Semi Supervised Multi Text Classifications for Telugu Documents

D Naga Sudha[1],Y Madhavee Latha[2]

[1]Research Scholar JNTUH College of Engineering, Hyderabad, Telangana, India
[2]Malla Reddy Engineering College for Women, Telangana, India

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: As the amount of information available on the internet grows at a rapid pace, text classification becomes critical. This data is in an unstructured state and will need to be digitized. Due to the digital nature of these documents, data must be organized by automatically assigning a collection of documents to predefined labels based on their content. To mitigate the growing impact of news text classification, keyword detection approaches based on mostly supervised classification methods have been proposed. However, in practice, the given data is largely unlabeled, necessitating the use of semi-supervised learning techniques instead. We examine the effectiveness of a semi-supervised method for Telugu news articles in this paper. It also addresses some of the most pressing issues in automated text classification, including dealing with unstructured text, handling large numbers of attributes using natural language processing techniques, and dealing with missing metadata due to Telugu's morphological complexity. After classification, semi-supervised clustering is used to classify patterns. Unlabeled texts are used to adapt the centroids, while unlabeled texts are used to capture text cluster silhouette coefficients. To that end, the aim of this study is to use semi-supervised learning methods to investigate the effect of n-gram feature selection on news article text classification. Statistical results classification rate, precision, recall and F-score for news articles are validated. The results show that the approaches significantly improve the performance.

Index Terms- Text classification, Logistic regression, Naive Bayes classifier, Support Vector machine, Shillloute Coefficient.

I.Introduction : As the amount of data grows at an exponential rate, it is important to analyses and classify it. As a result, the significance of text classification starts to emerge. Text classification is the process of assigning labels to documents based on their content and constructing a model using a training dataset. The classification of documents has certain flaws. It is primarily a big data problem, with high dimensionality (a large number of attributes) reducing the classifier's efficiency. Another critical aspect is feature selection, which is used to represent the document's features using a variety of methods, including binary representation and term frequency of occurrences. We carry out a comparison of classification algorithms and valuate variety of various feature sets with the goal of optimizing accuracy for the classification of news articles.

Text classification [1] is the method of automatically assigning one of the predefined labels to a paragraph or document. If any Di is a document in the entire set of documents D, and c1, c2, c3,..., cn is the set of all categories, the text classification assigns a category Cj to it. A corpus's articles can only belong to one category, and a text can only belong to one category.and A text can only be assigned to one of the categories. It's referred to as single-label, otherwise it's known as multi-label. A single-label text classification task is further divided into binary class and multi-class classification when a document is assigned to n mutually exclusive classes (Yaying Qiu and et al, 2010). Text classification will help us divide documents conceptually and has a wide range of applications.

As technology advances, the amount of text-based tools available grows. They can be Telugu-language online papers, blogs, or social media correspondence. For the corporation, organizations, and individuals,.a method for analyzing and classifying data was needed. Text classification can also be used to address the issue of categorization for Telugu language. There are several text-classification applications for foreign languages, but we attempted to create a new framework specifically for Telugu language.

News will be automatically classified into pre-defined categories on news websites, such as sport, business, education, research, and so on. The method can also be used to analyzes the mood of product reviews. For example, if the company posts a picture of a new product on Facebook, it will receive thousands of comments. Positive and negative feedback are classified by the programmers. In this way, it will assist Telugu businesses in improving customer service, identifying weaknesses, and developing implementation plans for correcting errors. Overall, our goal in creating Telugu text classification is to make it easier for news websites, organizations, and businesses to categories and classify their data.

The process of assigning classification to text documents is a key task in the text processing field. class labels to unseen documents using the model created during the training phase. 'Find class the process of assigning classification to text documents is a key task in the text processing field. labels' is a popular operation in many applications. A bank officer, for example, may want to examine loan data to determine which consumer loan applications are risky and which are not.

In machine learning systems, there are two general techniques: supervised learning and unsupervised learning. Traditional supervised learning algorithms need a sufficient amount of labelled data as a training set in order to construct a classification model that can predict the class memberships of unlabeled instances. Unsupervised learning, on the other hand, is focused entirely on unlabeled samples [6]. Massive amounts of accumulated data have been accumulated on the web, especially on blogs, forums, and social networks, and they continue to grow day by day without a doubt. Unfortunately, much of the data available lacks pre-assigned marks, limiting its use in a variety of realistic machine learning applications such as text classification, emotion recognition, and speech recognition. Furthermore, manually assigning labels to them can be time-consuming, boring, and costly. Learning a classifier with just a few labelled training data, for example, could not yield sufficient results. Many algorithms have proposed manipulating and using unlabeled data to help the learning process for better classification in circumstances where labelled data is insufficient.

II.RELATED WORK : On text classification and classifiers, the model is similar to Vandana Korde and C Namrata Mahender[2]. Text mining's main purpose is to allow users to extract information from textual tools. It involves operations such as extraction, classification (supervised, unsupervised, and semi-supervised), and summarization. The technologies of Natural Language Processing (NLP), Data Mining, and Machine Learning are all based on natural language processing (NLP). They evaluate the efficacy of various text classifiers. According to Y. H. LI and A. K. Jain [3,] the paper investigates four different methods for document classification: naive Bayes, nearest neighbour, decision trees, and a subspace form. These were tested on Yahoo news groups in seven categories (business, culture, health, foreign, politics, sports, and technology) and in combination. Automatic Text Classification is a semisupervised machine learning task that automatically assigns a given document to a collection of pre-defined categories, according to Mita K. Dalal and Mukesh A. Zaveri's research paper [4]. Mowafy M, Rezk A, and El-bakry HM[5] describe An Efficient Classification Model for Unstructured Text Documents using multinomial nave Bayes MNB with TF IDF and KNN for both news and non-news documents.

III. PROPOSED MODEL: In this paper is to categorize news into specific categories and evaluate the category predictor's performance. Data is obtained and preprocessed first, and then feature extraction algorithms such as unigrams, bigrams, and trigrams transform the content of a text document (Dj) into useful features (w1j... wkj). The extracted features are converted to numeric data, and the optimal number of clusters to use as machine learning inputs is determined. Finally, the MLmodels are trained using these transformed features, and the results are tested using the test dataset. Figure2 depicts the proposed process.

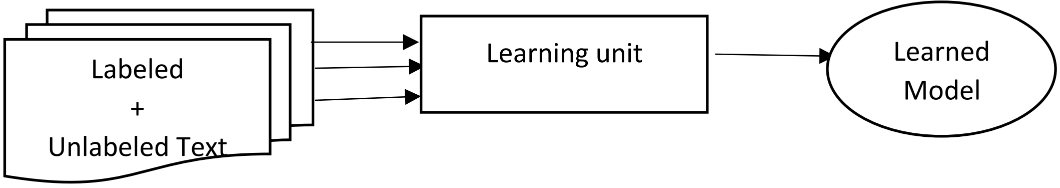

Fig.1The general framework of proposed semi-supervised method

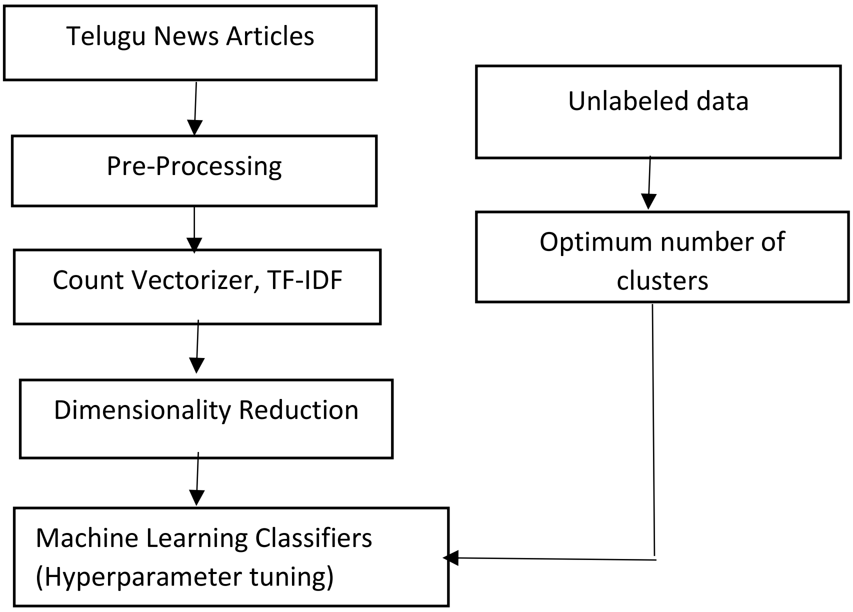

Figure 2 . Proposed method of Multi class text classification

A collection of news articles from daily haunt was used . It contains 3000 Telugu news documents divided into five categories: industry, sports, technology, politics, and films. 60% of the documents are used for preparation, 20%t for testing, and 20% are used for validation.

Multi-class classification is used to forecast news categories. The dataset used in this analysis, for example, is divided into five categories: Business, sports, technology, politics, and movies. Each class is labelled to aid comprehension and is often labelled in words. Label encoding is used to convert labels into numeric values in ML model.

The initial step is to Text preprocessing, also known as text cleaning, is a preliminary and critical phase in news classification that reduces the amount of space needed and improves classification efficiency [18]. The dataset is usually unstructured, with a mix of useful and useless information. Stop words, punctuation, special characters, meaningless sentences, quotes, and dates do not add any predictive power to the classifier. To reduce the distortions added to the model, a cleaning process should be performed before removing any element from the raw dataset.

There's a lot of data out there that hasn't been classified, and human labelling is a time-consuming and costly process. The advantages of semi-supervised methods are numerous. Preprocessing, semi-supervised clustering, classifiers and hyperparameter tuning are the four main modules of the proposed method. In semi-supervised clustering, both labelled and unlabeled data are used to cluster. The marked texts are used to form the silhouettes of text clusters, while the unlabeled texts are used to capture the centroids of the text components.

Preprocessing: In the preprocessing stage, the text document is represented in a word format. Tokenization and stop word elimination are two steps in the preprocessing process. Individual terms are tokenized from the document contents. Stop Word Removal: Stop words are common words such as posts, prepositions, and pronouns, and they are eliminated. Clustering is broken down into three steps: initialization, clustering, and output creation. We use both labelled and unlabeled text as input to the clustering process, which is semi-supervised. Each text that is associated with a label is given that label during the initialization step; if a text does not have a label, it is marked as unlabeled. The documents that aren't labelled are added to a list. Then we measure the distance between one unlabeled document in the list and all other unlabeled documents in the list, as well as the labelled cluster centroids. If an unlabeled document has the shortest distance to a labelled cluster, it is connected to that cluster and assigned that cluster's name. The cluster's centroid has moved up. It's also called "unlabeled." The procedure is repeated until all unlabeled records are labelled. Finally, if any cluster has less than three documents as members, it is marked as irrelevant and removed from the model.

It consists of two steps: feature extraction, which extracts the specific features/patterns, and feature representation, which numerically represents each feature. Text features are usually created using Count Vectorizer and N-grams [1, 22]. The simplest function extraction technique is Count Vectorizer. It essentially divides a document's words,all without regard for the order N-grams, on the other hand, are simply a text series of N tokens (words). It considers the order and relationship between words as well as the frequency of a word in the text. Depending on the number of tokens considered, the N-grams may be Unigram (N=1), Bigram (N=2), or Trigram (N=3). The values of the extracted features can be represented by Binary Representation (BR), Term Frequency (TF), Term Frequency Inverse Document Frequency (TF-IDF), and Normalized TF-IDF are some of the techniques used to display data [23, 24]. The Count Vectorizer and the TF-IDF feature representation technique are used in this article. The TF-IDF technique eliminates the most common words from the text and extracts only the most relevant function words [13].

IV. MACHINE LEARNING ALGORITHMS/CLASSIFIERS: A classifier is a machine learning model that assigns input data to one of many categories. The algorithms Naive Bayes, Logistic Regression, and SVM are used to train a model that can classify news articles into categories in this analysis.

Logistic Regression: On the basis of independent variables, discrete values (such as 0/1, yes/no, true/false) are calculated in logistic regression[9,15]. The probability of an event occurring is predicted using the logit equation, and the output values range from 0 to 1. The model is being tested and evaluated. This stage involves applying the model to the documents in the test set and evaluating their real class labels. In this step, the model is applied to the test set documents, and the real class labels are compared to the predicted labels. In this step, the document labels are only used for evaluation, while the class labels are used by the learning algorithm in the validation step..

Naive Bayes classifier: In the construction of classifier methods, Naive Bayes [6,10]is a technique for assigning labels to problem instances, where feature values are expressed by vectors and class labels are taken from a finite set. All of the algorithms considered in this step are based on a general theory.In this Nave Bayes classifier, the specific features are considered to be independent of other features[5].

Support Vector Machine: The SVM[12] is used as linear classifiers. It is based on statistical learning algorithms and attempts to find a hyperplane that divides the classes with the largest margins by mapping the documents into feature space. The SVM can be thought of as a perceptron extension. It maximizes the geometric margin while minimizing the classification error and ensures that the geometric margin is maximized. SVM's main goal is to find a hyper plane (linear or non-linear) that maximizes the margin.

V.Conclusions: For more accurate text classification, used TF-IDF N- gram model , and count vectorizer are used as for text Classification. The TF-IDF Vectorizer outperformed the count-based vectorizer, as anticipated, since it uses the TF-IDF Transformer to consider the value of a word in the text. Furthermore, selecting an appropriate classifier is just as crucial as gathering data. We switched our focus to Logistic Regression and Support Vector Machine after exploring the Naive Bayes method. The Neural Network performed worse than the SVM. The Artificial Neural Network is far more efficient than SVM, according to scholarly sources, but it cannot demonstrate its maximum strength for the text classification problem in our case. The proposed model compares various machine learning methods for categorising Telugu news stories, and it shows that support Vector Machine works well with a 93.64 % classification rate.

Documents	No. of Test documents	Classification Rate		F1-Score	
		Count	TF-IDF	Count	TF-IDF
Business	280	92.5	93.3	0.71	0.71
Technology	160	91.2	92.5	0.75	0.782
Education	80	92.9	92.2	0.70	0.77
Sports	40	92.6	92.6	0.694	0.681
Movies	40	90.6	91.5	0.681	0.71

Figure 3. Classification rate and F1-score of Telugu test documents

Documents	No. of. Test documents	Precision		Recall	
		Count	TF-IDF	Count	TF-IDF
Business	280	0.721	0.754	0.715	0.714
Technology	160	0.774	0.798	0.718	0.769
Education	80	0.650	0.756	0.790	0.804
Sports	40	0.680	0.718	0.710	0.726
Movies	40	0.650	0.714	0.713	0.705

Figure.4 . Precision and Recall of Telugu test documents

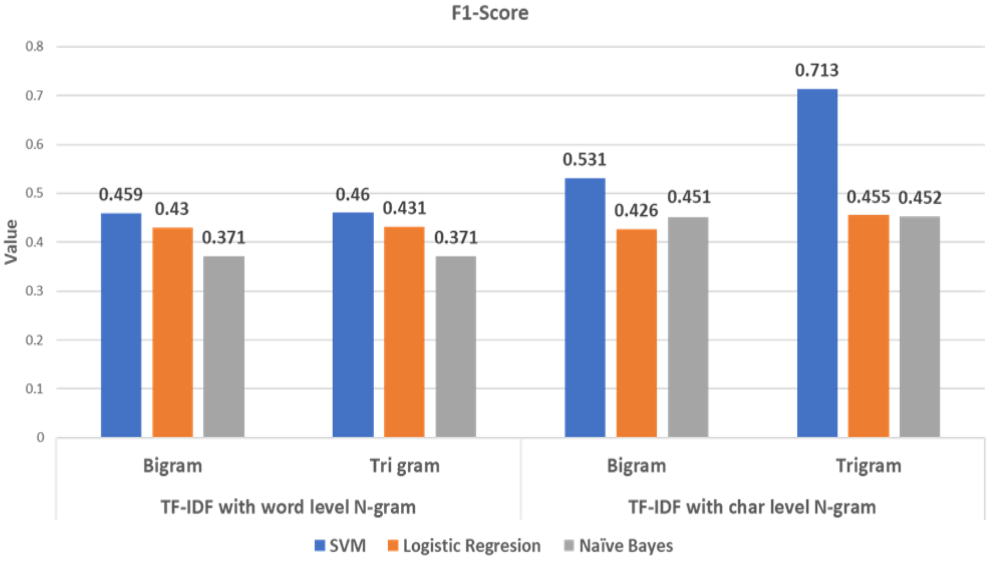

Figure 5. F1-score of Telugu News articles with TF-IDF with word, Char bigram and Trigram

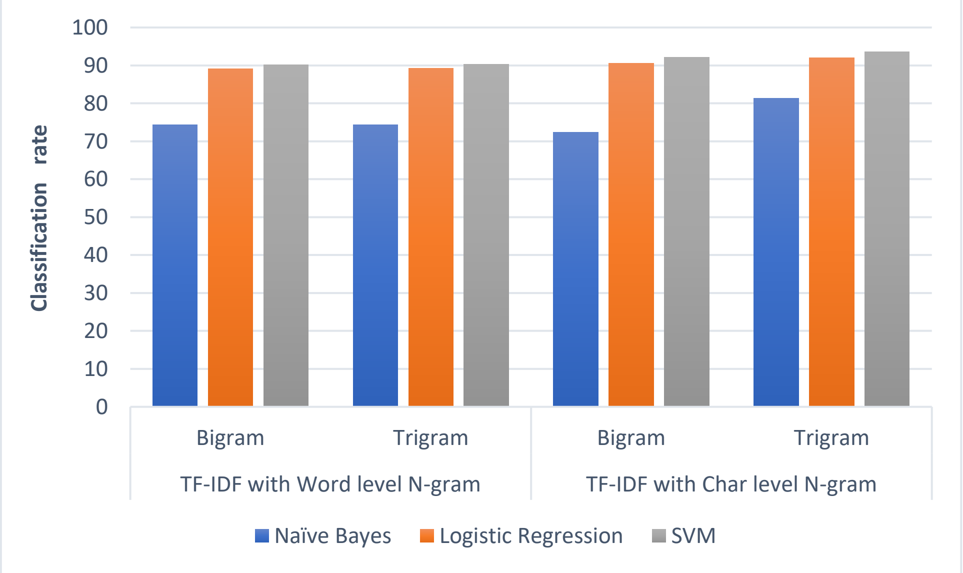

Figure 6. Classification Rate of Machine learning classifiers with TF-IDF using word, Char bigram and Trigram.

VII.REFERENCES

1. Mita K. Dalal, Mukesh A. Zaveri "Automatic Text Classification: A Technical Review", International Journal of Computer Applications (0975 – 8887),Volu me 28– No.2, August(2011).
2. K. Naleeni, Dr.L.Jaba Sheela,"Survey on Text Classification", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 6 July (2014). [3]Kratarth Goel, Raunaq Vohra, Ainesh Bakshi, "A Novel Feature Selection and Extraction Technique for Classification", IEEE International Conference on Systems, Man, and Cybernetics, October 5-8,(2016).
3. [4]Mowafy M*, Rezk A and El-bakry HM "An Efficient Classification Model for Unstructured Text Document by using multinomial naïve Bayes MNB with TF IDF and KNN and both for news articles"American Journal of Computer Science and Information Technology(2018).
4. [5]Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naïve Bayes Text Classification", IEEE transactions on Knowledge and Data Engineering, VOL. 18, NO. 11, November (2006).
5. Ioan Pop,"An Approach of the Naïve Bayes Classifier for the document classification", General Mathematics Vol. 14, No. 4 (2006).
6. George Tsatsaronis ,Vicky Panagiotopoulou, "A Generalized Vector Space Model for Text Retrieval based on Semantic Relatedness", Association for Computational Linguistics, Athens, Greece, 2 April (2009).
7. Y.H. Chen,Y.F. Zheng, J.F. Pan, N. Yang,"A hybrid text classification method based on K-congenernearest- neighbors and hypersphere support vector machine", International Conference on Information Technology and Applications, (2013).
8. Hosmer, D. and Stanley, L. (1989). Applied Logistic Regression, John Wiley and Sons, Inc
9. S.L.Ting,W.H.Ip, Albert.H.C.Tsang ,"Is Naive Bayes a Good Classifier for Classification?",International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, (2011)
10. Thorsten Joachims (2006).Text Categorizationwith Support Vector Machines: Learning with Many Relevant Features
11. A. K. Mourya, S. U. Ahsaan, and H. Kaur, "Performance and Evaluation of Different Kernels in Support Vector Machine for Text Mining," in Advances in Intelligent Computing and Communication, Singapore, 2020, pp. 264–271, doi: 10.1007/978-981-15-2774-6_33.
12. Liu X., Mou L., Cui H., Lu Z., Song S. Jumper: learning when to make classification decisions in reading IJCAI 2018, pp. 4237– 4243.
13. Fu Sun, Linyang Li, Xipeng Qiu, and Yang Liu. 2018. U-net:Machine reading comprehension with unanswerable questions.arXiv preprint arXiv:1810.06638.
14. N.A. Zaidi, G.I. Webb, M.J. Carman, F. Petitjean, J. Cerquides, ALRn: accelerated higher-order logistic regression, Mach. Learn.104 (2–3) 2016, 151–194