ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.7

News Text Classification Model Based on Topic Model

Prajakta Jaybharat Jadhav[1], Prof. Torana N. Kamble[2]
ME Student[1], Assistant Professor [2]
Department of Computer Engineering
Bharati Vidyapeeth Collage of Engineering, Navi, Mumbai India

Abstract:
Text-based classification is a technique which may be used to identify different types of data from the applications point of view. In this project we are going to develop, a text-based classification can be used to classify input text into categories, as defined by the user. The classifier is first trained with an initial dataset using historical data. After the training process is complete, the classifier makes use of the trained data in order to classify any new input text that may be provided. The proposed model also offers an incremental approach to text classification as it dynamically trains the classifier from a new set of data provided by the users.

Keywords: latent dirichlet, text classification, Topic, Model, stop word process, stemming process.

I. INTRODUCTION

In recent years, with the continuous development of information technology, information data of internet increases explosively. The major news websites have become the main platform for human to get news information now. However, the news data of news portal is increasing, which also brings some challenges to the site. The traditional text classification methods have been unable to meet the needs of the current social development. So the research on text classification model is always a hot topic in the field of the text mining in recent years.

The news text classification system can quickly handle with all text data fast, and make accurate prediction of the classification labels. So automatic classification can help to complete text classification function for news platform with high efficiency, and it can also help the company to save expenses. In the era of big data, the research on automatic text classification plays an increasingly important role. Many of the classic text classification algorithms have been proposed and widely used. For example, Support Vector Machine, Naive Bayes and Decision tree and so on. And they are very commonly used classification algorithms.

However, each kind of classification algorithms has different strengths and weaknesses. It is the strengths and weaknesses that decide different classification algorithms for different scene. With the rapid development of social media sites, a lot of user generated content is being shared in the Web, leading to new challenges for traditional media retrieval techniques. An event describes the happening at a specific time and place in real-world, and it is one of the most important cues for people to recall past memories. The reminder value of an event makes it extremely helpful in organizing human life. Thus, organizing media by events has recently drawn much attention within the multimedia research community. Along with the development of information technology, all kinds of information resources' stock and growth have shown massive feature. And text data always occupies a very important position.

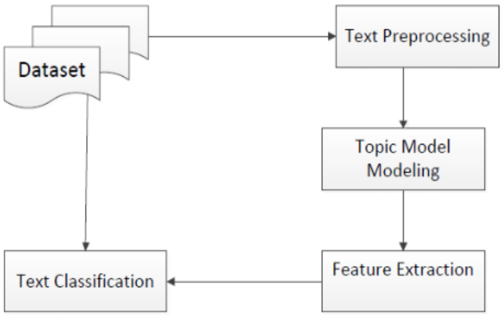


Figure.1.Example of Classification Process

People have huge needs about managing and using the text information effectively, which promotes automatic text classification technology's rapid development and extensive application. The massive amount of digital multimedia content available to us today necessitates some good methods for retrieval, organization, and management. For example, imagine that we are given an extremely large collection of text documents. It would be desirable for each document to have some sort of "short description" that could quickly tell us what it is about. Going further, it would also be useful for these short descriptions to have representations that are consistent across the entire collection. This way, we may use it to determine how closely related one document is to another. We may even use it to visualize where a particular document stands relative to all other documents in the corpus in terms of content similarity. One can see that having these short descriptions would be invaluable for searching and indexing a large collection of text data. Latent Dirichlet Allocation (LDA) is an algorithm that specifically aims to find these short descriptions for members in a data collection. Originally proposed in the context of text document modeling, LDA posits that one way of summarizing the content of a document quickly is to look at the set of words it uses. Because words carry very strong semantic information, documents that contain similar content will most likely use a similar set of words**.** As such, mining an entire corpus of text documents can expose sets of words that frequently co-occur within documents. These sets of words may be intuitively interpreted as topics and act as the building blocks of the short descriptions. The objective is to cluster the documents using the keywords to improve the quality of cluster to great extent. We have selected technologies, entrainment, business as domains for clustering the news in the domains. A concept dictionary maintained which consist of domain specific keywords. Then LDA algorithm is used for clustering from word accuracy.

II LATENT DIRICHLET ALLOCATION

In natural language processing, latent Dirichlet allocation (LDA) is a Generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery by David blei, andrew ng and michael I. jordan in2003. Latent Dirichlet Allocation (LDA) is a kind of topic model algorithm based on probability model. The algorithm thinks that each article is composed of a plurality of topic mixture. It can identify potential hidden information topic in large-scale document set. The algorithm assumes that each word in the corpus in an article is through by "with a certain probability to choose a topic, and then from this subject with a certain probability to select a word". For each text, it chooses a topic from topic distribution and then chooses a word from the word distributions from corresponding topic. Finally it repeats the above procedure until it makes the traversal of the document every word.

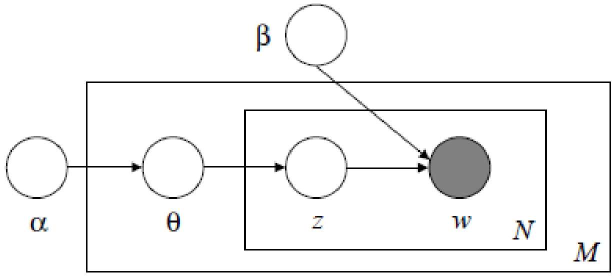


Figure. 2. Graphical model representation of lda.

The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

M= number of documents

N= total number of words in all documents; sum of all N_d *values*

α =collection of all k values, viewed as a single vector

β = collection of all, viewed as a single vector

Z= identity of topic of all words in all documents

W= identity of all words in all documents

The LDA model is represented as a probabilistic graphical model in Figure 2. As the figure makes clear, there are three levels to the LDA representation. The parameters a and b are corpus level parameters, assumed to be sampled once in the process of generating a corpus. The variables qd are document-level variables, sampled once per document. Finally, the variables zdn and wdn are word-level variables and are sampled once for each word in each document. It is important to distinguish LDA from a simple Dirichlet-multinomial clustering model. A classical clustering model would involve a two-level model in which a Dirichlet is sampled once for a corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable. As with many clustering models, such a model restricts a document to being associated with a single topic. LDA, on the other hand, involves three levels, and notably the topic node is sampled repeatedly within the document. Under this model, documents can be associated with multiple topics.

III.COMPARING WITH OTHER ALGORITHM

K-Nearest Neighbor: The K-Nearest Neighbor algorithm (KNN) is among the simplest of all classification algorithms: KNN is a type of instance-based learning; it classifies objects based on the k closest training examples in the feature space. An object is classified by a majority vote over its neighbor's classes, with the object being assigned to the class most common amongst its k nearest neighbors. Despite its simplicity, KNN has been successful in a large number of classification and regression problems. It is often successful in classification situations where the decision boundary is very Irregular. kNN stands for k-nearest neighbor classification, a well-known statistical approach which has been intensively study in pattern recognition for over four decades. kNN has been applied to text categorization since the early stages of the research. It is one of the the top-performing methods on the benchmark Reuters corpus (the 21450 version, Apte set); the other top-performing methods include LLSF by Yang, decision trees with boosting by Apte et al., and neural networks by Wiener et al. The kNN algorithm is quite simple: given a test document, the system _nds the k nearest neighbors among the training documents, and uses the categories of the k neighbors to weight the category candidates. The similarity score of each neighbor document to the test document is used as the weight of the categories of the neighbor document. If several of the k nearest neighbors share a category, then the per-neighbor weights of that category are added together, and the resulting weighted sum is used as the likelihood score of that category with respect to the test document. This simple method does not allow the system to assign multiple categories to any document and is not necessarily the optimal strategy for kNN, or any classi_er, because documents often have more than one category.

IV.IMPLIMENTATON AND RESULT

For the implementation we used java software that is JDK with net Beans IDE. For the results we used three parameter that is precision, recall, F1-Measures. Precision is the percentage of relevant documents and the detection of all documents. Recall is the ratio of the total amount of related literature and literature in retrieval system. Among them, F1-Measure is the key index of the experimental result. F1_Measure is the harmonic mean of precision and recall.

$$\text{Recall}=\frac{Number\ of\ correct\ perdication\ in\ domain}{total\ number\ of\ documents}$$

$$\text{Precision}=\frac{number\ of\ correct\ perdiction\ in\ domain}{Return\ number\ of\ documents\ in\ domain}$$

$$\text{F1}=2*\frac{recall*precision}{recall+precision}$$

As shown in table I, in this we are taking 5 domain that is Business, Crime, Entrainment, Technology, Medical .we collect all the documents, so we have to check the performance using precision,recall,F1-measures.for that we need training dataset and test dataset.

Table .1. Dataset of experiment

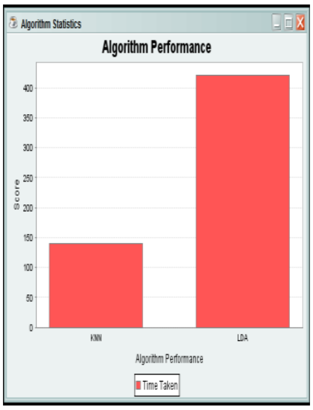
Domain number	Domain name	Number of documents
1	Business	140
2	Crime	50
3	Entrainment	70
4	Technology	58
5	Medical	100

As shown in table II, here on documents apply the precision, recall, f1-measures for KNN and LDA.

Table .2. Dataset of experiment

	KNN	LDA
Precision	53.8462	95
Recall	100	91.6666
F1-measures	70	93.3036

In this experiment, it tries different topics on that we apply precision, recall, f1-measures.It can be seen the obvious and different result of classification model according to different topics and here the results found as well as news text classification performs well.



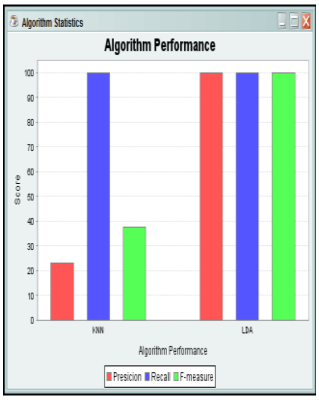


CHART I. ALGORITHM PERFORMANCE CHART II. ALGORITHM PERFORMANCE

V. CONCLUSION

The news text classification system can quickly handle with all text data fast, and make accurate prediction of the classification labels. Automatic classification can help to complete text classification function for news platform with high efficiency, and it can also help the company to save expenses. To reduce the features dimension of the news text and get good classification results. To produce good quality clusters. To improves the scalability and efficiency.

VI. REFERENCES

[1]. S. Jinshu, Z. Bofeng, and X. Xin, “Advances in machine learning based text categorization,” *Journal of Software*, vol. 17, no. 9, pp. 1848–1859, 2006.

[2]. G. Salton and C. Yang, “On the specification of term values in automatic indexing,” *Journal of Documentation*, vol. 29, no. 4, pp. 351–372, 1973.

[3]. Z. Cheng, L. Qing, andL. Fujun, “Improved VSM algorithm and its application in FAQ,” *Computer Engineering*, vol. 38,no. 17,pp. 201–204, 2012.

[4]. X. Junling, Z.Yuming, C. Lin, andX. Baowen, “An unsupervised feature selection approach based on mutual information,” *Journal of Computer Research and Development*, vol. 49, no. 2, pp. 372–382, 2012.

[5]. Z. Zhenhai, L. Shining, and L. Zhigang, “Multi-label feature selection algorithm based on information entropy,” *Journal of Computer Research and Development*, vol. 50, no. 6, pp. 1177– 1184, 2013.

[6]. L. Kousu and S. Caiqing, “Research on feature-selection in Chinese text classification,” *Computer Simulation*, vol. 24, no. 3, pp. 289–291, 2007.

[7]. Y. Yang and J. Q. Pedersen, “A comparative study on feature selection in text categorization,” in *Proceedings of the 14th International Conference on Machine Learning*, pp. 412–420, Nashville, Tenn, USA, July 1997.

[8]. Q. Liqing, Z. Ruyi, Z. Gang et al., “An extensive empirical study of feature selection for text categorization,” in *Proceedings of the 7th IEEE/ACIS International Conference on Computer and Information Science*, pp. 312–315, IEEE, Washington, DC, USA, May 2008.

[9]. S. Wenqian, D. Hongbin, Z. Haibin, and W. Yongbin, “A novel feature weight algorithm for text categorization,” in *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE ’08)*, pp. 1–7, Beijing, China, October 2008.