
ISSN: 2319-345X
IJMRBS
International Journal of
Management Research and
Business Strategy
www.ijmrbs.org
E-mail
editor@ijmrbs.org
editor.ijmrbs@gmail.com

International Journal of Management Research and Business Strategy

ISSN 2319-345X www.ijmrbs.com

Vol. 10, Issuse. 2, April 2022

Sentiment Analysis of Web Discourse Data using Ensemble Classifiers

M.Govindarajan

Abstract— It's popular for people and organizations to post their thoughts on blogs, microblogs, review sites, Twitter, Facebook, etc. Opinion mining can be employed to glean useful information from a text. When it comes to novel ensemble classification methodologies, the accuracy of classifiers like arcing and bagging is examined to see how well they work. Naive Bayesian Bayes (NB), Support Vector Machines (SVM), and Genetic Algorithms (GA) are the three most common categorization algorithms (GA). Blogs, newsgroups and Twitter are used to show the practicality and advantages of the recommended approaches.. Preprocessing, document indexing, feature reduction, classification, and aggregation are the five important parts of the suggested approach. Data from blogs, news groups, and Twitter are regularly the focus of comparison studies in the academic community. Blogger, newsgroup, and twitter data sets were used to compare the accuracy of simple classifiers using homogenous and heterogeneous models. As compared to individual classifiers, the proposed ensemble techniques provide a significant gain in accuracy, and heterogeneous models perform better than homogeneous models for the datasets of blogs, newsgroups, and Twitter, respectively.

Index Terms—Precision, Arcing, Bagging, Genetic Algorithm, Naive Bayes, Sentiment Mining, SVM.

INTRODUCTION

1.1 Sentiment analysis and natural language processing go hand in hand. For example, it tracks how individuals feel about a product or service they're utilizing. People can express themselves and get feedback through various online mediums, including blogs, review sites, and social networking sites like Twitter and Facebook. Sentiment analysis and opinion mining can be used to obtain and analyze customer feedback on a product or service.

1.2 **BloggerData**

Businesses and individuals are increasingly turning to social media content (forum l decision-making. Use sentiment analysis if you want to know what people think about a specific issue. Many documents and languages, such as books, movies, and even products, are difficult to categorize because they might be classed positively or negatively (Buche, A et al., 2013). Since there are so many websites out there, it is difficult to sift through them all for helpful data. People are flooding the internet with knowledge, whether it's in the form of long blogs or forum posts. Web page browsing can be tedious and time-consuming for the ordinary human, especially if you're looking for something specific. This necessitates the use of automated sentiment

AssistantProfessor,DepartmentofComputerScienceandEngineeringAnnamalaiUniversity,AnnamalaiNagar,TamilNadu,India
e-mail:govind_aucse@yahoo.com

analysis tools. Liu B. (2012) provides more information on opinion mining, which seeks to extract the attributes and components of an item that have been commented on in documents. Making decisions often necessitates soliciting the input of others, and this is why hearing what others have to say about a topic is so valuable. Individuals and organizations alike can benefit from this. The primary goal of opinion mining is to understand people's feelings, attitudes, and opinions by analyzing their sentiments (Poornima Singh et al., 2015).

1.3 **Newsgroupsdata**

A major aspect of document processing is document classification, which is becoming increasingly important due to the explosion of text documents available via the internet and in the media. Only a few of the various applications of document classification were indexed, analyzed, filtered, distributed, and archived (F. Sebastiani, 2002 and N. Chen et al., 2007). Because it is more efficient, automated categorization is preferred to manual accurate and time-efficient, especially when dealing with large volumes of information (N. VasfiSisi et al., 2013 and Nidhi et al., 2011). Documents are automatically classified using natural Approaches to data mining, machine learning, and language processing.

There are a variety of machine learning approaches that can be applied to the classification of documents. Some examples of these are the Bayesian classifier, decision tree, K-nearest neighbor (SVM), support vector machines (SVM), genetic algorithm, and genetic programming (Nidhi et al. 2011, Bhumika, S. S. Sehra et al., 2013). (GP). In supervised machine learning, GP is a regularly used evolutionary algorithm for computer programs (D. Kumar et al., 2013). A solution to a problem can be expressed using the GP's functions and terminals.TwitterData

Using Twitter, you may reach people all over the world and spread awareness about a wide range of topics and issues. Healthcare, politics, or education are just a few examples. Twitter's massive and unstructured data contains a wealth of useful information. It's no secret that Twitter is an increasingly popular place for people to air their views on a wide range of topics. It is possible to use this microblogging technology to keep people informed about important occurrences. Tracking internet thoughts and attitudes and evaluating whether or not the general public likes or dislikes them is done through sentiment analysis. Analysis of feelings. Unstructured (textual) data can be mined for useful numerical indexes using text mining techniques. It's possible to learn more about the dataset this way. a year later (Verzani, 2011).

Using Twitter, you may reach people all over the world and spread awareness about a wide range of topics and issues. Healthcare, politics, or education are all possibilities. Twitter's massive and unstructured data contains a wealth of useful information. It's no secret that Twitter is an increasingly popular place for people to air their views on a wide range of topics. Use this microblogging tool to keep the public informed about important occurrences. It is possible to use sentiment analysis to find out whether or not a piece of content's intended audience likes or dislikes the ideas and attitudes expressed therein. Analysis of feelings. Unstructured (textual) data can be mined for useful numerical indexes using text mining techniques. As a result, the dataset can be better understood. the subsequentyear (Verzani, 2011).

2.1 **BloggerData**

I. **RELATEDWORK**

It is common for blogs to be published in a timely manner and to include people's thoughts and feelings. Because of these capabilities, a great deal of effort has been put into using blogs to gather trends, opinions, or emotions on a wide range of topics.

As the internet becomes more extensively utilized, blogging and blog pages are quickly becoming the most popular means of

expressing one's opinions. The term "blogosphere" refers to the entire universe of blog sites. Bloggers use blogs to chronicle their daily lives and express their opinions on a wide range of issues. Many studies using sentiment analysis use blogs as a source of opinion. There are a lot of product reviews, subjects, and more on these sites. Blogs, as defined by D. E. O'Leary (2011), are a particular type of website in which the authors provide both information and their own personal ideas.Tests on sentiment analysis in web-based texts, such as forums, blogs, and reviews, were undertaken by Boiy and Moens using machine learning (2009). Handwritten annotations show if a language or statement is positive, negative, or neutral with respect to a given entity.. This survey is interested in finding out how people feel about the products they've used in the past. Additional categorization models are learned and tested as part of a cascaded pipeline process. Input texts are noisy; emotions are ascribed to a specific entity; and the training set is restricted. Researchers Ye et al. compared Nave Bayes, SVM, and character-based N-gram models for seven important tourism destinations in the United States and Europe (2009). A number of empirical investigations have shown that SVM and N-gram approaches outperformed the Nave Bayes strategy, and all three algorithms achieved at least 80% accuracy when training datasets contained numerous words. reviews.

2.2 "Node of Attraction" (NoA) is a term coined by Mustafa Hajeer et al. (2012) to describe a node in a network community that is the most active. If a post or communication attracted additional nodes and eventually formed a cluster, this NoA would be regarded as the originator. When analyzing an OSN dataset, genetic algorithms (GA) are employed to locate clusters of network communities. One or more thematic notions may be used to generate clusters based on a wide range of conversation subjects within our OSN (e.g., comments, emails, chat phrases, and so on).

2.3 2.3

2.4 For the sake of representing the syntactic connections between opinion targets and the sentences that reflect such opinions, Asad and colleagues developed a suffix tree data format in 2012. To begin with our SRT settings and observed language attributes are used in the initial labeling of samplers. You can also try different combinations of the POS, role, and word characteristics to see what works best. An increase in accuracy can be achieved by combining more features.

2.5 **Newsgroupsdata**

A wide range of classification approaches have been investigated in the text categorization literature. Naive Bayes classifiers and decision trees, as well as decision trees and neural networks and learning algorithms are examples of neural networks (Yan-Shi Dong et al., 2004). Svetlana Kiritchenko et al. (2001) devised a method for making machine learning more approachable. A lack of labeled data and the expensive expense of tagging material that does exist are the key difficulties in text categorization.

Text categorization research has looked at a range of strategies. Naive Bayes classifiers; decision trees; and neural networks and learning algorithms are all instances of naive Bayes classifiers (H. Schutze et al; 1995). (Yan-Shi Dong et al., 2004). Svetlana Kiritchenko et al. (2001) devised a method for making machine learning more approachable. It's difficult to classify text because of the scarcity of data that has been labeled and the high expense of labeling the data that has been tagged.While Suresh Kumar and colleagues (2015) tried SVM classifier on unlabeled data initially, they subsequently tested Naive Bays classifier. As a result, SVM outperformed Naive Bayes in this study. The results of the experiments also demonstrated that the effectiveness of co-training is dependent on the type of learning employed.

(N. Priyadharshini et al., 2013) state that they employed a method previously used to segment images and identify document parts as either text or images. Smearing rules are employed to divide the document image into blocks, and features are extracted from each block. For this genetic programming classifier, the Discipulus tool was used.

Since natural languages are so intricate and document feature spaces are so vast, this classification problem has proven to be incredibly difficult to solve. The suggested work by Saad M. Darwish et al. (2015) provides classifiers based on multi-tree representation of documents that can classify documents into more than two categories at once (multi-class classification) using genetic programming and multi-objective techniques. This combination has the ability to reduce errors due to the fact that each class is represented as a different target. For the most part, text categorization research has focused on binary problems, in which a document's content can be used to classify it as either relevant or irrelevant to a specific topic. Text data sources like Internet news, e-mail, and digital libraries may be challenging to categorize due to the wide range of topics they cover.

The most common approach is to segregate binary classification problems for each class in a multi-class text categorization system. It is necessary to apply all binary classifiers and integrate their predictions into a single decision when classifying a new document. The themes are ranked at the conclusion of the list. Observations from the World Wide Web

In comparison to other social media networks, such as Facebook and LinkedIn, Twitter is less formal and has a more varied lexicon. With the help of emoticons and other symbols, people can express themselves on a wide range of topics and in numerous ways (Agarwal et al. 2011). Emotional annotations were included in the Twitter corpus prepared by Pak and Paroubek (2010). Carvalho et al. employed genetic algorithms to find subsets of a set of paradigm words in order to improve classification accuracy (Carvalho et al., 2014). Xia et al. (2011) used an ensemble framework with two feature sets and three base classifiers to develop the ensemble framework for sentiment classification in their study. For SA ensemble learning, the present state of the art mostly includes of conventional approaches like Majority Voting, Bagging and Boosting (see below) (Wang et al., 2014). The Bayesian Model Averaging ensemble approach suggested by Fersini et al. outperformed both standard classification methods and ensemble methods in terms of performance (Fersini et al., 2014).

2.6 **BaggingClassifier**

The ensemble methodology, which integrates the findings of multiple fundamental classification models into a single output, is a widely used classification method (T. Ho, 1994; J. Kittler,, 1998). The ensemble technique has been employed by a number of academics to improve the accuracy of topical text classification. K-NN, Relevance feedback, and Bayesian classifiers all perform better than a single classifier in early studies (L. Larkey et al, 1996).This bagging technique works well with "unstable" learning algorithms, such as those in which even minor adjustments to the training data cause significant shifts in predictions. Neural networks and decision trees are examples of unstable learning algorithms, according to Breiman (1996a).

Some data sets using decision trees show that even after ten members have been added to an ensemble (Schapire, Freund, Bartlett and Lee, 1997), the test-set error can be further reduced (and they note that this result also applies to bagging).

Classification accuracy can be improved by mixing different types of machine learning algorithms. In the early days of ensemble machine learning, one of the first methods used was Bootstrap aggregation, also known as Bagging (N. Anitha et al, 2013). Inverse Document Frequency was proposed by Saraswathi et al. (2012), and bagging techniques were used to classify the viewpoints.

2.7 **ArcingClassifier**

It is hoped that the ensemble framework developed by Xia et al. (2011) will improve sentiment classification accuracy by taking advantage of many feature sets and classification methods.

According to Freund and Schapire (1995,1996), an algorithm was suggested that uses adaptive resampling and combination (thus the name "arcing") to boost weights for cases that are more frequently misclassified

during resampling and weighted voting to combine them. Basic classifiers benefit from a hybrid model's ability to outperform them (Tsai 2009).

In this study, a new hybrid approach to the sentiment mining problem is proposed. As a way to improve performance, we've developed a new architecture using arcing classifiers and methods like neural networks, support vector machines (SVMs), and generalized additive models (GAs) in conjunction with the classification methods neural network (NB), support vector machine (SVM), and generalized additive model (GA). In the context of sentiment mining, the proposed bagged (NB, SVM, GA) and hybrid classifiers outperform standalone classifiers while heterogeneous models outperform homogeneous models.

PROPOSEDMETHODOLOGY

Several academics have looked into the use of an ensemble classifier, which combines the strengths of multiple classifiers (D. Tax et al, 2000). It is vital to combine redundant and complementary classifiers to improve robustness, accuracy and overall generalization. Studying ensemble approaches for sentiment classification tasks is the goal of this research project. In order to forecast classification scores, this study first constructs base classifiers such as Naive Bayes (NB), Support Vector Machine (SVM), and Genetic Algorithm (GA). The accuracy of all categorization experiments was evaluated using a 10 10 cross-validation. To improve generalization, well-known homogeneous and heterogeneous ensemble approaches are used with base classifiers. Blogger, newsgroups, and twitter datasets, which are commonly utilized in sentiment categorization, are used to demonstrate the feasibility and benefits of the proposed methodologies. Finally, some in-depth discussion and conclusions are formed about the effectiveness of the ensemble technique for sentiment classification after conducting a wide range of comparative trials.

New hybrid methods for sentiment mining are proposed in this study. In order to generate better results, a new architecture based on the coupling of classification algorithms employing bagging and arcing classifiers is defined. In order to achieve the best classification results, the suggested approach combines five distinct phases: preprocessing, document indices, feature reduction, classification, and combining.

A. Pre-processing of the data

The noise in our data collection was reduced using a variety of pre-processing approaches. We were able to develop a more accurate classifier in less time by reducing the size of our data collection. The following are the primary actions to be taken:

Pre-processing, feature extraction / selection, model selection, and training and testing the classifier are all part of the process.

Reduces the size of the supplied text documents greatly through data pre-processing. Sentence boundary identification, stop-word deletion, and stemming are all part of this process. For example, "a," "the," "an," and "of" are examples of stop-words, which are functional terms that appear frequently in the text but are not useful for categorization. It is the act of reducing words to their root or basic form, which is called stemming. Algorithms like Porter's Stemmer are commonly used in the English language to remove the suffixes from an English word and reduce the vocabulary of a training set.

approximately one-third smaller than before. This would result in "generalizations generalization generalize general" being the stemmed version of the English word "generalizations."

In other words, it's "gen". Additional pre-processing is required in circumstances when the source documents are web pages.

HTML and script tags can be altered and modified.

It is possible to determine the most important words in a document by extracting/selecting features. A variety of techniques are employed to achieve this goal, including TF-IDF, LSI (latent semantic indexing), multi-word, and others. When we talk about the "features or properties" of a text, we're

referring to things like key words and phrase patterns that are common to that text.
In order to train a classifier for text, a suitable machine learning technique is used to transform the document into a document vector. A test set of text documents is used to evaluate the trained classifier. Text documents are classified using this model if the trained classifier's classification accuracy is deemed adequate for the test set.

A. **DocumentIndexing**

B. It is possible to determine the most important words in a document by extracting/selecting features. A variety of techniques are employed to achieve this goal, including TF-IDF, LSI (latent semantic indexing), multi-word, and others. When we talk about the "features or properties" of a text, we're referring to things like key words and phrase patterns that are common to that text.

C. In order to train a classifier for text, a suitable machine learning technique is used to transform the document into a document vector. A test set of text documents is used to evaluate the trained classifier. Text documents are classified using this model if the trained classifier's classification accuracy is deemed adequate for the test set.The "bag of words" strategy is commonly referred to as this. When the word is present, it's vital to consider what values to utilize. Weighting each present word based on its frequency in the document and/or in the training corpus as a whole is perhaps the most prevalent strategy. The tfidf (term frequency-inverse document frequency) weighting function is the most frequent, but there are others. A binary weighting function is employed in most sentiment categorization studies. It has been found that assigning a value of 1 if the word appears, and 0 otherwise, is the most successful method.DimensionalityReduction
Techniques for reducing the number of dimensions in a dataset have been proposed. Using this method, the original data is transformed into a low-dimensional form. Data analysis becomes more efficient and accurate as the number of dimensions is reduced.
Steps: Pick a dataset to work with.

Pre-process the data by discretizing it.

Using the Best First Search algorithm, you can exclude attributes that are duplicated or overly frequent.

Apply a classification method to the duplicated attributes and compare their results.

Pick the One That's Right for You.
1)BestfirstSearch
Classifier assessment models are used by Best First Search (BFS) to estimate the relative merits of characteristics. For classification purposes, attributes with a high merit value are termed prospective attributes. A backtracking facility is used to search the subsets of attributes. The best way to begin is to begin with a blank set of attributes and search ahead, or to begin with a full set of attributes and search backward.

D. ExistingClassification Methods

1) NaiveBayes(NB)

2) Word feature text categorization is well-suited to Nave Bayes assumption of attribute independence. It is possible to learn each attribute's parameters separately when the number of characteristics is huge, which substantially simplifies the learning process.

3) Depending on the type of event, there are two models. The binary occurrence of words is a property of the event in the multi-variate model, which makes use of a document event model. Rather of taking into consideration the fact that words can appear several times in a document, this model simply ignores this fact. Multinomial models should be utilized instead, where a multinomial distribution is employed to account for many instances of a word. In this case, the words take on the role of the actions.

4) SupportVectorMachine(SVM)
Multi-dimensional function approximation can be achieved using the support vector machine (SVM) technique, which was recently created. Regression functions and classifiers are the primary goals of support vector machines, which aim to find a classifier or regression

function that minimizes empirical risk (that is, training set error) (which corresponds to the generalization or test set error).
Given a collection of N linearly separable training examples S x,
Each example is a member of the Rn I I = 1,2,..., N set.
The SVM learning algorithm seeks the best hyperplane for one of the two classes represented by yi 1,1.Decidedly positive and negative examples are separated by a margin of w.x + b = 0. Linearly separable data classification is based on the following decision function:
f(X) sign W.X b

(1)Where
wandbarefoundfromthetrainingsetbysolving aconstrainedquadraticoptimizationproblem.T hefinaldecisionfunctionis
f(x) sign aiyi(xi..x) b
5) The function is based on the examples used during training for which
6) Non-zero s is the case. These instances are called support
7) vectors. Support vectors are often only a small fraction of the entire quantity of data collected in a study. By utilizing a nonlinear kernel in a non-linear setting, the SVM concept can be extended to a high-dimensional feature space. Linear classification can be performed in this large feature space. The SVM classifier is widely used in practical applications such as text classification and pattern recognition.In contrast to SVM, which is utilized in classification issues, support vector regression provides an alternative loss function that incorporates distance metrics. There are several variables that can affect how well a regression model performs: C, tube width, and mapping function. There will be polynomial degrees that range from 0 to 5 in this study. The best polynomial kernel for this study has epsilon=1.0E-12 parameters and one constant parameter. Genetic Programming (GA)
The genetic algorithm is built on metaphors for some of nature's evolutionary mechanisms as a machine learning paradigm. A population of chromosomes, a set of character strings, is created in a machine to accomplish this goal.
Individuals represent possible solutions to the optimization problem. Individuals in genetic algorithms are represented by binary vectors. This search yielded an n-dimensional boolean space. The quality of each possible solution is considered to be measurable using a fitness function.
Genetic algorithms use fitness-dependent probabilistic selection of people from the present population to create the next generation of individuals. Those chosen for the following generation are subjected to genetic operators' acts. Mutations and crossings are two of the most commonly used operators in genetic algorithms that represent individuals as binary strings. When two strings are combined, two new strings are created, however when a single string is changed, only one string is affected. Other genetic representations necessitate the use of the correct genetic operators.
The process of selecting and applying genetic operators to produce successive generations is repeatedly repeated until an acceptable result is attained. An evolutionary algorithm's performance is affected by many factors, including the choice of genetic representation and operators, the fitness function used in selection and various user-specified parameters, such as the size of the population and how likely it is that different genetic operators will be applied.. It is in this section that we'll go over how the genetic algorithm in general works terms:
Procedure:
t - 0
Initiate P. (t)
as long as (this is not a case of the program terminating)

cross-select P(t) from P(t - 1) (t)
P(t) must be mutated in order to evaluate P. (t)
Hence, the end.
In order to create a meaningful impact, we must utilize all of the tools at our disposal. Bigrams are used first, then grammatical categories are tightly defined, and finally the voting mechanism is used to improve the

accuracy of each classifier. In the end, the competition is fierce.

e. Ensemble Classifiers using Bagged Ensembles

1.The following is the procedure for bagging a given collection of data (Breiman, L. 1996a). I = 1, 2,...K iterations of the training set Di are sampled and replaced with d tuples from the original training set D. The bootstrap sample Di delivers the same results when it is repeatedly sampled with a replacement from the specified training data set D. An example from the specified training set D may appear more than once in any replication training data set Di, or it may not appear at all. Mi is a classifier model that is generated for each training set (Di). One vote is cast for each of the classifiers, each of which provides an estimate of the unknown tuple, X. X is assigned to the class with the most votes as a consequence of Mvote *'s counting.

1. Bagged ensemble classifiers are an algorithm that uses bagging as input.
2. A collection of d tuples, D
3. Three models make up the ensemble.
4. Base Classifier Output: • A Bagged (NB, SVM, GA) and M* Method: • Base Classifierfori=1tokdo//createkmodels
5. Create a bootstrap sample, Di, by sampling D with replacement, from the given training data set Drepeatedly. Each example in the given training set D may appear repeated times or not at all in any particularreplicate trainingdata setDi
6. UseDitoderiveamodel,Mi;
7. Classify each example d in training data Di and initialized the weight, Wi for the model, Mi, based on theaccuraciesofpercentageofcorrectlyclassifiedexampleintrainingdataDi.
8. endfor

Tousethebaggedensemblemodelsonatuple,X:

1. ifclassificationthen
2. leteachofthekmodelsclassifyXandreturnthemajorityvote;
3. ifpredictionthen
4. let eachofthek models predictavalue for Xandreturntheaveragepredicted value;

- Arcing-based Heterogeneous Ensemble Classifiers.
- If you have two training sets of tuples (D and Di), arcing (Breiman. L, 1996) can be used to sample from one of them and replace the training set's tuples with those from the original (D) set. Some samples from the D dataset will occur many times in the Di training dataset. Thus, the test dataset contains just those samples that did not make the cut for training. A classifier model (Mi) is then built from the training examples (d) in the training dataset Di. Mi is a classifier model that is generated for each training set (Di). One vote is cast for each of the classifiers, each of which provides an estimate of the unknown tuple, X. M* makes use of a hybrid classifier to figure out which category is the most popular (NB-SVM-GA).
-
- Hybrid NB-SVM-GA with Arcing Classifier as input
- D,asetofdtuples.
- k=3,thenumberofmodelsintheensemble.
- Base Classifiers (NB, SVM, GA)Output:HybridNB-SVM-GAmodel,M*.Procedure:

1. / Create k models for I = 1 to the power of k
2.Create a new training dataset, Di, by sampling D and replacing it with a new dataset. There may be multiple instances of the same example from the given dataset D in the training dataset Di.
3.Make a model out of Di, Mi.
4.Use the percentage of correctly classified examples in the training data Di to set the weight, Wi, of the model for the model, Mi.5.endfor

Hybrid models can be used to any type of data set.

it follows that if classification is the case, thenvote for the most popular classification of X in each of the k models

Assuming the prediction is correct,Let X be predicted by each of the k models, and then calculate the average of those predictions.As with bagging, some of the original D tuples may not be included in Di, while others may appear more than once in Di.

II.

PERFORMANCEEVALUATIONMEASURES

A. CrossValidationTechnique

B. Cross-validation, also known as rotation estimation, is a statistical test that determines whether or not the findings of one set of data may be applied to another. A predictive model's accuracy can be estimated using this method, which is most often employed in situations where the goal is to forecast the future. Cross-validation with a 10-fold dilution is common. It is important to pick the folds in stratified K-fold cross-validation so that the mean response value is about equal in all the folds.

C. **Criteriafor Evaluation**

In order to evaluate a classifier's performance, the percentage of test samples that are correctly classified is the most important parameter. To measure a classifier's accuracy, we look at how well it can predict the label of previously unknown or new data (i.e. tuples without class label information). In the same way, the accuracy of a predictor relates to how well a particular predictor can forecast the value of the predicted attribute for new or previously unseen data.

A. **BloggerDatasetDescription**

III. EXPERIMENTALRESULTS

There are 100 blogs included in this study. From the UCI Machine Learning Repository web page: https://archive.ics-uci-ml/datasets/BLOGGER

B.Description of the Newsgroups Dataset

A total of 20 newsgroups' worth of Usenet articles make up the data set. Those were retrieved from the KDD database at https://kdd.ics.uci

A description of the Twitter data set

For each issue, the data collection includes a total of 2000 tweets (1000 positive and 1000 negative tweets). There are tweets in both MSA and Jordanian dialects in this collection. Downloads were made possible using the UCI Machine Learning Repository's online repository.:http://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis#B. **ResultsandDiscussion**

Inthissection,newensembleclassificationmethodsareproposedforhomogeneousensembleclassifiersusingbagging and heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed intermsofaccuracy.

A)

HomogeneousEnsembleClassifiersusingBagging

The blogger, newsgroups, twitter datasets are taken to evaluate the proposed Bagged NB, SVM and GAclassifiers.

TABLEI.THEPERFORMANCEOFBASEANDPROPOSEDBAGGEDNBCLASSIFIERFORBLOGGERDATA

Dataset	Classifiers	Accuracy
BloggerData	ExistingNBClassifier	71.00%
	ProposedBaggedNBClassifier	76.00%
	ExistingSVMClassifier	73.00%
	ProposedBaggedSVMClassifier	77.00%
	ExistingGAClassifier	77.00%
	ProposedBaggedGAClassifier	81.00%

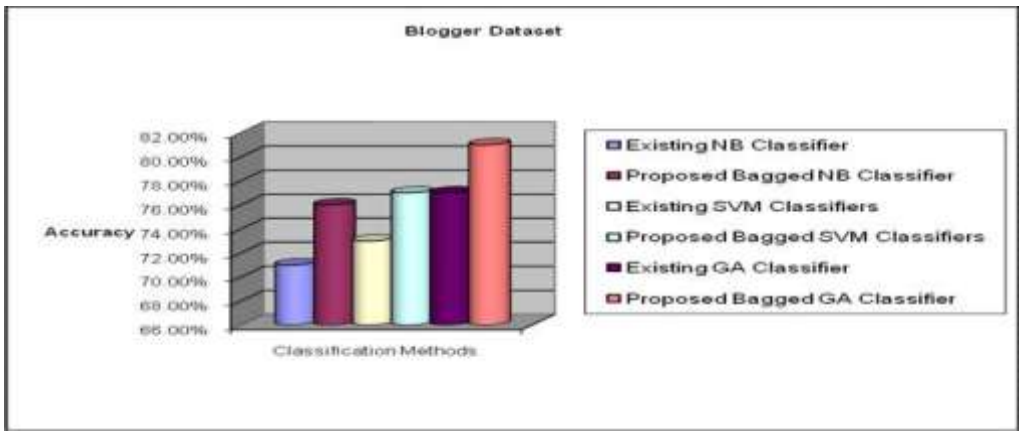

Fig.1. Classification Accuracy ofBase andProposedBagged EnsembleClassifier usingBlogger Data

TABLEII.THEPERFORMANCEOFBASEANDPROPOSEDBAGGEDNBCLASSIFIERFORNEWSGROUPSDATA

Newsgroup Dataset	**Classifiers**	**Accuracy**
misc.forsale	ExistingNBClassifier	97.50%
	ProposedBaggedNBClassifier	98.50%
	ExistingSVMClassifier	97.90%
	ProposedBaggedSVMClassifier	98.20%
	ExistingGAClassifier	97.80%
	ProposedBaggedGAClassifier	98.70%

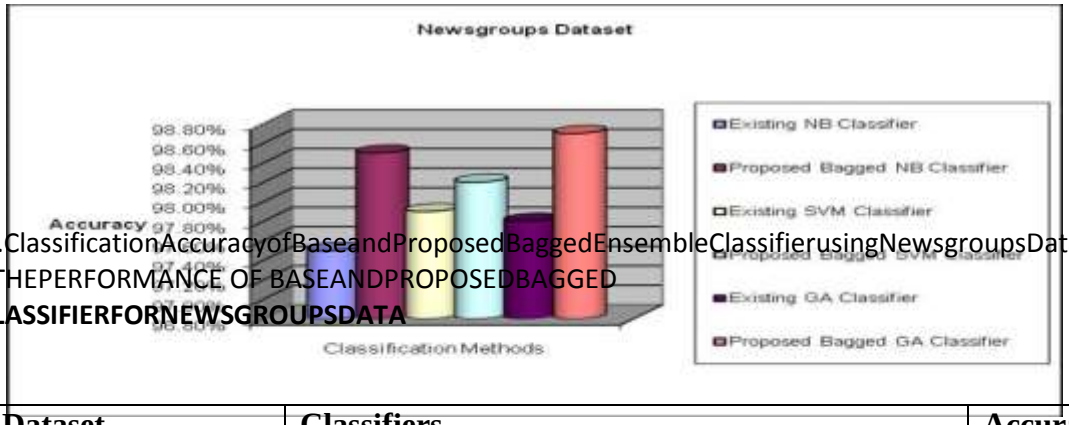

Fig.2.ClassificationAccuracyofBaseandProposedBaggedEnsembleClassifierusingNewsgroupsDataTABL

EIII. THEPERFORMANCE OF BASEANDPROPOSEDBAGGED

GACLASSIFIERFORNEWSGROUPSDATA

Dataset	**Classifiers**	**Accuracy**
TwitterData	ExistingNBClassifier	97.81%
	ProposedBaggedNBClassifier	98.36%
	ExistingSVMClassifier	97.26%
	ProposedBaggedSVMClassifier	98.90%
	ExistingGAClassifier	96.72%
	ProposedBaggedGAClassifier	97.81%

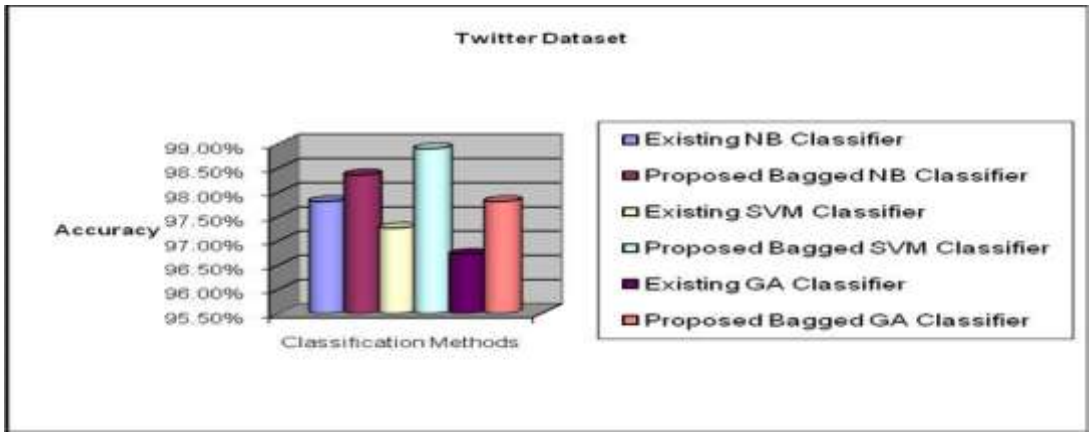

Fig.3.ClassificationAccuracyofBaseand ProposedBaggedEnsembleClassifierusingTwitterData

B) Using a bagging classifier in conjunction with NB, SVM, and GA as base learners, a new ensemble classification approach is developed and the performance is evaluated in terms of accuracy. The base classifiers are built using NB, SVM, and GA, and classification accuracy is tested using the 10-fold cross validation (Kohavi, R, 1995) technique. To achieve excellent classification performance, bagging is carried out using NB, SVM, and GA. Analysis of results shows that proposed bagged NB, SVM, and GA are superior to individual approaches in terms of classification accuracy for the datasets of bloggers, newsgroups, and Twitter. Following Fig. Improved classification accuracy can be seen in one to three proposed combined models compared to the base classifiers. Thus, the combined methods outperform the individual methods for the blogger, newsgroups, and twitter datasets, proving their superiority.

C) Theblogger,newsgroups,twitterdatasetsaretakentoevaluatetheproposedhybridNB-SVM-GAclassifier.TABLEIV.THEPERFORMANCEOFBASEANDPROPOSEDHYBRIDCLASSIFIERFOR BLOGGERDATA

Dataset	**Classifiers**	**Accuracy**
TwitterData	NaiveBayes	97.81%
	SupportVectorMachine	97.26%
	GeneticAlgorithm	96.72%
	ProposedHybridNB-SVM-GA	99.45%

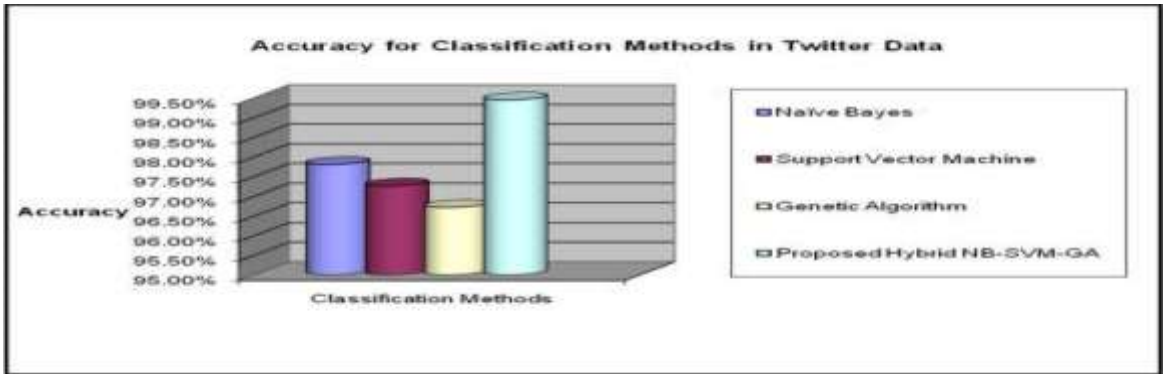

Fig.6.ClassificationAccuracyofBaseandProposedhybridNB-SVM-GAClassifierusingTwitterData

An investigation into new hybrid classification methods for heterogeneous ensemble classifiers employing arcing classifiers has been conducted. Section V details the results of a comparison of the performance of base and hybrid classifiers on this dataset. Cross-validation was utilized to evaluate the accuracy of classification. Classifiers such as NB, SVM, and GA are built one at a time in the suggested method in order to maximize generalization performance. The second step is to generate the NB, SVM, and GA ensemble. Using the ensemble approach, it is determined on a weight (0–1 scale) based on generalization performance, as illustrated in Tables IV to VI. Figures 4 to 6 indicate that the proposed hybrid models outperform base classifiers statistically in terms of classification accuracy.

Hybrid SVM-GA is more accurate in classifying blog, newsgroup, and twitter datasets than any of the individual techniques...

IV. CONCLUSION

Classification of homogeneous ensemble classifiers using new bagging-based methods has been evaluated on a variety of datasets, including blogs and newsgroups. The advantages of their underlying components are integrated into the suggested bagged NB, SVM, and GA classifiers. For heterogeneous ensemble classification, a new hybrid SVM-GA model is constructed and assessed in terms of accuracy using the NB, SVM, and GA models as the base classifiers.

The experiment's results are outlined in this section.GA outperforms SVM and NB in main accuracy measures when it comes to blogging datasets.

SVM outperforms GA and NB in terms of overall accuracy for newsgroup datasets.

Both SVM and GA are outperformed by NB on the Twitter dataset.

Base classifiers are substantially less accurate than the proposed bagged techniques.

The hybrid NB-SVM-GA classifier is more accurate than the base classifiers.

There appears to be a lower critical value for the 2 statistic than previously thought.0.455. In this case, p = 0.50. This is less significant than the 0.05 or 5% significance level that is commonly accepted. As may be seen from a 2 significance table, this value has a degree of freedom of 1 and is therefore significant. Classifiers proposed in this study are statistically more significant at p 0.05 than existing classifiers.

Homogeneous and heterogeneous models are evaluated for the accuracy of base classifiers, and heterogeneous models outperform homogeneous models for blogger, newsgroups, and twitter datasets.

For both homogeneous and heterogeneous models, the blogger, newsgroups, and twitter datasets could be accurately identified.

The future research will be directed towards developing more accurate base classifiers particularly for theblogger,newsgroups,andtwitterdatasets.

ACKNOWLEDGMENT

Author thanks Annamalai University officials for giving the resources and encouragement needed to execute this project. The government of India's Department of Science and Technology, New Delhi, offers financing for this initiative under the Fast Track Scheme for Young Scientists. –

REFERENCES

The following papers were presented at the Workshop on Language in Social Media (LSM 2011), held in Portland, Oregon, by the Association for Computational Linguistics: Agarwal, A. Xie, B. Vovsha, I. Rambow, O. Passonneau, R. PA, 2011, pp. 30–38. Stroudsburg

IJIET, 3(1), 2013, pp. 22-31, "Sentiment Classification Approaches – A Review," N. Anitha, B. Anitha, S. Pradeepa, "Sentiment Classification Approaches – A Review,"

[1] “The classification of patterns using least squares twin support vector machines,” Expert Systems with Applications, 36(4), pp. 7535–7543, M. Arun Kumar and M. Gopal [1].

[2] (Amy Weinberg, Asad B. Sayeed, and Jordan Boyd-Graber), "Grammatical structures for word-level sentiment detection," Proceedings of the 2012 Conference on Human Language Technologies, Montreal,

Canada, June 3-8, 2012, pp. 667-676 of the North American Chapter of the Association for Computational Linguistics.

[3] This article, "A study of machine learning techniques for classifying text documents," was published by the Journal of Advances in Information Technology, 1(1), 2010.

[4] International Journal of Application or Innovation in Engineering & Management, Volume 2(3), Number 3, 2013, Pages 90-99, Bhumika, S. S. Sehra, and A. Nayyar, "A review article on algorithms used for text classification,"

[5] 5] Breiman L "Bias, Variance and Arcing Classifiers" Technical Report 460, University of California at Berkeley, Department of Statistics, 1996.

[6] Bagging predictors [6] Breiman, L., Machine Learning, 24(2), 1996a, p.123–140. [6]

[7] A machine learning approach to sentiment analysis in multilingual web writings, by E. Boiy and M. F. Moens, Information retrieval, 12(5), 2009, pp. 526-555.

[8] According to the International Journal on Natural Language Computing 2(3), 2013, pp. 39-48, "Opinion mining and analysis: a survey" by Buche, D. Chandak, and A. Zadgoonkar.

[9] A statistical and evolutionary approach to sentiment analysis is used in this method. [9] An International Joint Conference of the IEEE/WIC/ACM Computer Societies on Web Intelligence and Intelligent Agent Technologies (WI-IAT '14), pp. 110–117 (Carvalho, Prado and A Plastino). The IEEE Computer Society. According to an article published in the International Journal of Document Analysis and Recognition (IJDAR) in 2007, "A overview of document image classification: problem statement, classifier architecture and performance evaluation."

[10] [10]

[11] Bayesian ensemble learning in sentiment analysis: Fersini, Messina and Pozzi (2014) In Decision Support Systems, issue 68 (2014) pp 26–38.

[12] [11]

[13] An application of boosting to a decision-theoretic generalization of on-line learning is presented in the proceedings of the Second European Conference on Computational Learning Theory (pp.23-37).

[14]

[15] Proceedings of the Thirteenth International Conference on Machine Learning (ICML'96), Bari, Italy (pp.148-156). [12] Freund Y. and Schapire R.

[16]

[17] In IEEE Transactions on Pattern Analysis and Machine Intelligence, 16 (1994), pp. 66–75, T. Ho, J. Hull, & S. Srihari present "Decision combination in multiple classifier systems."

[18]

[19] "Combining classifiers: A theoretical framework," Pattern Analysis and Applications, vol. 1, no. 1, 1998, pp. 18–27, by J. Kittler.

[20]

[21] Proceedings of the International Joint Conference on Artificial Intelligence, Vol. 2, Montreal, Quebec, Canada, August 20-25, 1995, pp. 1137–1143 Kohavi R. "A study of cross-validation and bootstrap for accuracy estimation and model selection"

[22]

[23] Classification based on a genetic algorithm and programming: A survey, Journal of Theoretical and Applied Information Technology, 54(1), 2013, p. 48-58 (D. Kumar and S. Beniwal).

[24]

[25] Proceedings of the ACM SIGIR Conference, pp. 289–297, 1996, by L. Larkey and W. Croft, "Combining classifiers in text categorization."

[26]

[27] The third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, 1994, pp. 81–93, contains a comparison of two learning algorithms for text categorization.

[28]
[29] According to Liu, "Sentiment analysis and opinion mining" in Synthesis Lectures on Human Language Technologies, 5, pp.1-167 in 2012, "[19]
[30]
[31] [20] Mustafa H. Hajeer, Alka Singh, Dipankar Dasgupta, and Sugata Sanyal, "Clustering Online Social Network Communities using Genetic Algorithms," at WORLDCOMP'12: The 2012 International Conference on Security and Management (SAM'12), July 16-19, 2012, Las Vegas, USA.
[32]
[33] In their article, "Recent trends in text categorization approaches," Nidhi and V. Gupta report on the latest developments in the field.D. E. O'Leary, “Blog mining-review and extensions: From each according to his opinion”, Decision Support Systems,vol.51,2011,pp.821-830.
[34] Pak A. Paroubek, P., Twitter as a corpus for sentiment analysis and opinion mining, In Proceedings of the SeventhConference on Language Resources and Evaluation(LREC10). Valette, Malta. May 2010, European Language ResourcesAssociation.
[35] Poornima Singh, Gayatri S Pandi, “Opinion Mining Techniques for Social Network Analysis: A Survey”, InternationalJournalforScientificResearch&Development,2(12),2015,pp.350-354.
[36] N. Priyadharshini and V. MS, “Genetic programming for document segmentation and region classification usingdiscipulus,”InternationalJournalofAdvanced ResearchinArtificialIntelligence,2(2),2013,pp.15-22.
[37] Saad M. Darwish, Adel A. EL-Zoghabi, and Doaa B. Ebaid, "A Novel System for Document Classification UsingGeneticProgramming",JournalofAdvancesinInformationTechnology,6(4),2015,pp.194-200.
[38]
G.SaltonandM.McGill,“IntroductiontoModernInformationRetrieval”,McGraw-Hill,1983.
[39] Saraswathi.K, Tamilarasi.A,(2012),AModifiedMetaheuristicAlgorithmforOpinion mining, International Journal ofComputerApplications, 58(11), 2012,pp.43-47.
[40]
Schapire,R.,Freund,Y.,Bartlett,P.,andLee,W."Boostingthemargin: Anewexplanationfortheeffectivesofvotingmethods”,InproceedingsofthefourteenthInternationalConferenceonMachineLearning,1997,pp.322-330,Nashville,TN.
[41] H. Schutze,D. Hull,andJ. Pedersen,“AComparison of Classifiers andDocument Representations for the RoutingProblem”,InSIGIR’95,WashingtonD.C., 1995,pp.229–237.
[42]
F.Sebastiani,“Machinelearninginautomatedtextcategorization,”ACMComputingSurveys(CSUR),34(1),2002,pp.1-47.
[43]
SureshKumarandShivaniGoel,"EnhancingTextClassificationbyStochasticOptimizationmethodandSupportVectorMachine",InternationalJournalofComputerScienceand InformationTechnologies,6(4),2015,pp.3742-3745.
[44] SvetlanaKiritchenko,Stan Matwin,“EmailClassification withCo-training”,Proceedings ofthe2001 conferenceoftheCentreforAdvancedStudiesonCollaborative Research,2001,CASCON'01,pp.8.
[45]
Tsai,C.F.,Lu,Y.F,"CustomerChurnPredictionbyHybridNeuralNetwork",ExpertSystemswithApplication,39,2009,pp.12547-12553.
[46]
D.Tax,M.Breukelen,R.Duin,andJ.Kittler,“Combiningmultipleclassifiersbyaveragingorbymutliplying?”,PatternRecognition, Vol.33,2000,pp.1475-1485.
[47]
Tsai,C.F.,Lu,Y.F,CustomerChurnPredictionbyHybridNeuralNetwork,ExpertSystemswithApplication,39,2009,pp.12547-12553.
[48]
N.VasfiSisiandM.R.F.Derakhshi,“Textclassificationwithmachinelearningalgorithms,”JournalofBasicandAppliedScientificResearch,3(1),2013, pp.31-35.

[49] Verzani,J.GettingStartedwithRStudio.CA.O'ReillyMedia,Inc,2011.
[50] G.Wang,J.Sun,J.Ma,K.Xu,J.Gu,Sentimentclassification:thecontributionofensemblelearning,DecisionSupportSystems,57,2014,pp.77–93.
[51] R.Xia,C.Zong,andS.Li,“Ensembleoffeaturesetsandclassificationalgorithmsforsentimentclassification”,InformationSciences,181(6),2011, pp.1138–1152.
[52] Yan-ShiDong,Ke-SongHan,“AComparisonofSeveralEnsembleMethodsforTextCategorization”,Proceedingsofthe2004IEEEInternationalConferenceonServicesComputing,Shanghai,China,Sept.15-18,2004,pp.419-422.
[53] Ye,Q.,Zhang,Z.,&Law,R,“Sentimentclassificationofonlinereviewstotraveldestinationsbysupervisedmachinelearningapproaches”,ExpertSystemswithApplications,36(3),2009,pp.6527-6535.