Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 6, June -2017

N-Gram and KLD Based Efficient Feature Selection Approach for Text Categorization

Smita Shedbale[1], Dr. Kailash Shaw[2]

ME Student, Department of Computer Engineering, DYPCOE, Akurdi, SPPU, Pune, India[1]
Associate Professor, Department of Computer Engineering, DPCOE, Akurdi, SPPU, Pune, India[2]

Abstract — *Automated categorization of text into a set of predermined categories has become one of the important approach for handling and organizing enormous amount of web document. Text categorization method is used in a wide variety of applications such as news article categorization, spam filtering. To deal with a major challenge in text categorization that is high dimensionality of the feature space, feature extraction and feature selection plays most important role. In this paper, TF N-gram based and unigram TF-IDF based method will be carried out for feature extraction. Then from extracted feature set, feature selection is done by using Kullback-Leibler divergence measure (KLD). We evaluated the proposed approach on document collections BBC that is news article dataset, originating from BBC News, using classification algorithm, Naive Bayes. Proposed method outperforms in terms of accuracy of text categorization.*

Keywords *- Dimensionality Reduction, Feature Selection, Filter, Information gain, Jeffreys divergence, Kullback-Leibler Divergence, Maximum Discrimination, Text Categorization, Wrapper*

I. INTRODUCTION

As the availability of documents in digital format is increasing remarkably in recent years, it is not possible to organize and exploit manually such huge amount of information. Various machine learning techniques and statistical theory based methods have been extensively applied to text categorization. Text categorization is the process of modeling and building automatic text classifier which assigns one or more thematic categories to new document based on its content. Since early 90's different machine learning algorithms to text categorization has become popular and becomes dominant over period of time [1][2][3].

II.

Automated text categorization has been used in variety of application which includes automatic indexing for Boolean information retrieval systems, document organization, document filtering, word sense disambiguation, and in most of the internet application [4].

III.

In text categorization, each document should be represented in a way that is useful for learning phase of the classifier. "Bag-of-words" is the most commonly used representation of texts within document, where text (Sentence) is represented in the form of collection of features. Features correspond to particular word (term) whose value indicates its importance according to appropriate feature measurement.

The major challenge associated with automated text categorization is learning from high dimensional data. To perform dimensionality reduction it is necessary to reduce feature size [5]. Because document may consists of large number of features ranging from hundreds to few thousands which may result into heavy computational load for the learning process of text categorization. It is possible that many of the terms in a document may be repetitive and may not be relevant so learning process with these all may adversely affect overall performance of the classifier. Hence it is very important to reduce the entire feature space so that the reduced form of feature space contains most useful features which can be used for training the classifier effectively rather than entire original large feature space.

Most common approach of feature reduction in the field of text categorization is feature extraction and feature selection [6]. The feature extraction builds a new feature set by combinations or transformations of the original feature set which will not be same as the original set, whereas the feature selection, which is the most commonly used method in the field of text categorization, selects the optimal feature subset from the original feature set depending on some kind of evaluation criteria. These selected features are then given as input to learning algorithm. Feature selection along with classification algorithm helps to improve overall accuracy.

Variety of feature selection algorithms have been proposed in literature for the application of text classification. These are categorized into two main categories: filter approach and wrapper approach. Filter approach is responsible for selecting the features based upon general characteristic of data [7]. It doesn't include learning algorithm to evaluate the importance of feature. While in case of Wrapper feature selection algorithms,

usefulness of selected features is based upon evaluation criteria specified by learning algorithm [8]. For small dataset wrapper approach usually performs better than that of filter approach as because it receives feedback from learning algorithm about its importance.

Though wrapper approach outperforms than filter but as the number of documents from the dataset increases, evaluation criteria which measures the importance of feature, as it is based upon learning algorithm, it will create heavy load on entire process. Complexity increases as with number of documents in a data set. So, in most of the text categorization approach, mostly filter approach is used because of its simplicity. It usually ranks the feature and then selects top ranked features for classification. It is hard to determine which filter feature selection algorithm performs better when particular learning algorithm is specified.

Many filter feature selection approaches have been ad-dressed well in literature. Section I defines introduction about Automated text categorization and feature selection, section II includes Literature Review about filter selection approaches and section III includes System overview, section IV describes Result analysis and section V includes conclusion.

II. RELATED WORK

Literature Review focuses on following major areas:

A. Representation of Documents

Because of the increasing growth and usage of the internet, there is a need to develop most efficient and useful tools or software's which will assist users to search through internet. Large amount of information on the internet is in the form of textual format. Text categorization is one of the crucial research field within text mining which aims to recognize, understand and organize the volumes of text data or documents. In automated text categorization, term is said to be important parameter for representation of text within document.

There are various document representation models developed in literature that have evolved over through research work in diverse domains, of which "bag-of-words" is one of the widely used approach for text representation. It mainly focuses on representing document using frequency count of each term within that document. There are certain limitations of this representation like ignoring the context of words; some important words may be ambiguous.

To avoid these limitations, term weighting methods are used which assigns appropriate weights to the term according to its importance so that to achieve performance of text classification. Recently, document representation using neural networks have shown greater performance in the application of classification and clustering [9][10]preserving meaning as well as ordering of the words. The modeling approach is the language model which uses n-gram models to capture more contextual information than standard bag-of-words approaches, and employs better smoothing techniques [11]. N-gram Based Text Categorization is a simple method based on statistical information about the usage of sequences of words [12].

B. Naive Bayes Classifier

Most popular classifier in machine learning applications is Naive Bayes model. It allows each feature to contribute towards the final classification independently and equally from other features. This simplicity allows improving over-all efficiency. Naive Bayes allows competitive performance for text categorization compared with other classification method [13][14] such as neural network, support vector machine. Naive Bayes classification method is based upon Bayes rule with strong independence assumptions between the features [15]. For some types of probability models, naïve Bayes classifiers can be trained very efficiently in a supervised learning setting.

In many practical applications, naive Bayes models uses the method of maximum likelihood which means one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods. The assumptions on distributions of features are called the event model of the Naive Bayes classifier. In TC, both Binary and Real-valued feature models (multinomial distributions) have been widely used. In case of Binary-valued feature model (Bernoulli distributions), value of each feature is either 1 or 0 indicating whether particular term occurs in the document or not. In Real valued feature model, generally the feature refers to the term frequency (TF) which is defined as the number times that a particular term appears in the document. For discrete features like the ones found in document classification, multinomial and Bernoulli distributions are popular [15].

Multivariate Bernoulli Naive Bayes the binomial model is useful if feature vectors are binary. One application would be text classification with a bag of words model where the 0s 1s are "word occurs in the document" and "word does not occur in the document". The multinomial naive Bayes model is typically used for discrete counts. Gaussian Naive Bayes is used for features that follow a normal distribution.

Usually when Bernoulli model, multinomial model have been incorporated into the Bayesian framework and have resulted in classifiers of Bernoulli naive Bayes (BNB), multi-nominal naive Bayes (MNB), respectively. For large vocabulary size, extensive experiments on real-life benchmarks have shown that the MNB usually outperforms the BNB. When term frequency is used to represent document, MNB would be one of the best-known naive Bayes classification approach.

C. Feature Dimensionality Reduction

Once feature vectors are generated using N-gram technique, for high value of N, the dimensionality of the feature vectors may be intractable in terms of memory and time requirements. Dimensionality reduction can be carried out by selecting the most relevant features. Total number of features which rep-resents particular document can be reduced by selecting only useful and efficient features for classification and discarding non relevant, redundant features. Main objective of feature selection is to reduce curse of dimensionality which result into increased classification accuracy and to avoid wasting time in processing unnecessary features.

Feature selection is one of the important step of dimensionality reduction which is widely considered in many applications such as classification and prediction to improve overall performance. Feature selection focuses on selecting only subset of relevant and important features and ignores redundant and irrelevant features from the original feature space. Hence it reduces the overall original feature space and gives selected features as input to learning algorithm [16]. Feature selection along with classification algorithm improves overall accuracy of system.

In the literature, feature selection algorithms are broadly categorized into two main approaches, wrapper and filter feature selection. Wrapper approach depends on the feedback from classifier or classification algorithm which finally going to make use of those selected feature. Wrapper feature selection approach used in [17] where a feature is either added or removed at one step which aims to select optimal subset of features. After generating new set of features every time, classifier is trained with those selected features and tested on a validation data set. This approach leads to select and generate better feature set which will help to improve performance of classification. However, this process creates high computational complexity.

While in filter approach, each feature has assigned some value based on its importance measures and features with high importance score are selected and it will be used for classification. As this approach doesn't include learning algorithm for measuring the goodness of selected feature. This approach has less complexity than wrapper. Wrapper approach is generally suitable for small data set. But in case of large dataset, in text categorization filter feature selection approaches are used because of its simplicity.

Some of the filter feature selection approaches that are widely used in literature are based on information theory measures.

1. Document Frequency (DF) is one of the simple and effective feature selection method, which counts the total number of documents in which a particular term occurs. The idea behind document frequency is that the rare terms are not useful for category prediction and maybe degrade the global performance. So if the number of documents in which a term occurs is the largest, the term is retained. The document frequency of a term is calculated as follows:

$$DF(t_k, c_i) = P(t_k, c_i) \qquad (1)$$

2. Mutual Information (MI) is defined as measure of the mutual dependence between two variable [18]. High mutual information indicates a large reduction in uncertainty; low mutual information indicates a small reduction; and zero mutual information between two random variables means the variables are independent. MI measure between term t_k and category c_i can be defined as

$$MI(t_k, c_i) = \log \frac{p(t_k, c_i)}{p(t_k)p(c_i)} \qquad (2)$$

where, $p(t_k, c_i)$ indicates probability of term t_k in a document and document belongs to category c_i. If term and category are independent from each other then Mutual Information will be zero.

3. Information Gain(IG) measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution [19][20]

$$IG(t_k, c_i) = P(t_k, c_i) \log \frac{p(t_k, c)}{p(t_k)p(c)} + p(\overline{t_k}, c_i) \log \frac{p(\overline{t_k}, c)}{p(\overline{t_k})p(c)} \qquad (3)$$

4. Cross entropy for text(CET) [21] is given as

$$\text{CET}(t_k, c_i) = \text{P}(t_k, c_i) \log \frac{p(t_k, c_i)}{p(t_k)p(c_i)} \qquad (4)$$

5. Chi-square measure is proposed in [14] the lack of independence between the term t_k and category c_i which is modeled by using Chi-square distribution.

$$\text{CHI}(t_k, c_i) = \frac{\left[p(t_k, c_i)p(\overline{t_k}, \overline{c_i}) - p(t_k, \overline{c_i})p(\overline{t_k}, c_i)\right]^2}{P(t_k, c_i)P(t_k, \overline{c_i})P(\overline{t_k}, c_i)P(\overline{t_k}, \overline{c_i})} \qquad (5)$$

All of these filter approaches which are almost based on the information theory measures use binary variables, that is whether term t_k present or absent and whether particular document belongs to category c_i or not. Most of the recent papers [22][23] on the use of class specific features which can be used for categorization of texts. Recently, maximum discrimination (MD) [25], is first employed to calculate feature importance for each individual class, and then a global function, such as sum or weighted average, is applied to rank features to select a common feature subset for all classes.

III. SYSTEM OVERVIEW

Main objective of proposed system is to extracts features from the documents and among the extracted features select most relevent features for classification which will improve overall performance of classification.

IV.

In automated text categorization, input will be given as dataset which contains collection of different documents with specified topics.

A. System Architecture

Figure 1 shows system architecture. First, Document preparation and pre-processing is done and then feature extraction and feature selection process will be carried out.

1. Document Preprocessing Phase: At this step, documents are collected, cleaned, and properly organized. The goal of preprocessing phase is to reduce the number of features which was successfully achieved by using number of techniques like stopword removal, text segmentation etc.

2. Feature Extraction Phase: Feature extraction is one of the dimensionality reduction approaches. It usually involves generating new features which are composites of existing features. Generally, information about sentiment is conveyed by adjectives or more specifically by certain combinations of adjectives with other parts of text. Models that assign probabilities to sequence of words are called language models (LMs). The simplest model that assigns probabilities to sentences or sequence of words is called N-gram. Whether estimating probabilities of next words or the whole sentence, the n-gram model is one of the most important tools in speech and language processing. N-gram is simply a consecutive sequence of words of a fixed window size n. In this paper, we focus on a different but simple text representation. In particular here, each feature is considered as a bag of word 4-grams, that is, words of length 4 as a single word. N-gram frequency profile was generated for each document as the first step of feature extraction and then Term Frequency Inverse Document Frequency (TF-IDF) is calculated for each feature as a weight vector of each feature(4-gram) to determine what words in a corpus of documents might be more favourable to consider it as a probability values which will be given to the feature selection algorithm. This phase consists of following steps:

   - In the first step, Feature extraction will be carried out using N-gram term frequency(TF) feature extraction technique.

   - Then unigrams Term frequency-Inverse Document Frequency (TF-IDF) feature extraction will be carried out.

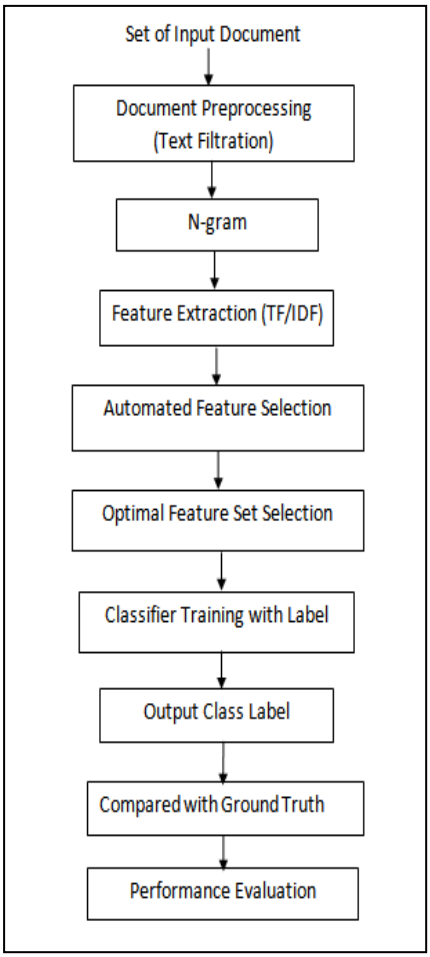

Fig.1. System Architecture

3. Feature Selection Phase: Feature selection is the process of selecting a subset of the original feature set. In this phase symmetric form of KL-divergence measure is used to obtain more efficient feature set. The method proposed in this paper is based on the symmetric Kullback- Leibler divergence, which is well known in Information Theory. In paper [24] text categorization is performed using this distance between the probability distribution of the document to classify and the probability distribution of each category. In information retrieval, the Kullback- Leibler divergence is used for query expansion.
Tang, Bo [25] introduced new divergence measure called JMH measure, and implemented feature selection approach which selects optimal feature set. Feature Selection aims at selecting features which are highly discriminative. It requires an understanding of what aspects of the dataset are important in document categorization, and which are not.
4. Classifier Training with Label Phase: Optimal Feature Set generated in the previous phase will be given as an input to classifier for classifier construction. Naive Bayes is used as an Classifier in this step which will give output as a class label which will be compared with ground truth. Naive Bayes is easy to model and build and mostly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Naive Bayes is a conditional probability model. Using Bayes' theorem, the conditional probability can be decomposed as

$$p(Ck \mid x) = \frac{p(c_k)\,p(x \mid c_k)}{P(x)} \qquad (6)$$

where x=(x1,..xn) represents n features, it assigns to this instance probabilities p(Ck | x1,..xn) for each of K possible outcomes or classes Ck. Above equation can be interpreted as

$$\text{posterior} = \frac{prior \times likelihood}{evidence} \qquad (7)$$

5. Performance Evaluation Phase: Some of the more popularly used measures will be used for estimating the performance of the classification system : Accuracy, F1-measure, precision, Recall etc. Formula's of which is provided in result's section.

IV. ALGORITHM

Proposed System implements following algorithm:

Algorithm 1: N-gram feature extraction and KLD based Feature Selection Approach for categorization

---

Input: Text from data set
Output: A reduced feature set

1. Preprocessing and Feature Extraction using Four-gram technique
2. For i is 1←M (feature) {
3. Denote p_{ic} as the class distribution where c = {1..,N}
4. For c is 1←N (feature) {
5. Form two distributions: p_{ic} and $\overline{p_{ic}}$,where $\overline{p_{ic}}$ is the one grouping all remaining N-1 classes;
6. Calculate KL divergence between p_{ic} and $\overline{p_{ic}}$ }
7. Calculate the JMH-divergence
8. Sort features using cosine similarity.
9. Write wordhashmap.
10) Train Features.

The general pseudo code for N-gram which is considered here as will be as given below:

Pseudocode for N-gram

Input: Text, N (Text- inputText Words, N=2, 3, 4) Output: N-gram

1) for i = 1 → Size(Text) - N + 1 {
2) for n = 0 → n < N {
3) ngram + = Texti + n }
4) nGram.add(ngram) }
5) return nGram (output with N-gram)

IV. RESULTS AND DISCUSSION

A. DataSet

Most widely used dataset for text categorization that is 20-Newsgroup, BBC Datasets is considered here. The 20-NEWSGROUPS benchmark consists of about 20, 000 documents collected from the postings of 20 different online newsgroups or topics. BBC dataset consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005. Class Labels: 5 (business, entertainment, politics, sport, tech).

B. Results

Generally, performance measures such as accuracy, precision, recall and F1-measure is used to evaluate performance of classification. The accuracy metric defines overall classification performance which is defined as,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

The precision metric is the percentage of documents that are correctly classified as positive out of all the documents that are classified as positive. Recall is the metric which defines percentage of documents that are correctly classified as positive out of all the documents that are actually positive. The metrics of precision and recall are defined as,

$$precision = \frac{TP}{TP + FP} \quad (9)$$

$$recall = \frac{TP}{TP + FN} \quad (10)$$

where TP denotes the number of true positive, TN denotes the number of true negative, FP denotes the number of false positive, and FN denotes the number of false negative. Precision and Recall have an inverse relationship with each other. F1 measure is one of the most popular among those measures that attempt to combine precision and recall as one single measure and is defined as,

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (11)$$

Above definitions are defined for binary class classification. For multi class classification, F1 measure is obtained by averaging the F1 measure of each class weighted by the prior class.

TABLE : I
FEATURE EXTRACETD FROM ORIGINAL DOCUMENT WITH MAXIMUM DISCRIMINTION APPROACH

No.of Input Words	Selected Features
1747	1150
4083	3006
9757	5400
14578	7330
19270	9059

Table V.1 shows details about Number of input words from datasets and selected features(instances) obtained after feature extraction phase using MD method.

Proposed approach which is based upon N-gram based feature extraction and KLD based feature selection approach shows significant reduction in features that are selected after execution of this method. Table 8.2 shows original features after N-gram typically using 4-gram approach and shows original number of features and selected number of features after applying feature selection approach. Results of which are given in following table:

TABLE II

FEATURE EXTRACETD FROM ORIGINAL DOCUMENT WITH N-GRAM AND KLD BASED APPROACH

No.of Input Words	Selected Features	Feature Reduction
1720	53	96.91%
4655	128	97.26%
9607	340	96.47%
14353	490	96.59%
18969	624	96.72%

Total Feature reduction obtained using 4-gram and KLD based approach is shown in following figure 2 based on the

results obtained which clearly shows there is high amount of reduction in number of features that are selected and these features will be given as an input to naive bayes learning algorithm.

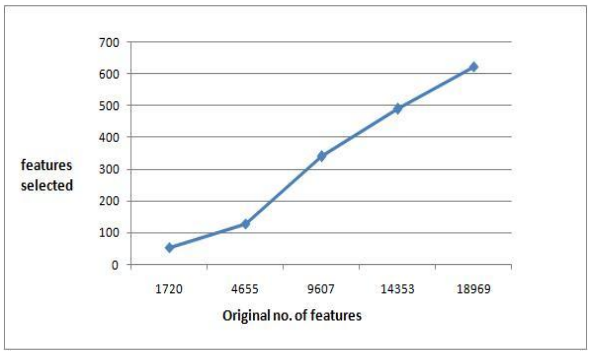

Fig. 2. Feature reduction using N-gram and KLD approach

To compare the performance of proposed feature selection method, with F1-measure is compared with different number of instances(features) along with Maximum Discrimination method. Overall accuracy of proposed method increases as compare to maximum discrimination method which tested over document size. Feature selection method is tested when naive Bayes is used as an classifier. Figure 3 shows the final graph of F1-measure comparison of MD feature selection technique and N-gram(4-Gram) & KLD based approach of feature selection algorithm, where x-axis represents document size(original feature instances) and y-axis represents F1-measure where it clearly shows significant increase in accuracy over existing method.

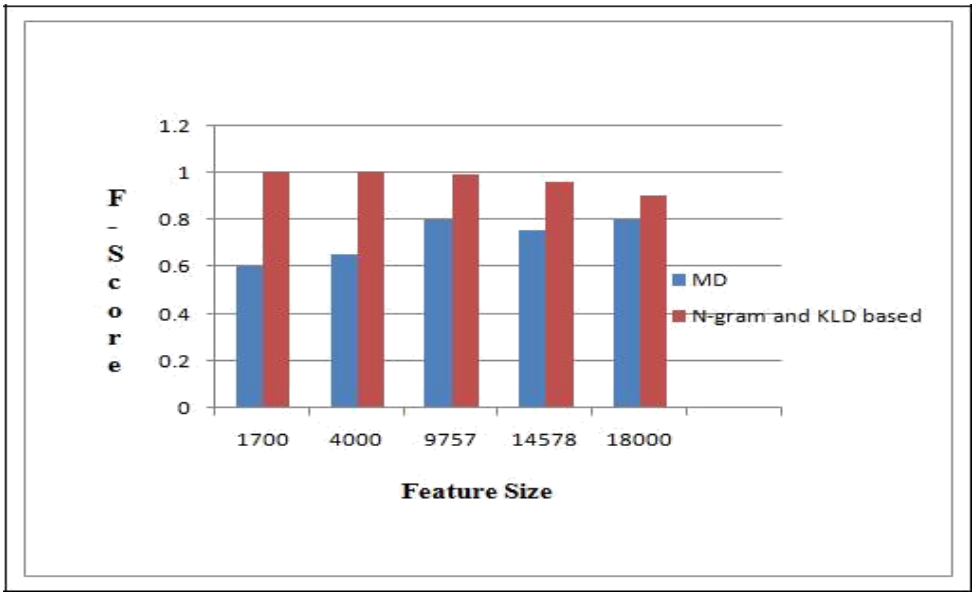

Fig. 3. Comparison of f1-measure of existing and proposed approach

V. CONCLUSION AND FUTURE WORK

Large number of documents are increasing rapidly, there-fore, to organize it in electronic form, text categorization becomes an important and challenging issue. A major issue for text categorization is its large number of features. Most of the features are irrelevant, noisy redundant, which may mislead the classifier. Hence, it is most important to reduce dimensionality of data to get smaller subset and provide the most gain in information.

Existing approach focuses on feature selection method based on the information measures for naive Bayes classifiers, which aims to select the features that offer the maximum discriminative capacity for text classification. Rather than considering only term frequency in calculating probabilities of each feature, implementing N-gram based feature extraction following term weighting approach and KLD divergence measure based feature selection it is possible to acheive more accuracy than existing approach. Approach which makes the use of N-gram and KLD feature selection increases F1 measure by performing feature reduction by around 90-95% than original features that are extracted using N-gram.

REFERENCES

1. Sebastiani, Fabrizio. ”Machine learning in automated text categorization.” ACM computing surveys (CSUR) 34.1 (2002): 1-47.
2. Lam, Wai, Miguel Ruiz, and Padmini Srinivasan. ”Automatic text categorization and its application to text retrieval.” IEEE Transactions on Knowledge and Data engineering 11.6 (1999): 865-879.
3. Joachims, Thorsten. ”Text categorization with support vector machines: Learning with many relevant features.” European conference on machine learning. Springer Berlin Heidelberg, 1998.
4. http://nmis.isti.cnr.it/sebastiani/Publications/ASAI99.pdf
5. Lewis, David D. ”Feature selection and feature extraction for text categorization.” Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992.
6. Nejad, A. Mohammad Behrouzian, et al. "Feature Selection Techniques for Text Classification." International journal of Computer Science & Network Solutions 2.
7. Chen, Jingnian, et al. ”Feature selection for text classification with Nave Bayes.” Expert Systems with Applications 36.3 (2009): 5432-5435.
8. Kohavi, Ron, and George H. John. ”Wrappers for feature subset selection.” Artificial intelligence 97.1 (1997): 273-324.
9. T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.
10. Q. V. Le and T. Mikolov, Distributed representations of sentences and documents, arXiv preprint arXiv:1405.4053, 2014.
11. Farhoodi, Mojgan, Alireza Yari, and Ali Sayah. ”N-gram based text classification for Persian newspaper corpus.” Digital Content, Multimedia Technology and its Applications (IDCTA), 2011 7th International Conference on. IEEE, 2011.

12. Wikipedia contributors. ”N-gram.” Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 4 Jan. 2017. Web. 4 Jan. 2017.
13. A. Genkin, D. D. Lewis, and D. Madigan, Large-scale Bayesian logistic regression for text categorization, Technometrics, vol. 49, no. 3, pp. 291304, 2007.
14. B. Tang and H. He, ENN: Extended nearest neighbor method for pattern recognition [research frontier], IEEE Comput. Intell. Mag.,vol. 10, no. 3, pp. 5260, 2015.
15. McCallum, Andrew, and Kamal Nigam. ”A comparison of event models for naive bayes text classification.” AAAI-98 workshop on learning for text categorization. Vol. 752. 1998.
16. Y. Yang and J. O. Pedersen, A comparative study on feature selection in text categorization, in Proc. Int. Conf. Mach. Learn.,vol. 97, 1997, pp. 412420.
17. R. Kohavi and G. H. John, Wrappers for feature subset selection,Artif. Intell., vol. 97, no. 1, pp. 273324, 1997.
18. Yang, Yiming, and Jan O. Pedersen. ”A comparative study on feature selection in text categorization.” ICML. Vol. 97. 1997.
19. Roobaert, Danny, Grigoris Karakoulas, and Nitesh V. Chawla. ”Informa-tion gain, correlation and support vector machines.” Feature Extraction. Springer Berlin Heidelberg, 2006. 463-470.
20. Caropreso, Maria Fernanda, Stan Matwin, and Fabrizio Sebastiani. ”A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization.” Text databases and document manage-ment: Theory and practice (2001): 78-102.

21. Mladenic, Dunja, and Marko Grobelnik. ”Feature selection for unbal-anced class distribution and naive bayes.” ICML. Vol. 99. 1999.
22. Tang, Bo, et al. ”A Bayesian classification approach using class-specific features for text categorization.” IEEE Transactions on Knowledge and Data Engineering 28.6 (2016): 1602-1606.
23. Tang, Bo, et al. ”EEF: Exponentially Embedded Families with Class-Specific Features for Classification.” (2016).
24. Bigi, Brigitte. ”Using Kullback-Leibler distance for text categorization.” European Conference on Information Retrieval. Springer Berlin Heidel-berg, 2003.

25. Tang, Bo, Steven Kay, and Haibo He. ”Toward optimal feature selection in naive Bayes for text categorization.” (2016).

