ISSN: 0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Leveraging Lexical and Semantic Features for Improved Classification of Short Texts

***Dr.C.Jaya Prakash*[1]*., Kammari Sindhuja*[2]**

1 Professor, Department of CSE, Malla Reddy College of Engineering for Women.,
Maisammaguda., Medchal., TS, India
2, B.Tech CSE (20RG1A05L7),
Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

Article Info

Received: 29-06-2023 *Revised: 17-07-2023* *Accepted: 28-07-2023*

Abstract

In this study, we offer a new method for categorizing brief texts by integrating lexical and semantic characteristics. We offer a refined metric for selecting lexical characteristics and then use a reservoir of prior knowledge that spans the domains of interest to identify relevant semantic features. When words and meanings are put together, it creates done by assigning varying weights to words in a map of themes. The number of features is reduced to the number of subjects in this manner. Our classification system, a Support Vector Machine (SVM), uses Wikipedia articles as training data. Results from our experiments demonstrate that, in comparison to other strategies for labelling brief texts, our approach is more successful.

1. Introduction

In a wide variety of fields, text categorization plays a crucial function. Web applications like social networks, online review systems, etc. have increased the amount of brief messages and news we encounter daily. For the automated categorization of brief texts, traditional text mining techniques have limitations. Texts, such as the absence of detail in a sentence's context and the casual language used to explain ideas.

In order to solve these issues when categorizing brief writings, it is usual practice to supplement the original texts with extra information. One technique is to use search engines, then use the information gleaned from those searches to create additional material that provides further context [1, 2, and 3]. Another option is to supplement your expertise with data from third-party sources (like Wikipedia or the Open Directory Project) [4, 5, 6, and 7]. Although these two approaches enhance short text categorization to varying degrees, there is a disadvantage in dealing with the quantity of irrelevant and noisy information if we naively enlarge original texts. In text mining, probabilistic latent topic models [6, 8, 9, and 10] have been employed successfully. These types of models often presume that each text has a multinomial distribution across the themes that have been learned from domain-specific datasets. Because there are so few subjects to cover, the texts' vector space is no longer sparse and their individual dimensionalities have shrunk. Because these models must guarantee that every text has some chance of being created by any of the subjects, we find that the probabilities of all topics are non-zero. This implies that there are connections between pretty much every subject and every paragraph. However, in practical contexts, a book may only be connected to a few of themes and may have no connections to others at all. When dealing with Brief texts, the limits of relying only on topic distribution become readily apparent.

To overcome these constraints, we present a topic model based method that takes into account both lexical and semantic aspects in order to classify brief texts. In order to learn subjects in relation to all target categories, we use a background knowledge library, similar to other current approaches. Once we have all of the subjects from the repository, we utilize Gibbs sampling to associate each word in the short texts with the appropriate learning themes. In other words, we would assign each occurrence of a word to a subject, and then use these topics to describe a brief paragraph. This allows us to see how some, but not all, of the words in a brief text may be mapped to subjects. We also use various mapping weights based on the discriminatory power of words.

We consider that the subject to which a set of words is allocated has a stronger connection to the target category if those words are consistent with that set. Therefore, we introduce the lexical evidence-based expected cross entropy approach for gauging the discriminative power of words in compact texts. Every now and again, we take a step back and assess the results and impact of our Tested the suggested method using the Google Snippet and Consumed datasets, utilizing Wikipedia as a reference. Our strategy outperforms conventional approaches, as shown by the experiments. Here is how the rest of the paper is structured. The context and related works are presented in Section 2. Our methodology is laid forth in Section 3. Section 4 demonstrates experiments and analysis of results on two real-world datasets. Section 5 contains the debate, while Section 6 provides some last thoughts.

2. Related Work

Over fitting is a common issue in text classification, and the large dimensionality of feature space is one of the key obstacles to this task. Over the last several years, many different feature selection methods have been proposed in an effort to lower dimensionality. TF-IDF, IG, MI, ECE, et cetera are all measures of informational gain and frequency [11]. In turn, these qualities are used to symbolize documents. To predict the category labels of new, unseen documents, we may apply a classification model (K-Nearest Neighbour, Naive Bays, or Support Vector Machine) to the training set and generate a classifier. The term "lexical-based classification" is used to describe this sort of categorization strategy. Once topic models gain traction in the semantic analysis community, a new field of semantic-based text categorization emerges. Using topic distribution settings for each document, [8] and [9] decreased the dimensionality of the feature space of a document to the number of topics, which was then paired with a conventional classifier to accomplish classification.

In [12], a classified label representing a subject was assigned to each document. In order to combine labelled and unlabeled data into a single probabilistic model, [13] introduced a new cross-domain text classification approach that builds upon the original PLSA algorithm. Topics and labels were mapped one-to-one in [14], making the method applicable to multi-label categorization. While the aforementioned lexical-based and semantic-based classification methods work well enough for lengthy texts, the emergence of new forms of short texts in recent years has created new challenges for categorizing them. By comparing the online search results of the candidate terms with the content of the blog, [15] suggested a technique for extracting significant subject phrases from a blog, therefore determining whether the site provides rich material. By using the L2 normalization of the centred of each short text, [1] provided a contextual vector to represent each text. Whole sets of results from a search engine. To aid in the comprehension of brief and badly written documents, the authors of [16] used TAGME, a strong tool for identifying significant terms for tagging such texts. Words' lexical weight and the connections of subjects that they belonged to were taken into account in [17]'s suggested themes based similarity assessment approach for choosing feature words. Short texts were evaluated in [10] with the assumption that they all pertain to the same subject.

3. Proposed Approach

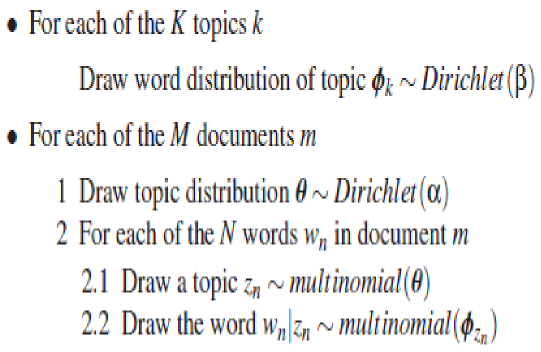
In the following, we will describe our method in depth. The following is the major step in our methodology. Pick a reputable external repository, and pull out some larger texts that pertain to the goal categories to use as context. Use a topic model to extract relevant information from these larger texts.

Third, using our refined predicted cross entropy, choose feature words that can be used to differentiate between alternatives. As the vector representations of brief texts, map the weighted words to appropriate subjects.

Five, use labelled data to train a classification model.

Category Topic Learning

Our method uses a knowledge base to discover information relevant to the desired subject areas it is crucial to choose a repository with sufficient material to exhaustively cover categories and their associated subjects. In order to learn themes, we first gather linked lengthy texts and then use a topic model. Inferred from a database of common sense information using a generative probabilistic model called Latent Dirichlet Allocation (LDA) [8], one may extract the semantic themes present in a corpus and use them to create an understanding of the data. The central concept is to model texts as multinomial distributions over latent themes, with each topic itself being described by a multinomial distribution over words. In order to generate LDA, the following steps are taken.

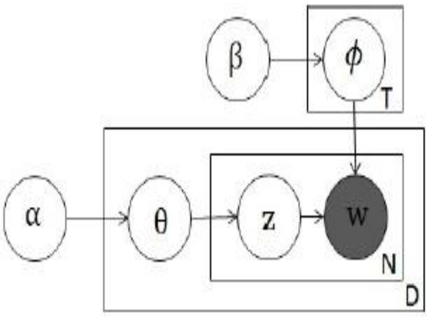


```
• For each of the $K$ topics $k$
        Draw word distribution of topic $\phi_k \sim Dirichlet(\beta)$
• For each of the $M$ documents $m$
    1 Draw topic distribution $\theta \sim Dirichlet(\alpha)$
    2 For each of the $N$ words $w_n$ in document $m$
        2.1 Draw a topic $z_n \sim multinomial(\theta)$
        2.2 Draw the word $w_n|z_n \sim multinomial(\phi_{z_n})$
```

The parameters of the distribution of documents and topics, respectively, are determined by the hyper-parameters and the topic-words parameter in the generating process. LDA's graphical representation is seen in Fig. 1. The distribution of documents' themes and the distribution of topics' words are both obtained by a Gibbs sampling technique. Within this framework, after all the words in short texts have been mapped to subjects using our method, we are interested in kit (the likelihood word t is given to topic k). Following instruction on necessary prerequisite knowledge, kit might be learned:

$$\phi_{kt} = \frac{n_{kt}+\beta}{n_k+V\beta} \qquad (1)$$

Where V is the total number of words in the vocabulary not is the number of times word t appears in topic k, and no is the total number of words in topic k. According to the description of the creative process, is the hyper-parameter.



Feature Selection

Word frequency and the association between words and categories are both taken into account by the feature selection metric known as expected cross entropy (ECE). A higher ECE value indicates that the associated word is more important in determining the category to which it belongs. The standard formula for determining the ECE of the word w is:

$$f(w) = p(w)\sum_{i} p(C_i|w)log\frac{p(C_i|w)}{p(C_i)} \qquad (2)$$

Where w stands for the word and Chi stands for the category it falls under. Here, we provide a two-stage enhancement to this feature selection metric. To begin, it has been observed that a representative term from category A may not play a significant role in category B. We choose alternative (B), where each category has its own weight for each word, as opposed to the second option (Equation 2), where each word has the same weight across the board. The following formula might be used to determine how much emphasis is placed on certain words:

$$f(w,C_i) = p(w|C_i)p(C_i|w)log\frac{p(C_i|w)}{p(C_i)} \qquad (3)$$

According to Equation (3), a term is more likely to have a high weight with respect to category I if it has a strong association with category I or if category I am of small size. Secondly, we anticipate that most unique terms will belong to a single group. In harmony with the whole. The M-ECE value of a given word is calculated using Equation (4), where M stands for the category we have chosen.

$$F(w,C_i) = f(w,C_i) - \sum_{j\neq i} f(w,C_j) \qquad (4)$$

Selecting the most distinctive N words from each group allows lexical characteristics to be represented. When we combine these feature words with semantic characteristics in the next step of our process, we provide various weights to the mappings for each word.

Words Mapping with Weight

In this work, we show how to extract meaning from text and map features onto a finite set of subjects. Then, we demonstrate how to describe a brief text using a combination of lexical and semantic characteristics while keeping the feature space's dimensions the same. We begin by relating the terms in these brief passages to previously taught themes. The Gibbs method of sampling is used. We utilize Formula (5) to repeatedly label every word in each text with a category.

$$p(z_i = k|z_{-i}, w_{-i}, \bullet) \propto \frac{n_{mk}^{-i}+\alpha}{n_m+K\alpha}\cdot\phi_{kt} \qquad (5)$$

$$\eta \propto F(w,C_i), \eta > 1 \qquad (6)$$

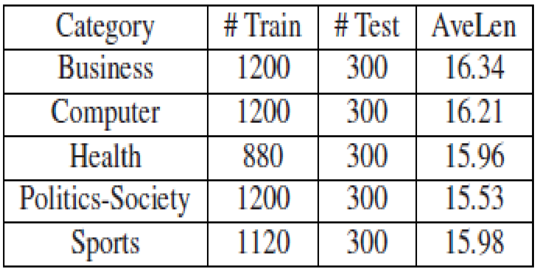
Where F (w, Ci) represents the M-ECE value of the word w for the given category Ci. In other words, the greater the significance of the term inside the category, the greater the mapping weight it will get. If a brief text is represented using themes, then the related subject will be highlighted. Short texts may still be represented by all these taught themes, despite the fact that the members of the vector are different when compared with only considering semantics.

4. Experiment and Analysis

Data Set

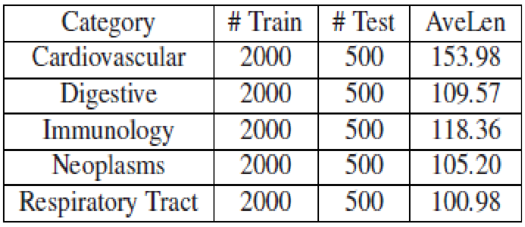
To test the efficacy of our method, we experiment on two different datasets. The Google Snippet 1 Dataset includes 8 distinct types of search engine results. To conduct our experiments, we selected 5 groups from the original dataset, as indicated in Table 1. The medical abstracts in Consumed 2 come from The 23 subheadings that make up 1991's Mesh (Medical Subject Headings) index. To supplement the primary dataset, we choose five categories from the original dataset and extract information from a subset of the abstracts. Table 2 displays some data about Consumed.

Table 1: Google Snippet Dataset



Category	# Train	# Test	AveLen
Business	1200	300	16.34
Computer	1200	300	16.21
Health	880	300	15.96
Politics-Society	1200	300	15.53
Sports	1120	300	15.98

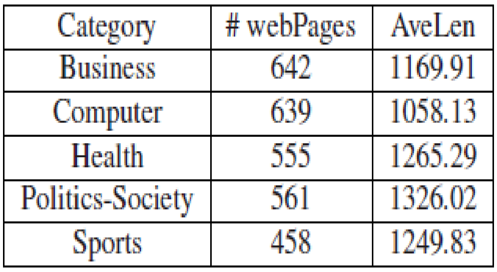
Table 2: Consumed Dataset



Category	# Train	# Test	AveLen
Cardiovascular	2000	500	153.98
Digestive	2000	500	109.57
Immunology	2000	500	118.36
Neoplasms	2000	500	105.20
Respiratory Tract	2000	500	100.98

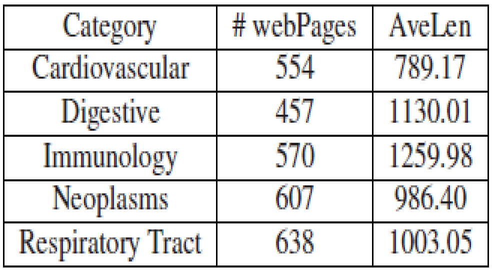
As we can see, the average length (Avenel) of texts in Google Snippet dataset is only ca. 16 after pre-processing. Although abstracts of Consumed dataset are much longer, they still contain less word co-occurrence.

Table 3: Background Dataset for Google Snippet



Category	# webPages	AveLen
Business	642	1169.91
Computer	639	1058.13
Health	555	1265.29
Politics-Society	561	1326.02
Sports	458	1249.83

Table 4: Background Dataset for Consumed

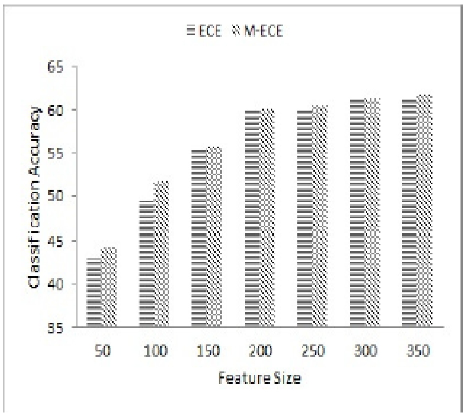


Category	# webPages	AveLen
Cardiovascular	554	789.17
Digestive	457	1130.01
Immunology	570	1259.98
Neoplasms	607	986.40
Respiratory Tract	638	1003.05

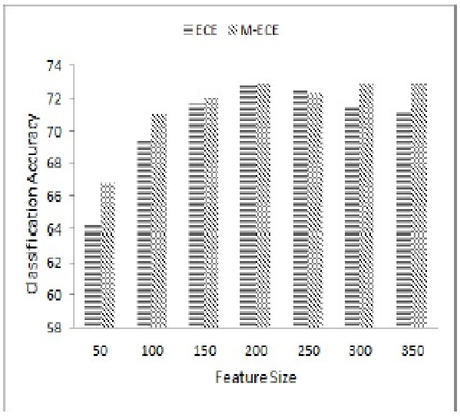
Results and Analysis

Evaluation of M-ECE

A series of tests were conducted to prove the use of our modified version of the standard ECE measure for feature selection. Both the Google Snippet and Consumed datasets were used in the analysis. To compare the effectiveness of standard ECE and our modified measure M-ECE, we applied both measures to feature sets of varying sizes, from 50 to 350. Classifying texts using a support vector machine (SVM). Figure 2 displays the results of the categorization.



(a) Google Snippet



(b) Consumed

Fig. 2: Classification Accuracy of Traditional ECE and M-ECE

As we can see in Fig. 2(a), M-ECE outperforms classic ECE in virtually all circumstances, with the exception of feature size 300, when they both perform similarly. In particular, this advantage becomes more apparent at low feature sizes, while the performance of classic ECE and M-ECE is only approximate at low feature sizes. Becomes bigger. We also found that accuracy remains almost constant when feature count increases above 200, suggesting that this dataset's lexical categorization benefits most from a feature size of 200. On the Consumed dataset (shown in Fig. 2(b)), M-ECE also outperforms classic ECE in most situations. The exception is when the feature size is 250. Furthermore, if we use conventional ECE as a selection metric, accuracy starts to drop down at feature size 200. On the other hand, M-ECE maintains its accuracy even as feature sizes increase. Thus, we infer that M-ECE is a more efficient and reliable method than ECE.

Impacts of Number of Topics

Here, we show how the categorization accuracy of our method would change as the number of subjects changed. The baseline dataset was subjected to LDA many times, with the number of topics varied between 40 and 160. We built feature spaces of varying dimensions while leaving the mapping process alone. According to the categorization we can observe that Google Snippet is quite accurate, with a maximum of 93.87% at subject number 60 and a minimum of 92.73% at topic number 160. As the number of topics in a Consumed dataset grows or shrinks, the accuracy varies somewhat. However, it remains almost constant when the number of topics exceeds 100. Because of this, we may say that our method's accuracy is relatively invariant over a wide range of subject matter.

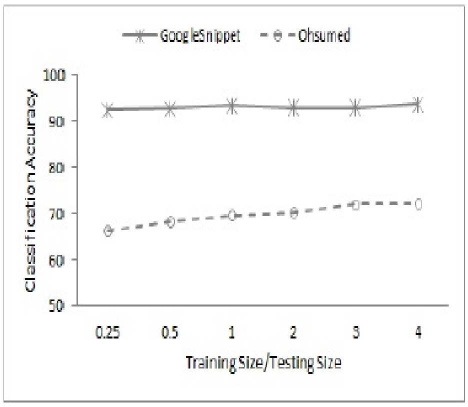


Fig. 4: The Effect of Different Training Sizes

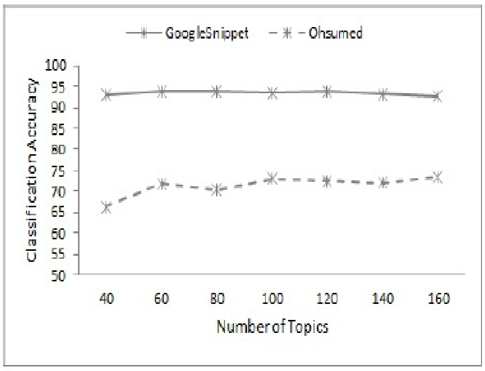


Fig. 5: The Effect of Topic Numbers

6. Conclusion

In this research, we provide a new measure technique to pick lexical characteristics from short texts, and we also describe a unique way to combining lexical and semantic features for short text classification. Results from experiments show that both feature selection and classification for short texts may be improved. The next step in our research will be to use our suggested method for text analysis and mining in different contexts. Moreover, we're curious about using correlated topic models as an extension of the basic latent dirichlet allocation model (LDA) to mine small texts for not only certain semantic elements, but also the correlations between these variables.

References

[1] Mehran Sahami , Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. Proceedings of the 15th international conference on World Wide Web, 2006.

[2] D Bollegala, Y Matsuo, M Ishizuka. Measuring semantic similarity between words using web search engines. Proceedings of the 16th international conference on World Wide Web, 2007.

[3] W. Yih and C. Meek. Improving similarity measures for short segments of text. Proceedings of the 22nd National Conference on Artificial Intelligence, 2007.

[4] Ou Jin , Nathan N. Liu , Kai Zhao , Yong Yu , Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. Proceedings of the 20th ACM international conference on Information and knowledge management, 2011.

[5] Xuan-Hieu Phan , Le-Minh Nguyen , Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Proceeding of the 17th

international conference on World Wide Web, 2008.

[6] Mengen Chen , Xiaoming Jin , Dou Shen. Short text classification improved by learning multi-granularity topics. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, p.1776-1781, 2011.

[7] Somnath Banerjee , Krishnan Ramanathan , Ajay Gupta. Feature Selection for Unbalanced Class Distribution and Naive Bayes. Proceedings of the Sixteenth International Conference on Machine Learning, p.258-267,1999.

[8] D. Blei , A. Ng , M. Jordan and J. Lafferty. Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, p.993-1022,2003.

[9] Yue Lu , Qiaozhu Mei , Chengxiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, Information Retrieval, v.14 n.2, p.178-203, 2011.

[10] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Volume 1, p. 536C544. 2012.

[11] Dunja Mladenic , Marko Grobelnik. Feature Selection for Unbalanced Class Distribution and Naive Bayes. Proceedings of the Sixteenth International Conference on Machine Learning, p.258-267, 1999.

[12] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In NIPS, volume 22, 2008.