New Media and Mass Communication
ISSN 2224-3267 (Paper) ISSN 2224-3275 (Online)
Vol.93, 2020
www.iiste.org

Text Document Categorization using Enhanced Sentence Vector Space Model and Bi-Gram Text Representation Model Based on Novel Fusion Techniques

Abdisa Demissie Amensisa
School of Computing, Bahir Dar Institute of Technology, Ethiopia

Abstract
The text document classification tasks passes under the Automatic Classification (also known as pattern Recognition) problem in Machine Learning and Text Mining. It is necessary to classify large text documents into specific classes, to make clear and search simply. Classified data are easy for users to browse. The important issue in usual text document classification is representing the features for classification of an unknown document into predefined categories. The Combination of classifiers is fused together to increase the accuracy classification result in a single text document. This paper states a novel fusion approach to classify text documents by considering ES-VSM and Bigram representation models for text documents. ES-VSM: Enhanced Sentence –Vector Space Model is an advanced feature of the sentence based vector space model and extension to simple VSM will be considered for the constructive representation of text documents. The main objective of the study is to boost the accuracy of text classification by accounting for the features extracted from the text document. The proposed system concatenates two different representation models of the text documents for designing two different classifiers and feeds them as one input to the classifier. An enhanced S-VSM and interval-valued representation model are considered for the effective representation of text documents. A word level neural network Bigram representation of text documents is proposed for effective capturing of semantic information present in the text data. A Proposed approach improves the overall accuracy of text document classification to a significant extent.
Keywords: ES-VSM; Fusion, Text Document Classification, Neural Network, Text Representation, Machine learning.
DOI: 10.7176/NMMC/93-03
Publication date: September 30th 2020

1. Introduction

In the past ancient time, a document that supports online recording depends on the features that are demonstrated in technological science [18].Text Document classification, the undertaking of natural language text document, depending on their content of some previously assigned classification is a crucial part that plenty of data structured and management tasks can be done. Text document classification is classifying the items or texts found in the document need to be categorized to the specified class which is earlier known.

Text document classification is the field that appeared recently on the theme of text mining. Text document categorization is gorgeous by eliminating the non-electronic way of regulating documents based on their content and provides good accuracy [3, 16], to complete text document categorization and the prime activity present the text in documents. Bag of words is one method of creating text document vectors that seen and available commonly in a specific document [19]. However, it takes care of the similarities between the low-level performances. Because of ambiguity, ambiguity can cause a person's choice of motives as a necessity and minimize the accuracy of the procedure [2].The Classification system takes thousands of text files in single input and, as output; it keeps all the files in different folders on the basis of the contents in it and helps how to manage the files in the computer when having no time to read each file separately and organize the documents in different folders.

The great point of this article depicts for addressing text classification and overcoming the current problem which may be prevailed in the single features what is called unigram representation model by having consideration of an Enhanced Sentence Vector Space Model (ES-VSM) and Bi-gram model representing document types using fusion based approach. A Bigram is two consecutive elements from a texts or tokens. For users, Bi-grams are often easier to interpret than single words and it reduces the computational complexity. As the computational complexity is decreased, the easier to categorize the text documents. This implies complexity reduction leads to good and accurate classification. Before categorizations are processed, text presented in the document should be structured in the format that the machine can read. Assignment of text documents manifested in the form that the input methods are saving and classification algorithm can handle since documents are set of connotations which can be challenged to hold and try. That is the point why featureless or unstructured data should be converted to be coherent in a classification algorithm. The one that is examined to construct text in documents in form of sentence is Enhanced Sentence Vector Space Model, whereas for powerful seizing meaning of information in the documents neural network form representation for text documents is suggested in [1].The first step to taking the textual algorithm is to compile the training document. Such sets of documents are prepared for evaluation purposes. In

advance set of textbooks for text editing can be used includes Vehicle Wikipedia, Google Newsgroup, 20 Newsgroup -small, 20 Newsgroup-Large [1].

The remaining part of this article categorized as follows. Section 2 reviews related research work on text categorization. Section 3 presents document categorization process. Section 4 describes text document Classification using fusion technique. Section 5 presents experimental results and final section 6 shows the conclusion and explores future works.

2. Literature Survey on Text Document Classification

To avoid human intervention and finding the remedy for the problem of automatic document management, researchers in early 1990's started moving towards the design of machine learning algorithms [12].There has been the amount of effort done to promote text categorization by Bi-grams and enhanced Sentence Vector Space Model. The idea of incorporating various components is an attempt to increase the effectiveness of individual components. Few of relevant work is as follows:

Text Document classification uses the text document representation model like Enhanced Sentence Vector Space Model and n-gram model by using fusion based techniques, according to [1], this technique is based on score level of the classifiers.

In 2017, S.N.Bharath B., Ajit D [1] text document classification is proposed by examining two text document representation models like Unigram and Sentence Vector Space Model based on the scored level of fusion technique for the text document. The classifiers Interval valued with representation model of Enhanced Sentence Vector Model that examined the efficacy representing text document, whereas neural network classifiers of unigram word level of representing text documents are suggested to capture semantic information in the document effectively. It is based on the integration of textual documents with both the episode and the volumes and the neural network divisions, with the consistent 96.48% text document representation of Unigram and Space pitch model results.

In 2011, A. Zelaia., et al [3] presented that Multi-line / Multiple-Identification Documentation Errors have a number of distribution systems based on the K-NN algorithm. A new approach will be provided, based on the Bayesian Website voting, to implement several tag labels. For categorization method the low-cost vector structured documents acquired by singular value decomposition that can be used for training and testing documents, and set of KNN classifiers to imagine the category of test documents; each KNN classifier uses an extracted database subsampled from the typical training database. Results of micro averaged are better than macro, mean that micro average calculated by collecting decisions for all classes, whereas macro average estimated by averaging over the results of the various categories.

In 2013, Ankit Bhakkad., et al [4] presented modern and effective presentation for calculating all existing two consecutive word incidence used for EV-SM (Enhanced vector space model) as a basis to make the text in a form of vector. The intensified Vector Space Model is a supplement to normal vector space model that can store the place of tokens added to the frequency document. The result obtained in time complexity to find the frequency of all n-1 existing bigram will be O (n).

In 2010, William B., et al [5] says that N-gram frequency methods provide invaluable and effective methods for cracking down on documents. This is done using the samples of the selected categories, rather than looking for complex, high-level descriptions of natural language analysis or detailed descriptions. In essence, this approach demonstrates the "classification by example" technique. Gathering illustrative and construction statements can usually be handled automatically. Also, this system can withstand various OCR (encoder ID) problems, because it is not available in most of the N-graph grammar features and with any particular word.

In 2016, Aytug., et al [6] examined that Five-Stroke Keyword Removal Methods Predictability (Generate keyword-based keyword routine, Frequency keyword-based query-based number of variable-length statements, Using a keyword-based word presser based on static statistical information, Emphasizing keyword-based non-central wording and Text numeric method)Stage Tactical Algorithms and Scientific Text Document Identity (Category) Coordinated Methods. Multiplexing Methods The performances of the projects have a more understanding of basic instructional algorithms such as Naive Bayes algorithm 83. 49% and 91.49%are an Emergency Abduction Tactical.

In 2016, A. Seara Vieira. et al [7] proposed a new technique to reduce input components to improve the effectiveness of text fragments. The method uses a document store to divide the information into groups and sets up a textual-based layout based on the Text Hidden Markov model. The approach is ideal for large datasets. The result obtained from the experiment are very satisfactory compared to commonly used techniques like Info Gain and the statistical tests performed to demonstrate the suitability of the proposed technique for the pre-processing step in a text classification task.

In 2016 Selvi, S.Thamarai, et al. [8] proposed a new hybrid text categorization model that combines both Rocchio algorithm and Random Forest algorithm to perform Multi-tagging classification. This model overcomes the disadvantages of Rocchio and Random Forest algorithm. By combining two single label algorithm of text

classification, Resolution from more than one label of text classification model has been created. Rocchio algorithm, being a supervised curriculum algorithm, takes vectors as inputs and gives the relevant categories. Random Forest algorithm is an incremental learning algorithm that selects the appropriate categories from the selected categories. Experiments on 20 Newsgroup dataset and RCV1 dataset demonstrate the accuracy of Hybrid Text categorization model. The fact-finding results show that the provided model is somewhat accurate from existing algorithms.

In 2016, Heng Zhang.et al [9] presented a general structure for a short essay, with words and hidden headings summarized together. By referring to a large-scale external data collection named “corpus” which is a topic suitable for bits to identify subjects and then using the corpus to build a topic model with Latent Dirichlet Allocation. Word counts for words and short articles will be displayed as new words and integrated into the text to enhance content. It is possible to include briefings based on training and organization on both words and topics. The learning vectors by words and topics with an open short text categorization datasets primarily used to minimize the categorization error when compared to learning word vectors only. Test results have appreciated the effectiveness of the word/topic integration process.

In 2014, M B Revanasiddappa., et al [10] proposed a new union encodings to split text documents. Specifically, figurative placement methods; a symbolic collection selected without behavior, a symbolic set based on the selection symbolic cluster based on the feature using a discrete scale; and symbolic Character Set method are suggested. Writing techniques are very powerful to decrease data stops of vector writing vectors.Test results show that the symbolic set of control is to achieve better alignment with the traditional approaches based on the genre.

In 2016 V. V. Gulin., A. B. Frolov [11] proposed to describe that the text document classification problem taking into account their structural features is formulated as a machine learning problem using text document features characterizing relations on the set of lexemes. This is a distinction from the conventional model in which only unary relations are used. The planned machine's effectiveness is examined through computer tests on the class of the Reuters_21578 collection with eight known classifiers. The Planned modification is appropriate for use in the classification of text documents with simple classifiers like classifier of Naive Bayesian, support vector machines, logistic regression, and the classification solution tree (C4.5).

In 2012 J Yun., et al [15] proposed a two-step presentation model to represent text data. In two-stage presentation model, one word for word (information), another for accounting information (contact information) and these Standards are linked to the relationship between literature and concepts. The linkage framework used to select an appropriate document using the link structure between contextual texts depends on the context. Authors are done a multi-layer coverage framework to use the vocabulary and context data used by the two-level representation models. A multilayer classification framework contains three classifiers. Among these, two analysts appear to be at the core level and comparative level. These results of two classified segments will be consolidated and the final point will be provided for the third classifier. To show you the experiments that the author of data sets approved like 20Newsgroups, Reuters-21578 and Classic3 have shown that the proposed two-level representation model and multiple classification style result improves the text categorization compared with the existing single model (Term-based VSM, Term Semantic Kernel Model, Conceptual basis Vector Space Model, Concept-Semiconductor Corner Model and Word and Concept Vector Space Model) plus existing categorization methods.

3. Document Representation

In the Machine Learning approach, the classification algorithms are designed electronically having some knowledge of activities that classification can be settled in pre-standardized categorized training documents. Text document against the specified Word document (Word overlaps) of the dictionary has been moved by the opposite. Each word is seen at least once in or input document should match.

Text documents of natural heritage cannot be read and readable through standard algorithms. Classifiers of text document categorization and assimilate algorithms never process text documents directly into a text form. In the pre-running process, documents will be converted into better presentations. Normally, the documents are organized in characteristic vectors. One character is an internal structure that does not have an internal structure. A document is structured as a vector and a series of features and weights [19]. Regrettably, machines cannot clearly perceive words as human beings, but literary documents need proper representation. These written documents are usually required to replace a group of sets, which are suitable for partial algorithms [1].People read texts. The text of documents consists of written sentences and sentence consists of words. Human beings can perceive the linguistic structure easily with their meaning, while difficult to understand for machines that could not hopefully on natural language comprehensive yet. It is feasible to address this issue by teaching some languages to the machine and plenty of algorithms are suggested in the published paper. A stunning model of representation will be presented to represent textual information after it has been converted from unencrypted text data to a structured format. In text classification problems, the way documents are organized can influence the accuracy of

categorization to certain classes. The rapidity of collaborative and electronic data caching has been the key to addressing and organizing text literacy. Overall, text files have not been properly formatted for normal databases when they have unset text files. Text documents, a common set of collections, cannot be supported by cellular algorithms. The document depicts depletion in the number of degradation techniques that can be classified within the characteristics of feature extraction and feature selection.

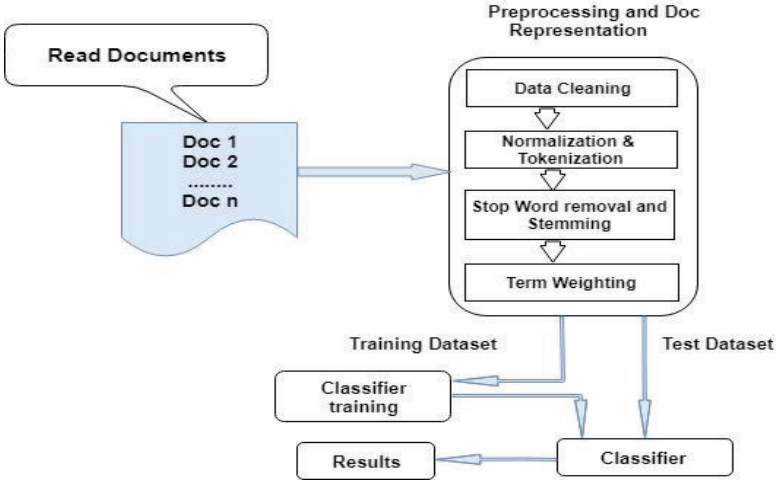

Figure 1. Document representation and classification process

3.1. Process for how to Extract Features

Pre-processing processes are used to investigate the boundaries of each language and to provide as many language dependents, tokens, words as possible, and to present written documents in clear word format. Dimension extraction is a pre-processing stage of the knowledge discovery. Basically, the process of writing predictive text sentences turns into a set of terms, and at the similar time enriches their meaning.

3.2. Token process

A document is considered a string, and then it is divided into the alternate listing list that will be input to the release process and is a process for clipping text to words, phrases, symbols or other useful items.

3.3. Document Tokenization Algorithm

Read document file
Read punctuations list
Read unnecessary characters list
For each token in file
If token ends with punctuation then
Remove punctuation from file
End if
If token is in characters list
Remove token from file
End if
End for

3.4. Stop word Elimination Process

The experienced words like "the", "a", "and" etc., so unpredictable terms have to be removed.

3.5. Algorithm for Eliminating Stop Word

Read document file
Read stop word list
For each token in file
If token is in stop word list then
Remove token from file
End if
If token is number then
Remove token from file
End if

End for

3.6. Stemming Word

Use an algorithm that can be used to shake rhythmic words into different syllables or grammatical forms. This is a process of transferring the tonic to their roots. To connect, to count on the computer, etc. and to sum up the whole basic word together.

3.7. Stemming Word Algorithm

Read document file
Read exception list
Read prefix list
Read suffix list
Assign the first 1, 2, 3…characters of the token to prefix
Assign the last 1, 2, 3…characters of the token suffix
For each token in file
If token is not in an exception list and prefix is in prefix list
Remove prefix from token
End if
If token is not in an exception list and suffix is in suffix list
Remove suffix from token
End if
End for

3.8. Process for Selecting Features

The step of clarifying words after feature extraction to build vector space is feature selection. This step reduces irrelevant word from documents to upgrade categorization efficacy and accuracy of a text classifier and reduces computational complexity. Selecting a character is a leading method to select the features that are most important from a data set, by removing anonymous and randomized attributes to improve the production of machine learning theme algorithms.

3.8.1. Text document representation using Bigram model

In an article [14] understands how to use scammers to distinguish between words where adults are exposed to an oral language. Compelling Practice Comparison of the Story Comparison of events in the story, the words surrounded by two words (Bigram) consist of repeatedly overlapping frequencies if this is the case with the temporal context. When the researchers come together in similarly large contexts, the unified words are united. These findings are particularly relevant because of the relative connectivity of Internet-based categories of information from different channels. Learners divide vocabulary words into the order and words into proper terms. It is then used to mention the exact wording of the term (Bigram), or the learner may use a submissive style design. The studies included the designs of the graphic in the artificial languages. If hypothesized, it is unlikely that it can expand its own vocabulary/divide/disaggregation education, due partly to the problematic identification of the appropriate system and to identify the widths and targets in the area.

4. Text Document Classification

Text assignment assignments are usually completed using single classifiers, which may result in unexpected results due to the disapproval of the selected clauses. But coming together of classifiers provide interesting result by assigning the document to the perfect class. To categorize the documents into certain predefined class the combined algorithm meet the better result. The methods for classification in this way can be two classifiers are designed with text document representation model of Enhanced Sentence Vector Space Model and Bigram model to fuse together the score level of classification algorithm result.

4.1. Text Document Categorization using Fusion Technique

Fusion refers to a purely unique approach to better comprehension and correct understanding of integrated compilers with different distribution divisions and related information with related datasets [13, 17].Indexing of reference classes was shown to reduce the amount of error in the collection's actions with the partition of a single fraction. Additionally, various techniques are used to make the final decision, making it harder for the individual to better understand the problem that is possible in each dataset [12].

5. Proposed Model

This paper proposes to address problem of classification of text documents using fusion based approach in a different environment by concatenating ES-VSM and Bi-gram representation model and feeds them as one input

to the classifier to address text document classification. Interval-valued classifier and Neural Network Classifier are the two-classifier fusion techniques designed based on representation models of the text documents. The score obtained from two classifiers develop the novel fusion based technique for text classification problem. Further score level fusion is employed for optimum decision by various combinations of classifiers for the classification of text documents. Score level fusion is robust enough to classify the documents.

Text representation model of ES-VSM and Bi-gram represents a suitable text document which will be input for the classification algorithms.

ES-VSM: Enhanced sentence vector space model would be presented to help in constructing a lower dimensional feature presentation of text documents for addressing that vector space model generates high dimensional feature matrix. It transforms the text document into the numerical vector by maintaining the sentence occurrence count in the document and stores the positions of tokens in addition to their frequency in document.

ES-VSM: Dj ={{ Y1,(x11,x12,x13……)},{Y2,(x21,x22,x23…)},……}

Whereas,

Dj is vector representation of document j,

Yi is the frequency of the ith term appears in document j.

(x1……xfi) represents the position at which ith term appears in document j.

The construction of ES-VSM is as follows.

Let kj, j = 1, 2, 3 . . . r be the different classes in the database which contains Dz, z= 1, 2, 3.. s documents in each class. Each document consists of tn, n = 1, 2, 3. . . n set of terms. All the terms tn from Dz documents are collected and a dictionary Dikj will be formed to represent the class kj and this process is applied for all the class kj, hence Dikj will be constructed for all the remaining r number of classes in the databases then ES-VSM will be constructed. Dictionary can be created and used where words from the inputted document are matched.The newly formed dictionary Dikj is subjected for dimensionality reduction.

ES-VSM will be constructed by calculating the probability between terms in dictionary Dikj and class kj. The calculation of the probability between terms in dictionary Dikj to the class cj is described as follows.

The numbers of classes in the databases contain a number of training documents and terms are extracted from each document by text pre-processing algorithm then the remaining terms are pooled for construction of a dictionary of the respective class. Let m be the number of terms in the dictionary. Each document can be given a term frequency vector representation of dimension m based on the frequency of occurrences of each of m terms in the class.

Naive Bayes theorem used to converts a matrix into a column vector of each dictionary and widely used for classification and clustering, but it is potential for general probabilistic modeling. To estimate the probability of a particular dictionary which is interpreted to a specific class say, Dikj, j = 1, 2 …j, we calculate the posterior probability of the text data interpreted as the specific dictionary is given by the formula.

$$Pr(Dikj|d) = \frac{P_r(Dik_j \mid t_1, t_2, \ldots t_n)}{n} \quad \text{--------- (1.1)}$$

Where t1, t2…tn is the term frequency vector representing d. The posterior probability of word ti from document d from the dictionary Dikj is given by

$$Pr(Dikj|Wl) = \frac{P_r(W_l|Dik_j)P_r(Dik_j)}{n} \quad \text{------ (1.2)}$$

$$Pr(Dikj) = \text{-----} \frac{\text{Number of terms in } Dik_j}{\text{Number of terms in training set}} \text{ - (1.3)}$$

Normalization is a process that converts a list of words to a more uniform sequence. To do normalization of the word tl, Pr (tl) is calculated by,

$$Pr(tl) = \frac{\Sigma \text{ Occurence of term } t_l \text{ in all dictionaries}}{\Sigma \text{ Occurences of all terms in all dictionaries}} \quad \text{------- (1.4)}$$

$$Pr(tl|Dikj) = - \frac{\text{Occurrences of terms } t_l \text{ in class } Dik_j}{\Sigma \text{ occurences of all terms in in class } Dik_j} \quad \text{-----(1.5)}$$

On the basis of the Bayes formula,prior probability Pr(Dikj) value is ,the likelihood pr(tl | Dikj) and pr(tl) which will be the evidence, along with the posterior probability for each term in the input document is d as pr(Dikj| tl),its posterior probability being annotated to the class Dikj ,pr(Dikj| d) can thus be measured using equation.

The posterior probabilities for all words tl, $1 < l < t$ present in document d with respect to the class Dikj, $1< j < z$ are calculated. From the obtained posterior probabilities of words in document d, the posterior probability of document d is calculated and being annotated to the class Dikj using Eq.1.Similarly, the posterior probabilities of the document d being annotated to all other classes Dikj, $1< j < z$ are calculated and presented like a vector which

provides the probability of the probability of the respective document belonging to each individual dictionary. These z values are used to approximate the document d in form of z level feature space and hence each dictionary will be provided with a z-dimensional belonging to each individual dictionary. These z values are used to approximate the document d in form of z level feature space and hence each dictionary will be provided with a z -dimensional vector representation with each dimensional value being crisp as shown below.

(Pr(Dik1|d),Pr(Dik2|d),…,Pr(Dikj|d) .... (1.6)

An interval value representation will be formed by considering minimum and maximum values of posterior probability of each dictionary Dikj. Based on the newly formed interval-valued data type a symbolic representative vector Symbj with j-dimension as given below.

Symbj=(I1,I2,I3…….Ij) …. (1.7)

Here Il is the interval which assimilates the lth feature value of all terms of the dictionary. Dikj is denoted as:-

Ijl=[f –jl , f +jl] ……(1.8)

Where f –jl =min {Pr (Dikj | ti) $\forall$ i =1…n} and f+jl=max {Pr (Dikj | ti) $\forall$ i =1…n} …… (1.9)

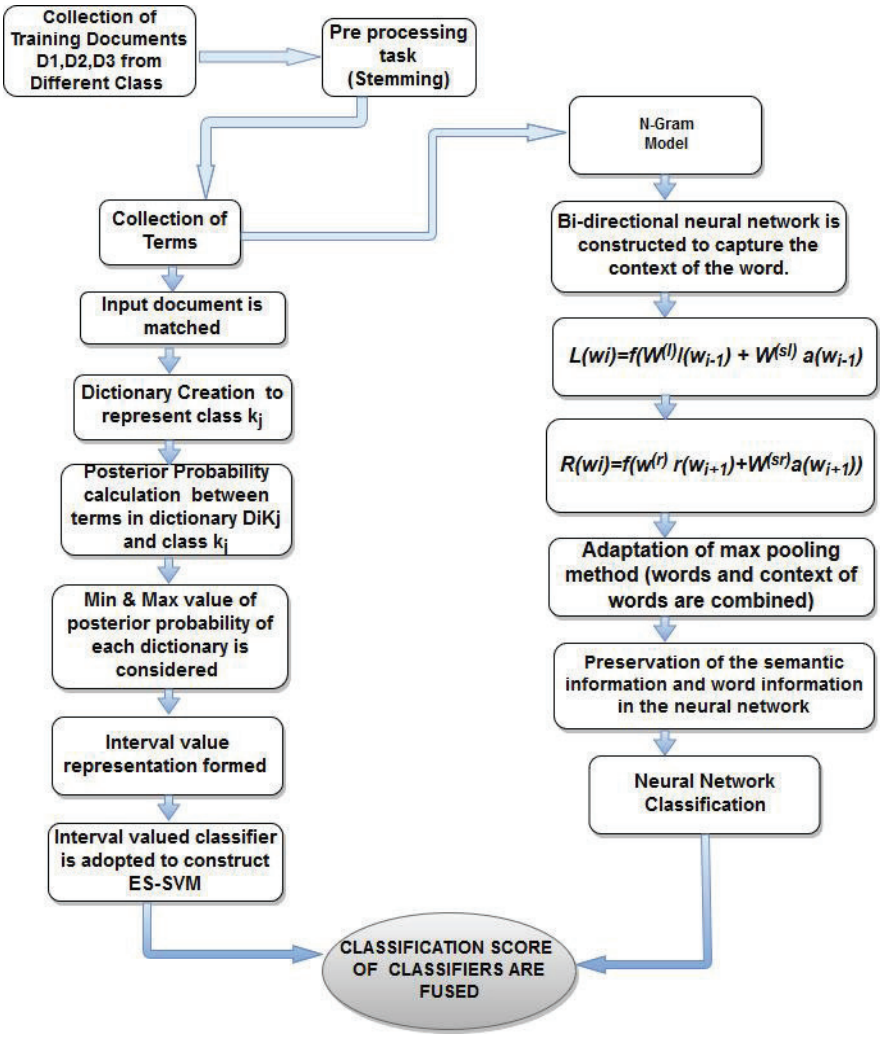

Figure 2. Different steps stages in proposed model.

Table 1. Term occurrences and their Posterior Probabilities respect to all j dictionaries.

Term	Probability dictionary 1	Probability dictionary 2	...	Probability dictionary j
t_1	$P_r(D_ik_1 \mid t_1)$	$P_r(D_ik_2 \mid t_1)$	.....	$P_r(D_ik_j \mid t_1)$
t_2	$P_r(D_ik_1 \mid t_2)$	$P_r(D_ik_2 \mid t_2)$	.....	$P_r(D_ik_j \mid t_2)$
t_3	$P_r(D_ik_1 \mid t_3)$	$P_r(D_ik_2 \mid t_3)$	.....	$P_r(D_ik_j \mid t_3)$
.	.	..... .		.
.	.	.....		
t_n	$P_r(D_ik_1 \mid t_n)$	$P_r(D_ik_2 \mid t_n)$	....	$P_r(D_ik_j \mid t_n)$

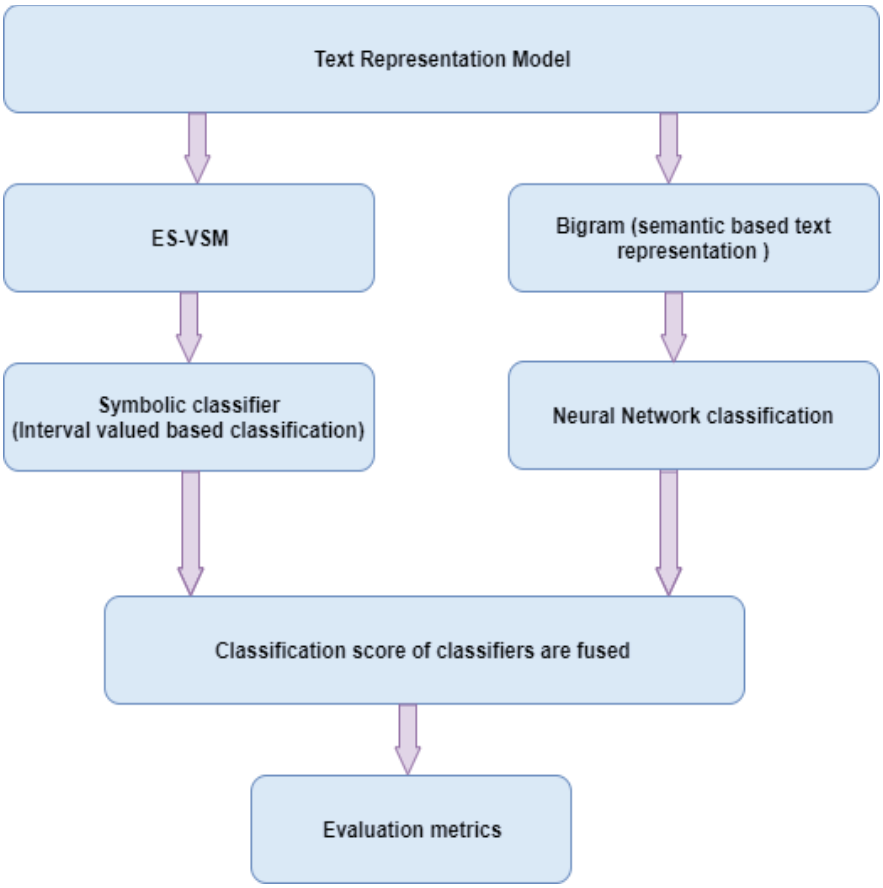

Figure 3. Block Diagram of the proposed fusion based approach for text classification

5.1. Word representation model

One of the popular methods for representing features of a text document is Bag of words representation model. The main limitation of a bag of words representation is it fails to preserve the semantic information of the documents and can be addressed using an enhanced bag of words representation with help of neural network to captures semantic information from the text documents. The neural network helps to construct a sentence level representation by capturing semantic information at a sentence level. One of the advantages factors of the neural network is that its ability to capture meaningful information of the document. But word-level model is having limitation is capturing the context information of word that means it gives more weight to words are not present in the beginning part of the document. Bi-directional recurrent structure contributes less noise during capturing the semantic information of the text documents. Then the max-pooling algorithm is used for selecting important terms of the text documents. The max-pooling method determines the different words in the document.

5.2. Neural network for classification

The neural network is designed to recognize patterns. Neural network addresses the issue of capturing the semantic information for the text documents using neural network model. The input for the neural network will be text document D, which contains the terms in the order t1, t2, t3 …..tn.

5.3. Term level text representation model

To preserve the more precise word information word and context of the words should be combined. A left (L(wi)) and right (R(wi)) neural network are constructed to capture the context of the word.

Two equations are used to preserve the semantic of left and right-hand side context of each word in the document.

$L(wi)=f(W(l)l(wi-1)+W(sl)a(wi-1))$ 2.0

$R(wi)=f(W(r)\ r(wi-1)+W(sr)\ a(wi+1))$ 2.1

a(wi-1) represents word perspective of the word wi-1.

l(wi-1) represents the left side (previous word) perspective of the word wi.

w(l) represents word data matrix which transforms the word perspective to the next layer.

W(sl) represents word matrix which is considered for preserving the semantic information between the present word wi with its previous word.

5.4. Max pooling Techniques for text representation

Max pooling method is adapted to preserve the semantic information and word information in the neural network.

6. Experiment Results

The evaluation of the proposed inference approach is carried out on each text document classification data set. Experimental results show the efficiency of the proposed approach with an average success rate of 95%. The expected accuracy is various with different 60 % of training and 40% test input class of text data sets. The evaluation performance provided from 20 News Group large data sets differ with the text files in particular class. The result of categorized text files are subjected to exact class where it is represented to.

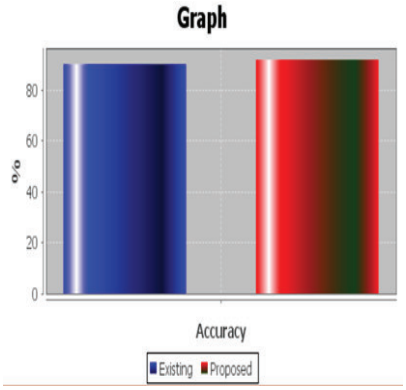

Figure 4. Evaluation performance for Base Ball class of 20Newsgroup large.

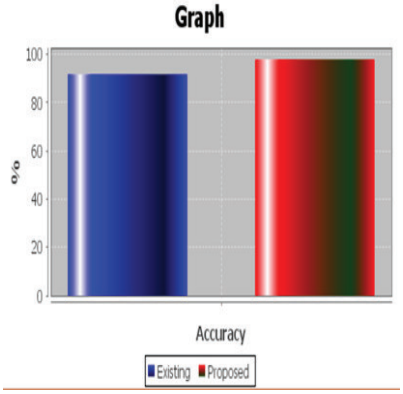

Figure 5. Evaluation performance for Motorcycle class of 20 News Group large.

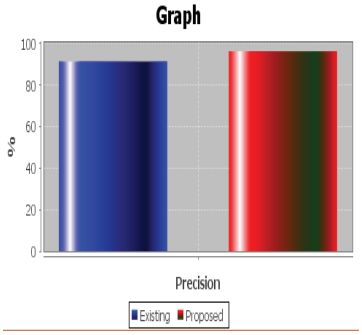

Figure 6. Evaluation performance for Space class of 20 News Group large.

7. Conclusions and Future Work

In this paper, the problem of text classification is addressed by considering two text document representation model by fusing the classifiers. Two different classifiers are designed based on the two different representation models of the text documents. An enhanced Sentence Vector Space Model and interval valued representation model is considered for the effective representation of text documents.

A word level neural network Bigram representation for text documents is proposed for effective capturing of

semantic information present in the text data.

The point ticked below of concern out of the text document categorization area, the machine learning field for the future would be resolved.

- Extending the n-gram model for text representation to more gram for accurately representing which text document can be classified.
- Building a lower dimensional feature by using a novel based Enhanced Sentence Vector Space Model (S-VSM)
- Protecting the information of a text document.
- Planning a Bag of words to represent consecutive terms in the document powerfully.
- Proposing document categorization depends on a combination of a novel technique.

The models discussed in this article are subjected for extensive experimentation on publically available datasets. The experiment results are the manifestations which presents the effectiveness of the proposed models publically available various class of 20 News Group large data sets is better than existing approach of unigram model. In future to provide and gain appropriate non assigned text document of text categorization data sets the more Bag of words can be recommended for which reducing the ambiguity.

References

[1] Bhushan, S. B., & Danti, A. (2017). Classification of text documents based on score level fusion approach. Pattern Recognition Letters.
[2] Li, C. H., Song, W., & Park, S. C. (2009). An automatically constructed thesaurus for neural network based document categorization. Expert Systems with Applications, 36(8), 10969-10975.
[3] Zelaia, A., Alegria, I., Arregi, O., & Sierra, B. (2011). A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension. Applied Soft Computing, 11(8), 4981-4990.
[4] Bhakkad, A., Dharamadhikari, S. C., & Kulkarni, P. (2013). Efficient approach to find bigram frequency in text document using E-VSM. International Journal of Computer Applications, 68(19).
[5] William B.Cavnar., John M. Trenkle (2010). "N-Gram-Based Text Categorization"vol.5 IJCSS
[6] Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, 57, 232-247.
[7] Vieira, A. S., Borrajo, L., & Iglesias, E. L. (2016). Improving the text classification using clustering and a novel HMM to reduce the dimensionality. Computer methods and programs in biomedicine, 136, 119-130.
[8] Selvi, S. T., Karthikeyan, P., Vincent, A., Abinaya, V., Neeraja, G., & Deepika, R. (2017, January). Text categorization using Rocchio algorithm and random forest algorithm. In Advanced Computing (ICoAC), 2016 Eighth International Conference on (pp. 7-12). IEEE.
[9] Zhang, H., & Zhong, G. (2016). Improving short text classification by learning vector representations of both words and hidden topics. Knowledge-Based Systems, 102, 76-86.
[10] Revanasiddappa, M. B., Harish, B. S., & Manjunath, S. (2014, November). Document classification using symbolic classifiers. In Contemporary Computing and Informatics (IC3I), 2014 International Conference on (pp. 299-303). IEEE.
[11] Gulin, V. V., & Frolov, A. B. (2016). On the classification of text documents taking into account their structural features. Journal of Computer and Systems Sciences International, 55(3), 394-403.
[12] Danti, A., & Bhushan, S. B. (2013). Document vector space representation model for automatic text classification. In Proceedings of International Conference on Multimedia Processing, Communication and Information Technology (pp. 338-344).
[13] Castanedo, F. (2013). A review of data fusion techniques. The Scientific World Journal, 2013.
[14] Mintz, T. H., Wang, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts. Cognitive psychology, 75, 1-27.
[15] Yun, J., Jing, L., Yu, J., & Huang, H. (2012). A multi-layer text classification framework based on two-level representation model. Expert Systems with Applications, 39(2), 2035-2046.
[16] Schmidt, S., Schnitzer, S., & Rensing, C. (2016). Text classification based filters for a domain-dspecific search engine. Computers in Industry, 78, 70-79.
[17] Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. Expert Systems with Applications, 80, 83-93.
[18] Liu, C., Wang, W., Tu, G., Xiang, Y., Wang, S., & Lv, F. (2017). A new Centroid-Based Classification model for text categorization. Knowledge-Based Systems, 136, 15-26.
[19] Pinheiro, R. H., Cavalcanti, G. D., & Tsang, R. (2017). Combining dissimilarity spaces for text categorization. Information Sciences, 406, 87-101.