International Journal of Engineering Research & Management Technology

ISSN: 2348-4039

January-2019 Volume 6, Issue-1

Email: editor@ijermt.org www.ijermt.org

INVESTIGATES AN ENSEMBLE APPROACH OF CLASSIFIER BASED TEXT MINING FOR DATA MINING APPLICATIONS

Gyanendra Kumar Shukla

Asst. Prof. Comm-IT Career Academy

(GGSIP University Delhi)

Mr. Deepak Rathore

Research Developers (Data Analytics)

ABSTRACT

In order to access information that is concealed inside databases, a number of data mining strategies have been developed. The mining of data and extraction of information become more difficult when the data to be mined is big, scattered, and diverse. In the field of data mining, classification is a job that sees significant use for the purpose of prediction. A huge variety of different approaches to machine learning have been developed specifically for this objective. The performance of individual classification algorithms may be improved by the use of ensemble learning, which combines many different basic classifiers. In particular, distributed data mining makes extensive use of ensemble learning as an important component. Therefore, research into ensemble learning is essential if one want to use it to solve data mining issues that are relevant in the real world. We present a method for constructing an ensemble of classifiers and analyse its performance using well-known learning methods on a variety of publically accessible datasets from the biomedical domain. These datasets are from the field of medical research. The automated division of text into a set of predetermined categories has been recognised for as long as anyone can remember as an important step in the management and processing of large quantities of digital documents, which are becoming more prevalent in today's society. This kind of information may be found on the web and is often referred to as digital or electronic information. It can be found in the form of papers, conference material, publications, journals, editorials, web pages, e-mail, and other forms. People no longer rely just on antiquated paper sources like books, periodicals, newspapers, and the like to get their knowledge; rather, they obtain it mostly via the many internet sources. The fact that this massive amount of information is not well organised, which makes it difficult to handle, is nonetheless the primary issue. Text categorization is widely acknowledged to be one of the most important strategies used for the purpose of arranging this sort of digital data. In this article, we have conducted research on the work that has already been done in the subject of text classification. This will enable us to have an accurate assessment of the advancements that have been achieved in this area up to this point. We have read the papers to the best of our ability and have made an effort to condense and synthesise all of the material that is currently available in an organised and comprehensible manner.

Keywords: *Ensemble Learning; Meta-Learning; Classifier Ensemble; Ensemble Method; Classification Performance; Meta-Classifier.*

INTRODUCTION

The task of assigning a document to one of many predetermined categories is known as text categorization. Text classification will give a document the category cj it belongs to if, to put it more properly, di is a document that is part of the full collection of documents D and c1,c2,c3,...,cn is the set of all the categories (Ikonomakis et al., 2005). Documents may be designated as belonging to one class, many classes, or neither class at all, based on the features they possess. The term "single-label" is used to refer to a document that has been assigned to only one class, while the term "multi-label" is used to refer to a document that has been allocated to more than one class (Wang & Chiang, 2011). If only one of the two classes is assigned to the document, a "single-label" text classification problem is considered a "binary class" problem. On the other hand, a "single-label" text classification problem is considered a "multi-class" problem if only N classes that are mutually exclusive are assigned to the document. Text classification involves representing the document, choosing which features to use or transforming those features, using a data mining technique, and, as a last step, evaluating the effectiveness of the data mining algorithm that was used. These days, the quantity of knowledge that can be found on the internet is staggering, and it is continuing to grow at an exponential pace. Since the advent of digital documents, automatic text categorization has been a key application and research area in order to effectively handle the vast amounts of data that are now accessible over the internet (Ikonomakis et al., 2005). It is based on methods of machine learning that automatically create a classifier by learning the features of the categories from a group of documents that have already been pre-classified (Sebastiani, 2002). The process of information extraction and summarization, as well as text retrieval and question answering, are all made much easier by its use. In most cases, the majority of the data that is used for categorization is of a heterogeneous character. This data may be obtained via the web, through newsgroups and bulletin boards, as well as through broadcast or printed news items, news reports, movie reviews, and ads. Due to the fact that they are compiled from a variety of sources, any document within the same category will often include a distinctive structure, its own unique vocabulary, and frequently a considerably unique writing style. Because of this, the automated categorization of text is very important. This article presents a comprehensive analysis of the work that has been done up to this point in the field of text classification. The research focuses on the difficulties that arise while attempting to categorise unstructured online content.

TEXT MINING METHODS AND TECHNIQUES

This section displays the chosen articles organised into groups according to the various text mining approaches and procedures that were used, as shown in Table 3. Text classification (which made up 32% of the total) and natural language processing (which made up 31% of the total) were the two that stood out the most, followed by information retrieval, text clustering, and text summarization. In addition, throughout our search, we came across theoretical publications. These studies just provide thoughts about the potential uses of text mining to educational settings. In conclusion, additional procedures and techniques that are used less often include information extraction, machine translation, and text synthesis, among many others.

APPLICATION OF A DATA MINING ALGORITHM

After the features have been selected and transformed, the documents are able to be represented in a manner that is straightforwardly compatible with a data mining technique. Either statistical methods, collectively referred to as the statistical method, or machine learning methods, collectively referred to as various supervised and unsupervised techniques of machine learning, can serve as the foundation for a data mining method. The statistical method is the more common of the two. There are several different text classifiers that make use of machine learning strategies such as decision trees (DT), naive-bayes (NB), rule induction, neural networks (NN), K-nearest neighbours (KNN), and support vector machines (SVM) (SVM). Both their architecture and the strategy that they take are unique to them. Table 3 provides a summary of some of the most well-known data mining approaches.

WIDELY USED DATA MINING METHODS

(RQ3) SVM, KNN, NB, ANN, Rocchio algorithm, and Association rule mining were found to be the most popular data mining techniques, according to a comprehensive literature review conducted in the field of text classification. Other popular approaches were association rule mining and Rocchio algorithm (ARM). Table 5 displays various machine learning algorithms and approaches, together with the number of published publications that make use of these methodologies. The chart makes it very evident that the SVM is the tool that is used by researchers in their work the most often. Quite a few of the writers have worked on the SVM method and suggested an enhanced version of it in order to further boost the application of this approach, which ultimately results in improved text classification performance. KNN algorithm is the second prominent approach utilised by the researchers as it is used in 31 articles. In a manner analogous to that of SVM, the authors of these papers have proposed various variants of KNN. They have also compared the performance of their proposed KNN algorithm to that of other machine learning algorithms in order to demonstrate that the new KNN algorithm performs better than traditional algorithms. Last but not least, we have the NB algorithm, which has been cited in 23 articles and is thus ranked third. It is also easy to observe that 65% of the papers use the KNN algorithm and the SVM algorithm to classify the text, while only 35% of the articles use other approaches outside the KNN and SVM algorithms. This is something that can be seen very clearly. Figure 5 is an illustration of this pattern. This unequivocally demonstrates that the SVM and KNN algorithms are among the most often used machine learning algorithms that are utilised by the academics. The SVM and KNN algorithms were used in 88 of the total 132 publications, while the other data mining approaches were utilised in 44 of the studies.

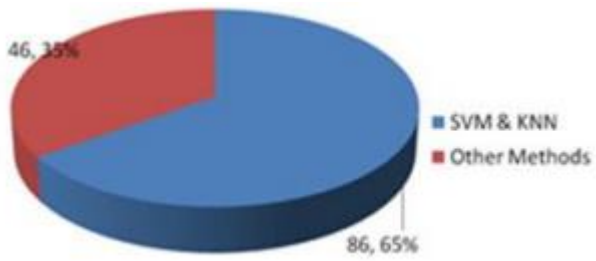

Figure 5. Distribution of machine learning methods (Number of papers/ percentage of total papers)

Table 5. Most important data mining methods used

Rank	Data Mining Methods	# Papers
1	SVM	55
2	KNN	31
3	NB	23
4	ANN	10
5	Rocchio Algorithm	9
6	Association Rule Mining	4

Methods of data mining in its many forms (RQ4) Figure 6 illustrates the distribution of the data mining strategies that have been used in the various articles. The approaches have been classified into the following categories: statistical methods, machine learning based methods, statistical methods and machine learning based methods, and the category of "unknown." When statistical techniques and machine learning based methods are utilised in the same model, the approach of that publication is referred to as "statistical methods Plus machine learning based methods." Some of the authors (Tao et al., 2005; Altinacy & Erenel, 2010; Luo et al., 2011) have not used any of the data mining methods in their paper because they have proposed a new term-weighting method and have compared its performance with the performance of the existing term-weighting schemes. Consequently, they have not used any of the data mining methods. The methodology of that particular study is referred to as "unknown procedures." It has been observed that 86 percent of papers used machine learning based methods, whereas only 6 percent of papers used statistical methods such as the Hidden Markov Model (Frasconi et al., 2002), Logistic Regression (Zhang and Oles 2001; Yen et al. 2011), Partial Least Squares (Namburu et al., 2005), and Linear Least Square Fit (Yang &Pedersen, 1997; Yang, 1999; Zhang &Oles, 2001). It is encouraging to note that more academics are studying the potential of machine learning approaches to forecast text categorization modules. This is because statistical methods are seen as black-box solutions, and these models are very reliant on data. According to the data shown in figure 6, out of a total of 132 studies, 8 papers used statistical techniques, 113 papers utilised machine learning methodologies, and 8 papers utilised both statistical and machine learning methodologies.

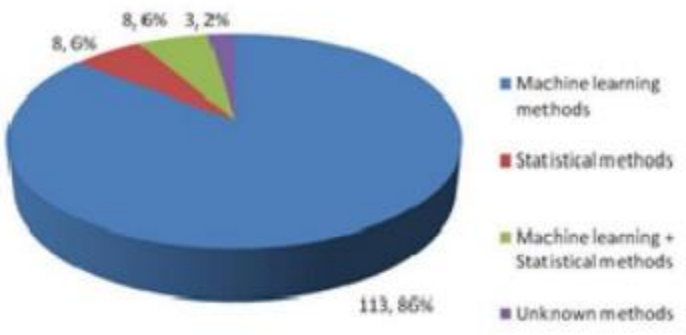

Figure 6. Distribution of data mining methods (Number of papers/ percentage of total papers)

Objectives of the study

1) To study of the data mining algorithm.
2) To study of the data mining methods.

EVALUATION

TM has been used extensively in a variety of situations for the purpose of evaluating the performance of students, particularly for the purpose of evaluating essays and online projects (Dikli, 2006; Kadupitiya, Ranathunga, & Dias, 2017). Several pieces of literature advocated for the evaluation of essays based on superficial characteristics, such as word counts (Dikli, 2006; Rudner, Garcia, & Welch, 2006). However, it is essential to move beyond this analysis (Crossley et al., 2015; Ericsson & Haswell, 2006), and there are a large number of publications that are centred on the application of semantic methods (Hughes, Hastings, Magliano, Goldman, & Lawless, 2012; Simsek et al., 2015), writing style (Oberreuter & Velásquez, 2013; Snow, Allen, Jacovina (2017). Approaches lexical and semantic are used in a manner similar to that previously described in the assessment of online assignments (Cutrone & Chang, 2010; Prevost, Haudek, Urban-Lurain, & Merrill, 2012; Ramachandran & Gehringer, 2011). However, in this particular instance, the works tend to be more focused on solving specific problems such as plagiarism (Adeva, Carroll, & Calvo, 2006), analysing short answer (Saha, Dhamecha, Marvaniya, Sindhgatta, & Sengupta, 2018), and classifying the questions. [Citations from: Adeva, Carroll, & Calvo, 2006; Saha, Dhamecha, Marvaniya, Sindhgatta (Godea, TulleyPatton, Barbee, & Nielsen, 2018). Finally, it could be used in formative evaluation to assist educators in establishing a pedagogical basis for decisions in order to maintain the environment (Gibson et al., 2017; Lehman, Mills, D'Mello, & Graesser, 2012) and evaluate interactions on educational online discussions. (Gibson et al., 2017; Lehman, Mills, D'Mello, & Graesser, 2012). (Rubio & Villalon, 2016; Yoo & Kim, 2014).

Goal	Number of papers (%)
Evaluation	95 (27.69%)
Student support/motivation	48 (13.99%)
Analytics	45 (13.11%)
Question/content generation	22 (06.41%)
User feedback	18 (05.24%)
Recommendation systems	9 (02.62%)
Others	106 (30.90%)

EVALUATION OF A TEXT CLASSIFIER

The effectiveness of a text classifier may be evaluated with the use of an assessment measure. Ckwe are able to construct a confusion matrix for each category, which is depicted in Figure 2. In this figure, a represents the number of true positive classifications, b represents the number of false positive classifications, c represents

the number of false negative classifications, and d represents the number of true negative classifications. If the classifier was flawless, both b and c would have a score of zero.

		Predicted Class	
		C_k	Not C_k
Actual Class	C_k	a	c
	Not C_k	b	d

Figure 2: Confusion matrix for Category Ck

The prediction accuracy may be calculated using the value (a+d)/(a+b+c+d). Recall and accuracy are often used as the performance metrics for evaluating text categorization systems. The percentage of documents in category Ck that are properly predicted is what is meant by the term "recall," which is defined as a/(a + c). The ratio of the number of documents that are predicted as being in category Ck to the actual number of documents that are in that category is known as the precision of the classification system. Each degree of recall has a corresponding amount of accuracy that goes along with it. In general, increasing the recall will result in a decrease in accuracy, and vice versa (Yang &Pedersen, 1997). The break-even point, sometimes referred to as the BEP, is the point at which recall and accuracy are equal. This point is frequently used as a single summary statistic for comparing findings. There are certain circumstances in which a genuine BEP is not present. It is usual practise to combine Recall and Precision into a single performance measure known as the F1 Score. This score is defined by the formula F1 = 2PrecisionRecall/(Precision+Recall), and it is derived from the combination of the two performance measures. This is the product of accuracy and recall divided by their average, and it serves as yet another helpful statistic that is utilised for analysing the efficacy of classifiers. After computing these scores for the binary judgements on each particular category, the results are then averaged across all of the categories. When dealing with many classes, there are two different approaches to average these measurements, namely the macro-average and the micro-average. Both of these averages have their advantages and disadvantages (Antonie & Zaiane, 2002). When doing the macro-averaging, one confusion matrix is utilised for each class. The performance measures are calculated on each of the confusion matrices, and then the results are averaged. In micro-averaging, there is just one contingency table that is utilised for all of the classes. Within each cell, an average of all of the classes is calculated, and the performance measures are derived from those cells. The macro-average metric assigns equal importance to each of the classes, irrespective of the total number of documents that fall within each category. The micro-average metric assigns equal weight to each document, giving more importance to overall performance on common classes.

DOCUMENTS, SOCIAL NETWORKS, BLOGS, AND EMAILS

Emails, social networks, and blogs are some examples of additional resources that may give information on the users' interactions inside educational settings. Text mining applications that make use of these resources are often tied in some way to text categorization, with a primary emphasis placed on sentiment analysis. In this

context, sentiment analysis has been applied in order to solve various problems, such as the following: to extract opinion about the educational environment (Kechaou, Ammar, & Alimi, 2011); to assist the instructor in improving the educational environment increasing the student engagement and creating an adaptive educational environment (Ortigosa, Martn, & Carro, 2014); and to aid in teaching evaluation and provide feedback based on the user interaction on social networks. In addition, sentiment analysis has been applied in order to extract opinion (Leong, Lee, & Mak, 2012).

Traditional statistical features such as TFIDF, Information Gain, Mutual Information, and CHI statistics can be combined with natural language processing features in order to extract sentiment from online educational platforms. Other features, such as TFIDF, Information Gain, Mutual Information, and CHI statistics, can also be used (Kechaou et al., 2011; Truong, 2016). In order to extract these characteristics successfully, it is necessary to first carry out a preprocessing phase (Leong et al., 2012). Text categorization is also used to extract information about students' behaviours, in addition to sentiment analysis (Tobarra et al., 2014). It provides assistance to educators in enhancing student input and preventing the loss of students. To extract user behaviour, Mansur and Yusof (2013) suggested the deployment of a hybrid approach that would be based on a mix of logs from educational settings and social network analysis. This approach would be implemented. In the similar vein, Tobarra et al. (2014) blends social network and forums interaction to develop students' models. Email analysis is another another application of text mining that may be used (Aghaee, 2015). Finally, it is also conceivable to automatically forecast each student's mark based on his or her participation in the online community that constitutes the class blogosphere (Gunnarsson & Alterman, 2012).

CRITERION

In order to properly evaluate each possible main research, systematic reviews need to have well defined inclusion and exclusion criteria. The inclusion and exclusion criteria for primary studies are determined by the research questions, and these criteria guide the selection of primary studies (Catal, 2011). The publications were taken into consideration for inclusion in our review if they described research on text categorization. This evaluation does not provide practitioners with a comprehensive description of all text categorization models or the methods that were used in the development of those models. Our objective is to categorise the articles according to the years they were published, the datasets they used, the various feature selection methods, the data mining algorithms, and an assessment metric. We included the articles that were published in reputable digital journals and conference proceedings such as ACM, IEEE, Springer, and Elsevier. These journals and portals are listed in the previous sentence. The publications that did not provide the results of any experiments were disregarded by our team. We did not omit the publications in which a new feature selection approach or some new evaluation measure was presented rather than a new data mining algorithm. In these papers, we did not offer a new data mining algorithm. To put it another way, we included all of the publications that were in some way connected to the subject of text categorization, whether directly or indirectly. The year of publication of the study or the procedures that have been employed were not taken into consideration prior to our exclusion.

CONCLUSION

The process of retrieving particular information from the internet is analogous to the classic search for a needle in a haystack. In the context of this study, the "needle" refers to a specific piece of knowledge that a user is looking for, while the "haystack" refers to the massive data warehouse that has been accumulated on the web over an extended period of time. Text categorization is quickly becoming one of the most significant approaches to solving this challenge. In this study, we analyse the progress that has been done in the field of text classification so far by conducting a survey of the publications on text classification that have been published between the years 1997 and 2012 in reputable conference proceedings and academic journals. This review would be helpful for future research that is based on the studies that have been done in the past. We have reviewed the articles with a particular emphasis on the different kinds of data mining methods, feature selection methods, the dataset, and the evaluation approaches that were employed by each research to carry out the outcomes.

REFERENCES

1) Altinçay, H., & Erenel, Z. (2010). Analytical evaluation of term weighting schemes for text categorization. Pattern Recognition Letters, 31, 1310–1323. Altinçay, H., & Erenel, Z. (2012).
2) Using the absolute difference of term occurrence probabilities in binary text categorization. ApplIntell, 36, 148–160. Amine, B.M., & Mimoun, M. (2007).

3) WordNet based cross-language text categorization. IEEE/ACS International Conference on Computer Systems and Applications, AICCSA. An, J., & Chen, Y.P.P. (2005).

4) Keyword Extraction for Text Categorization. Proceedings of the IEEE International Conference on Active Media Technology AMT. Antonie, M.L., & Zai'ane, O.R. (2002).

5) Text document categorization by term association. Proceedings of the IEEE International Conference on Data Mining, ICDM. Azam, N., & Yao, J.T. (2012).

6) Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Systems with Applications, 39, 4760–4768. Bakus, J., & Kamel, M.S. (2006).

7) Higher order feature selection for text classification. Knowledge Information System, 9(4), 468-491. Basu, A., Watters, C., & Shepherd, M. (2002).

8) Support vector machines for text categorization. Proceedings of the 36th Hawaii IEEE International Conference on System, HICSS'03. Cabrera, R.G., Gomez, M.M.Y., Rosso, P., & Pineda, L.V. (2009).

9) Using the Web as corpus for selftraining text categorization. Information Retrieval, 12, 400–415. Canfora, G., & Cerulo, L. (2005).

10) How software repositories can help in resolving a new change request. Workshop on Empirical Studies in Reverse Engineering. Catal, C. (2011).

11) Software fault prediction: A literature review and current trends. Expert Systems with Applications, 38, 4626-4636. Chang, Y.C., Chen, S.M., & Liau, C.J. (2008).

12) Multi-label text categorization based on a new linear classifier learning method and a category-sensitive refinement method. Expert Systems with Applications, 34, 1948–1953.