# Text Classification: A Survey on Techniques and Methods

Mukund N. Helaskar
Department of Computer Engineering
Pune Institute of Computer Technology, Pune

Dr. Sheetal S. Sonawane
Department of Computer Engineering
Pune Institute of Computer Technology, Pune

Abstract—Growth in the internet made immense impact on the data generation. Most of the data in the world is in textual format. There is need to access and use this data efficiently and easily due to this text classification is widely studied problem in research community. The applications of classification are also in diverse domains such as news filtering, opinion mining, information retrieval and so on. The problem is attempted to solve with Natural language processing, Machine learning algorithms. In the present work the brief literature survey of text classification work is presented.

Index Terms—Data Mining, Information Retrieval(IR), Machine Learning(ML), Natural Language Processing(NLP),Text classification, Word Embedding.

## I. Introduction

Text classification is broadly studied area in data mining. There is exponential increase in the online availability of information. Accessibility is from variety of sources like digital libraries, social network feeds, scientific literature, e-books and so on. Data present is in the structured and unstructured form. There is need to handle data of such magnitude efficiently. Main goal of the data mining is to extract useful information from textual data which deals with operations like information retrieval, classification.

There are many machine learning, Natural language processing approaches are proposed for text classification. Text classification problem is defined as follows for given set of documents $D = \{d_1, d_2, ....d_n\}$ each with assigned label from set $L = \{l_1, l_2, ....l_k\}$ based on the features of document d classification model should assign correct label l. Text classification is of two ways single label and multi-label.

Challenges are likewise there in text classification as it is a complex process. Representation of a text is one issue because machine will not understand the raw text data as input. There- fore representing text into machine understandable format i.e in numbers is necessary. Most common way is to represent text as Bag of Words(BOW)[1]. BOW considers the number of occurrences of the word in text. But there are disadvantages of BOW model. Order is not taken into consideration. For calculating the similarity exact term matching is needed. But two documents may have the same information using distinct words. This issue is addressed in Word2Vec[2] model which represents word as continuous vectors in distributed vector space. These vectors have given state of art result in many NLP tasks.

## II. System Architecture and Overview

Text classification system consists of various modules shown in figure 1 and described below:

### A. Dataset preparation

First step is the data collection. It depends on what problem needs to be solved and domain of the problem. For e.g news articles, opinion mining, medical data etc. The documents should be in text format.

### B. Pre-processing

Pre-processing of the text is key component. It has noticeable impact on the results. It consists on multiple steps such as:

1) Tokenization: Text is made up of words/phrases. Tokenization converts text into tokens(words/phrases) as basic elements.

2) Removing stop words: Every token(word/phrases) will not be of significance in the text. Hence in stop word such as "a", "and", "the" etc which occur frequently in text are removed.

3) lemmatization: lemmatization groups the words according to their grammatical context. There are different inflected forms of single words like playing,plays,played words represent play which is base word also known as lemma.

4) Stemming: Stemming converts different words into similar canonical form like computing, computation to compute which is known as stemm.

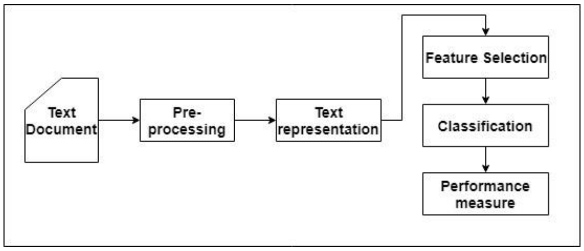

Fig. 1. Text classification System Architecture

### C. Text representation

Text is to be represented in the machine language to process as raw text can't be used as input for machine. Most commonly used method to represent text is Bag of Words(BOW)[1]. It represents text as group of words present in text using their occurrence frequency. It doesn't maintain any sequence. Text documents are represented as document matrix where each row is vector made of the words in documents. BOW has limitations like high dimensionality of the representation. Sparseness of the representation, semantic relations are also not inherited.

Latent Semantic Indexing[3] is another technique which uses singular value decomposition to understand structure of documents and identify hidden(latent) relationship between words. But it works better for small set of static documents which is well used in Information retrieval. Word2Vec[2] method gives continuous vector representation of word. These vector along with low dimensional representation inherit semantic relations as well as analogical relation between words. for e.g vec(Paris)- vec(France) + vec(Berlin) = vec(Germany). Word2Vec model overcomes problems in BOW model like high dimensionality of the representation, Better performance on word to word similarity.

### D. Feature Selection

After the pre-processing and text representation is done then next step is feature selection. Feature selection is necessary step which improves the efficiency of the classifier. Feature selection the process of selecting subset of feature from original text. Due to high dimensionality problem there are metrics for the feature selection also like term frequency/ inverse document frequency(tf/idf), Information gain and so on. In tf/idf the motive is to find important words to facilitate classification. For collection of N documents where d is individual document t is word in document d the tf/odf weight is given in equation (1):

$$W_{t_d} = f_{t_d}/n_{t_d} * \log(N/f_{d_t}) \qquad (1)$$ Where

$W_{t_d}$ is the tf/idf weight of the term t in document d.
$f_{t_d}/n$ is term frequency(tf) which is ratio of no. of occurrences of t in document d to total number of terms in d. $\log(N/f_{d_t})$ is inverse document frequency(idf) which is log of ratio of total no.of documents N to no.of document d in which toccurs. Words/features with high weight are used for classification.
Decision tree is built with top down approach which involves partitioning of the data into subsets. ID3[9] uses entropy and information gain to select the node as root node to split the data in homogeneous subsets.

$$\text{Entropy} = \sum_{i=1}^{c} -p_i * \log_2 p_i \qquad (2)$$

Entropy is given in equation(2). Gain is the difference between the entropy of all data points as a attribute and the entropy of single attribute with data points for more refer[9][10] Making dictionary of the important keywords and using those for classification task is also one of the approach. Important words for each category are selected at first. Then classification of text is done based on the presence/absence of those selected words. Probability of the word is also considered for classification.

### E. Classification

Classification of the text documents into the predefined categories is the objective. Motive is to automate this task in efficient way. Text documents can be classified by three ways supervised,unsupervised,semi-supervised methods. From last few years this problem is extensively studied with rapid progress like use of various machine learning algorithms such as Bayesian classifier, Decision tree, K-Nearest neighbor, Support Vector Machines(SVM), Neural Networks.

### F. Performance measure

Last step of the classification system is the evaluation of the results. To evaluate the results set of documents are set a side(test set). Classifier is trained on remaining set of documents(training set. After the training the we classify test set and compare the estimated result with original labels. The percentage of the correctly classified documents to total documents is called accuracy. There are other common metrics are also there along with accuracy which are precision, recall, F-1 score.

## III. Review of Literature

Number of methods are discussed in literature of text classification. Broadly these are classified in three ways which is supervised,unsupervised and semi-supervised methods shown in figure 2. In supervised methods there are many learning algorithms such as Naive Bayes(NB) classifier, Decision tree classifier, K-Nearest Neighbour(KNN), Support Vector Machine(SVM).

### A. Supervised Methods

Naive Bayes is probabilistic classifier which is based on Bayes rule. Goal is to calculate the probability of document D belongs to class C. There are two models used for Naive Bayes classification which are multivariate Bernoulli[4] and multinomial model[5]. In multivariate Bernoulli presence and absence of a feature or word in considered and frequency of word are ignored. In multinomial the frequency of word is considered. So if the vocabulary is large later model will perform better. In [6] authors propose model based on NB classifier which discriminatively learns feature for NB. Like- wise in [7] the pre-normalization of text and feature weighting is introduced to improve the NB classifier performance. For pre-normalization is done using Poisson distribution. [8] uses NB for feature selection for text classification.

Decision tree classifier divides data hierarchically based on the condition on the selected feature of text. ID3[9] and C4.5[10] are well known decision tree algorithms. These algorithms are used to select features and design a tree to

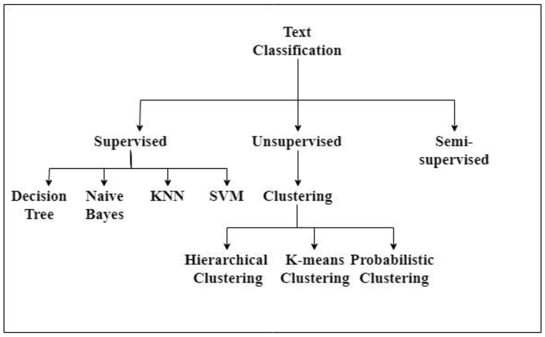

Fig. 2. Types of text classification

classify the text documents. Advantage of the decision tree is it is robust to noise, learns decision rules and easy to use. Due to higher dimensionality of text the performance of the decision tree is less efficient. Boosting[11] technique is applied to solve this issue. In [12] authors present decision tree based text categorization system. Fast decision tree algorithm proposed for text classification applications and method to convert decision tree to simplified rules.

K-Nearest Neighbor classifier uses distance measure for classification. Idea is that documents in a class will be similar to other documents in class. Distance measure used are Euclidean distance, cosine distance. Along with this Word Movers distance using word embedding is given state of art result in text classification. Optimal value of k is often a problem using KNN. In [13] author propose KNN based text classification. They solve the problem of selection of value k for different data and dependency of the result on k value with reasonable increase in accuracy. [14] proposes the model for improvement in the result of KNN by removing outliers from training set. The weight for each feature in KNN is identical. But for each class the importance of the each feature is different. Considering this issue [15] propose Weighted KNN algorithm which assigns weight for feature according to the class.

Support Vector Machines(SVM) is linear classifier. SVM attempts to find hyper-plane with maximum distance to clas- sify documents. Documents which are near hyper-plane know as support vectors which specify the location of the hyper- plane. Advantage if the SVM is it works well with high dimensional data. Support vector discriminate positive class from negative class. [16] introduce SVM for text classification. Authors present experimental and theoretical proof that it is well suited classifier for text classification. In [17] introduce new learning algorithm for SVM to handle unlabelled data effectively. SVM is used together with Bayes rule[18] and Hidden Markov Model[19]. Multi-label[20] text classification is also done using SVM.

For all the classifiers discussed the feature representation mostly done using BOW which has its disadvantages. Word embedding representation addresses issues in BOW and also gained popularity due to state of art result in many NLP tasks. These are low dimensional continuous valued vector representation of words. These representation inherit many semantic and analogical word relations. Word2Vec[2] and Glove[21] are two popular model which generate word embeddings. Word2Vec generated are excessively used due to better results. In [22] author propose a document distance measure which have given impressive results. In NLP community vector representation of word are widely studied in recent years. Text classification [23] [24] is also done using different vector representations.

### B. Unsupervised Methods

In unsupervised method the output labels are not available for learning of the classifier. Clustering is useful in such cases. Clustering is grouping the similar documents in a cluster using similarity function. Hierarchical clustering, K-means clustering and probabilistic clustering are most common text clustering algorithms. In Hierarchical clustering the grouping is done in the hierarchy i.e top-down or bottom up. Clustering is done based on the distance between the features. In K- means clustering algorithms groups n documents in k clusters. Finding optimal k value is main problem in this algorithm. In [25] efficient K-means clustering algorithm is given. In probabilistic clustering the documents are clustered based on the word occurrences. Topic modelling is done using these models. Two main topic models are probabilistic Latent Semantic Analysis(pLSA)[26] and Latent Dirichlet Allocation(LDA)[27]. Documents are made of different topics and topic is group of words. Finding the probabilistic distribution of words over the documents will help in topic modelling. Topic modelling is extensively applicable in Information re- trieval.

### C. Semi-supervised Method

Collection of text with labelled classes is difficult and time consuming process and it may not be feasible to do so. Meanwhile Unlabelled data collection is easy. Semi-supervised method uses large amount of unlabelled data with some amount of labelled data to design better classifiers. In [28] authors present extensive overview on semi-supervised text classification.

## IV. Dataset

There are some standard dataset which are mostly used for the comparison of the work in text classification which are described below.

### A. Reuters21578

The Reuters21578 collection consists of 21,578 news articles. Articles are categorized in 90 categories. This is standard dataset used for text classification research. Although partition of the articles is not uniform for all categories. Dataset can be downloaded from:
https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+ categorization+collection.

### B. 20Newsgroups

This is collection of articles of 20 different newsgroups. Each category contains about 1000 articles each labelled with newsgroup. This is uniformly partitioned in all categories. Dataset can be downloaded from:
http://qwone.com/ jason/20Newsgroups/.

### C. Stanford Sentiment Treebank(SST)

This dataset contains movie reviews in one sentence.10605 snippets are there in the dataset. The categories are very negative, negative, neutral, positive, very positive. The categories are based on the cutoff score of sentiment of snippet. This is used for sentiment classification which is broadly a text classification task. Dataset cane be downloaded from :
https://nlp.stanford.edu/sentiment/.

## V. Comparative observation

Brief review of all the present work done is in section III. The observation is that classifiers perform differently with respect to the dataset. Representation of the text is also an major factor for the efficiency of the task. In general classifiers SVM performs better due to it's ability to handle high dimensional data with BOW representation. The vector representation obtained by neural network have given state of art result. Many task specific vector representations are also presented with better results than those popular representations. Regardless of all work there is no general model for text classification task.

## VI. Application

- News Filtering
  The number of news generated is at par nowadays. Besides viewer needs a specific kind of news to read. Categorization of the new articles is necessary to ease the news broadcasting.
- Search Engines
  The content of search engines is in abundant size. Classification and grouping of relevant content is very necessary for search engine performance.
- Social media mining
  Any trending topic in social media will be in mention of majority of user posts. Classification of such topics or hashtags is also a application of text classification.
- Marketing
  In marketing the views, feedback, rating of the users for any product is classified using text classification. Classification brings simplification for marketing strategies of the product.
- Academics, Government, Medical
  The data generated in the these sectors is really huge. Organization of such data will be simplified if the contents are well categorized.

## VII. Conclusion

In this paper we attempted to give brief introduction of text classification approaches. Work presented in this area is compared based on the classifier used and the accuracy of result. Dataset used and the size of each text document in it also has an impact on the results. Using BOW representation of text SVM gives better results with these representations. Recent work in this area is mostly based on vector representation using neural networks. Using these as representation for the text has given state of art results for text classification task. Consideration of the task specific features also shown improvement in the results.

## References

[1] Harris, Zellig. Distributional structure Word, 1954.

[2] Mikolov, Tomas, et al. ”Efficient estimation of word representations in vector space.”arXiv preprint arXiv:1301.3781(2013).

[3] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G.W., and Harshman, R. A. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391407, 1990.

[4] Joe, Harry. Multivariate models and multivariate concepts Chapman and Hall/CRC, 1997.

[5] D. Lewis and W. Gale. A sequential algorithm for training In SIGIR-94, 1994.

[6] Yirong Shen and Jing Jiang Improving the Performance of Naive Bayes for Text ClassificationCS224N Spring 2003

[7] Sang- Bum Kim, et al, Some Effective Techniques for Naive Bayes Text Classification IEEE Transactions on Knowledge and Data Engineering, Vol. 18, November 2006.

[8] Dino Isa Text Document Pre-Processing Using the Bayes Formula for Classification Based on the Vector Space Mode, Computer and Informa- tion Science November, 2008

[9] Quinlan, J. (1986) Induction of decision trees. Machine Learning, 1, 81106

[10] Quinlan, J. (1993) C4.5:Programs for Machine Learning. Morgan Kaufmann, San Matteo, CA.

[11] Freund, F. and Schapire, R. (1995) A decision-theoretic generalization of online learning and an application to boosting. In Proc. Second European Conference on Computational Learning Theory, pp. 2337.

[12] D. E. Johnson F. J. Oles T. Zhang T. Goetz, A decision-tree-based sym- bolic rule induction system for text Categorization, by IBM SYSTEMS JOURNAL, VOL 41, NO 3, 2002

[13] Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, KNN Model-Based Approach in Classification, Proc. ODBASE pp- 986 996, 2003

[14] Kwangcheol Shin, Ajith Abraham, and Sang Yong Han, Improving kNN Text Categorization by Removing Outliers from Training Set, Springer-Verlag Berlin Heidelberg 2006.

[15] Fang Lu Qingyuan Bai, A Refined Weighted K-Nearest Neighbours Algorithm for Text Categorization, IEEE 2010.

[16] Joachims, T. Text categorization with support vector machines: learning with many relevant features. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, DE), pp. 137142 1998.

[17] Tong, Simon, and Daphne Koller. ”Support vector machine active learning with applications to text classification.” Journal of machine learning research 2.Nov (2001): 45-66

[18] Loubes, J. M. and van de Geer, S Support vector machines and the Bayes rule in classification, Data mining knowledge and discovery 6 259-275.2002

[19] Chen donghui Liu zhijing, A new text categorization method based on HMM and SVM, IEEE2010

[20] Yu-ping Qin Xiu-kun Wang, Study on Multi-label Text Classification Based on SVM Sixth International Conference on Fuzzy Systems and Knowledge Discovery 2009

[21] Pennington, Jeffrey, Richard Socher, and Christopher Manning. ”Glove: Global vectors for word representation.”Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[22] Kusner, Matt, et al. ”From word embeddings to document dis- tances.”International Conference on Machine Learning. 2015.

[23] Lenc, Ladislav, and Pavel Krl. ”Word Embeddings for Multi-label Document Classification.”Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. 2017.

[24] Xing, Chao, et al. ”Document classification with distributions of word vectors.”Signal and Information Processing Association Annual Summit and Conference (APSIPA),2014 Asia-Pacific. IEEE, 2014.

[25] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. 2002. An efficient k-means clustering algorithm: Analysis and implementation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 7 (2002), 881892.

[26] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 5057.

[27] DavidMBlei, AndrewY Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. the Journal of machine Learning research 3 (2003), 9931022.

[28] Nigam, Kamal, Andrew McCallum, and Tom Mitchell. ”Semi- supervised text classification using EM.” Semi-Supervised Learning (2006): 33-56.

[29] Zhu, Xiaojin Jerry. Semi-supervised learning literature survey. Univer- sity of Wisconsin-Madison Department of Computer Sciences, 2005.