018 pp. 1-4 doi: 10.1109/SPIT.2018.8350846.
[18] Rahman S Irfan M Raza M Moyeezullah Ghori K Yaqoob S Awais M. 2020 Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living.International Journal of Environmental Research and Public Health. 2020; 17(3):1082. https://doi.org/10.3390/ijerph17031082.
[19] Z. Chen F. Jiang Y. Cheng X. Gu W. Liu and J. Peng2018"XGBoost Classifier for DDoS Attack Detection and Analysis in SDN-Based Cloud" 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) 2018 pp. 251-256 doi: 10.1109/BigComp.2018.00044.
[20] Phan X.H. Nguyen L.M. Horiguchi S.: 2008 Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on the World Wide Web. pp. 91–100. ACM (2008).
[21] Remus R.: Asvuniofleipzig: 2013 Sentiment analysis in Twitter using data-driven machine learning techniques (2013).
[22] Saif H. He Y. Alani H.:2011Semantic Smoothing for Twitter Sentiment Analysis. In: Proceeding of the 10th International Semantic Web Conference (ISWC) (2011).
[23] Saif H. He Y. Alani H.: 2012 Alleviating data sparsity for Twitter sentiment analysis. (2012) In: Proceedings 2nd Workshop on Making Sense of Microposts (#MSM2012) in conjunction with WWW 2012. Layon France (2012).
[24] Saif H. He Y. Alani H.: 2012 Semantic sentiment analysis of Twitter. In: (2012) Proceedings of the 11th international conference on The Semantic Web. Boston MA (2012)
[25] Shamma D. Kennedy L. Churchill E.:2009Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the first SIGMM workshop on Social media. pp. 3–10. ACM (2009)
[26] Speriosu M. Sudan N. Upadhyay S. Baldridge J.: 2011Twitter polarity classification with label propagation over lexical links and the follower graph.In: (2011) Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP. Edinburgh Scotland (2011)
[27] Thelwall M. Buckley K. Paltoglou G.: 2012Sentiment strength detection for the social web.Journal of the American Society for Information Science and Technology 63(1) 163–173 (2012)
[28] Thelwall M. Buckley K. Paltoglou G. Cai D. Kappas A.: 2010 Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12) 2544–2558 (2010).
[29] Ritushree Narayan, Kesav Sinha "An Analysis and Detection of Misleading Information on Social Media Using Machine Learning Technique (051121-031256)" Published by IGI global https://www.igi-global.com/chapter/an-analysis-and-detection-of-misleading-information-on-social-media-using-machine-learning-techniques/299200

Authors' Profiles

Dr. Ritushree Narayan: Research scholar, Mathematics department, Magadh University, bodh Gaya, Gaya. Dr. Ritushree Narayan finished her Ph.D. degree from Bundelkahnad university, MP. Research Interest: Quantum Computing, Natural language processing, Machine Learning, and Application of Computer in other areas.

Dr. Pintu Samanta: Assistant Professor, Mathematics department, Magadh University, Bodh Gaya, India. Research Interest: Mathematics.

How to cite this paper: Ritushree Narayan, Pintu Samanta, "A Machine Learning Approach for Sentiment Analysis Using Social Media Posts", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.5, pp.23-35, 2024. DOI:10.5815/ijitcs.2024.05.02