Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online) DOI: 10.7176/JIEA
Vol.9, No.3, 2019

Feature Selection Techniques and Classification Accuracy of Supervised Machine Learning in Text Mining

Makara, Loise * **Ogada, Kennedy (PHD)** **Njagi, Dennis (PHD)**

School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, P. O. Box 62000-00200 Nairobi, Kenya

Abstract:

Text mining is a special case of data mining which explore unstructured or semi-structured text documents, to establish valuable patterns and rules that indicate trends and significant features about specific topics. Text mining has been in pattern recognition, predictive studies, sentiment analysis and statistical theories in many areas of research, medicine, financial analysis, social life analysis, and business intelligence. Text mining uses concept of natural language processing and machine learning. Machine learning algorithms have been used and reported to give great results, but their performance of machine learning algorithms is affected by factors such as dataset domain, number of classes, length of the corpus, and feature selection techniques used. Redundant attribute affects the performance of the classification algorithm, but this can be reduced by using different feature selection techniques and dimensionality reduction techniques. Feature selection is a data preprocessing step that chooses a subset of input variable while eliminating features with little or no predictive information. Feature selection techniques are Information gain, Term Frequency, Term Frequency-Inverse document frequency, Mutual Information, and Chi-Square, which can use a filters, wrappers, or embedded approaches. To get the most value from machine learning, pairing the best algorithms with the right tools and processes is necessary. Little research has been done on the effect of feature selection techniques on classification accuracy for pairing of these algorithms with the best feature selection techniques for optimal results. In this research, a text classification experiment was conducted using incident management dataset, where incidents were classified into their resolver groups. Support vector machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB) and Decision tree (DT) machine learning algorithms were examined. Filtering approach was used on the feature selection techniques, with different ranking indices applied for optimal feature set and classification accuracy results analyzed. The classification accuracy results obtained using TF were, 88% for SVM, 70% for NB, 79% for Decision tree, and KNN had 55%, while Boolean registered 90%, 83%, 82% and 75%, for SVM, NB, DT, and KNN respectively. TF-IDF, had 91%, 83%, 76%, and 56% for SVM, NB, DT, and KNN respectively. The results showed that algorithm performance is affected by feature selection technique applied. SVM performed best, followed by DT, KNN and finally NB. In conclusion, presence of noisy data leads to poor learning performance and increases the computational time. The classifiers performed differently depending on the feature selection technique applied. For optimal results, the classifier that performed best together with the feature selection technique with the best feature subset should be applied for all types of data for accurate classification performance.

Keywords: Text Classification, Supervised Machine Learning, Feature Selection

DOI: 10.7176/JIEA/9-3-06
Publication date:May 31st 2019

\

I. INTRODUCTION

Text mining also referred to as text classification is an automated process of assigning textual documents to a set of predefined categories (Aggarwal & Zhai, 2012). It is a discovery and extraction of interesting, non-trivial knowledge from free or unstructured text (Srivastava & Sahami, 2009). Text mining is a special case of data mining, which explores data in text files to establish valuable patterns and rules that indicate trends and significant features about specific topics(Kogan & Berry, 2010). Text mining works for unstructured or semi-structured collection of text documents such as emails, social media, blogs, corporate documents, Web pages, newsgroup postings, as well as survey feedbacks (Liu, 2012).

Text mining has been used for pattern recognition, predictive studies, sentiment analysis and statistical theories in many areas of academia, research, medicine, financial analysis, social life analysis, business intelligence other fields (Aggarwal & Zhai, 2012). It is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics as well as computational linguistics and highly rely on the natural language processing and machine learning concepts.

Natural language processing (NLP) is the study of human language so that computers can understand natural languages as humans do. NLP focus on the automatic processing and analysis of unstructured textual information, which has remained one of the oldest and most challenging problems in the field of artificial intelligence (Kumar, 2011).

Machine learning is a sub field of data science that focuses on designing and creating algorithms and programs which learns on their own and make predictions on data. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output value within an acceptable range (Alpaydin, 2016).

Machine learning algorithms are often categorized as being supervised, unsupervised or semi-supervised (Alpaydin, 2016). Supervised algorithms use patterns to predict the values of the label on additional unlabeled data. They require a training set of labelled documents and return a function that maps documents to the pre-defined class labels (Kogan & Berry, 2010). Supervised algorithms require humans to provide both input and desired output, in addition to furnishing feedback about the accuracy of predictions during training. Once training is complete, the algorithm will apply what was learned to new data. Such algorithms include Support Vector Machines (SVM), Naïve Bayes (NB), decision tree (DT), K-Nearest Neighbor (K-NN), Linear Regression and Neural networks (Marsland, 2014).

Most of the widely used data driven supervised machine learning algorithms include SVM, NB, DT, and K-NN(Chezian & Kanakalakshmi, 2015). K-NN classifier is an instance and case-based learning algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance. It can be used for both classification and regression problems but is more widely used in classification problems in the industry. KNN is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function. This method is highly used for its effectiveness, non-parametric and easy way of implementation (Marsland, 2014).

Naïve Bayes method is a probabilistic classification technique based on Bayes' theorem with an assumption of independence between predictors. It is a kind of module classifier under known priori probability and class conditional probability. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Its basic idea is to calculate the probability that document D belongs to class C.

Decision tree is a supervised learning algorithm that is mostly used for classification problems. It works for both categorical and continuous dependent variables, where the population is split into two or more homogeneous sets. This is done based on most significant attributes to make as distinct groups as possible. It is a simplistic in understanding and interpreting even for non-expert users and for that it is one of the most widely used classification techniques. Its classification accuracy is also competitive with other learning methods, and it is very efficient (Marsland, 2014).

Support vector machines (SVM) is another type of learning classifier, which has many desirable qualities that make it one of the most popular algorithms. It not only has a solid theoretical foundation, but also performs classification more accurately than most other algorithms in many applications, especially those applications involving very high dimensional data. It has been shown by several researchers that SVM is perhaps the most accurate algorithm for text classification. In SVM, each plot data is plotted as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a coordinate. The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. SVM classifier method is outstanding from other with its effectiveness (Marsland, 2014).

Unsupervised algorithms do not need to be trained with desired outcome data. Instead, they use an iterative approach called deep learning to review data and arrive at conclusions (Kogan & Berry, 2010). Unsupervised learning algorithms are used for more complex processing tasks than supervised learning systems. Unsupervised methods usually start off from unlabeled data sets, so, in a way, they are directly related to finding out unknown properties in them (Radovanovic & Ivanović, 2008). Unsupervised learning is used against data that has no historical labels. The system is not told the "right answer." The algorithm must figure out what is being shown (Yao et al.,2016). Such algorithms include K Mean and deep learning (Marsland, 2014).

Semi-supervised learning is used for the same applications as supervised learning. But it uses both labeled and unlabeled data for training. Typically, a small amount of labeled data with a large amount of unlabeled data, probably because unlabeled data is less expensive and takes less effort to acquire. This type of learning can be used with methods such as classification, regression and prediction. Semi-supervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process(Chezian & Kanakalakshmi, 2015).

In the areas of application, the different machine learning techniques yield great but varying results. The performance of these machine learning algorithms varies depending on various factors, such as the domain of the

dataset, number of classes in each dataset, the length of the corpus, data preprocessing done, feature selection techniques applied, quantity of the training data, estimation methods applied among others (Chezian & Kanakalakshmi, 2015).

Data preprocessing removes the noise and normalizes the data. If the data is not normalized, classification algorithm may be biased towards a particular set of attributes, affecting its accuracy. For supervised classification techniques, lack of diverse training data affects the performance of a classifier (Kogan & Berry, 2010). Training data should consist instances with different kind of information, to help the classier learn more kind of patterns. The quantity of the training data is equally important, this is because low quantities of training data could lead to overfitting, which generally degrades the accuracy of the classifier. Wrong estimation methods affect the accuracy of the classifier. Classification accuracy ideally should not be measured in a single experiment. Cross validation is a very good way of measuring accuracy which uses a leave one out kind of procedure and averages the accuracy for all the iterations (Chezian & Kanakalakshmi, 2015).

Data always contains more information than is needed to build the model, holds the wrong kind of information, or contain redundant attribute (Gurusamy, 2015). These are additional factors that affects the performance of the classification algorithm. They can lead to overfitting, mislead the modeling algorithms while making it more difficult to discover meaningful patterns. To reduce such effects, different feature selection techniques and dimensionality reduction techniques should be used. Feature selection is a data preprocessing step that chooses a subset of input variable while eliminating features with little or no predictive information (Liu, 2012).

Feature selection follows different approaches including ranking, filter, wrapper and embedded (Divya & Kumar, 2015). With ranking, features are ranked by some criteria and then those above a defined threshold are selected. Filters methods is the simplest to implement. It evaluates the quality of selected features, independent from the classification algorithm, where the features are selected first, then this subset is used to execute a classification algorithm. These methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. Wrapper methods require application of a classifier- which should be trained on a given feature subset, to evaluate its quality. These methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy. Embedded methods perform feature selection during learning of optimal parameters, as part a classification algorithm. They learn which features best contribute to the accuracy of the model while the model is being created(Divya & Kumar, 2015).

According to Bolon-Canedo et al (2015), the decision on the features to be used to create a predictive model is usually difficult and may require deep knowledge of the problem domain (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2015). Gurusamy (2015) suggests that it should be made possible to automatically select those features in the data that are most useful and/or most relevant for the problem being worked on. Different feature selection techniques are available, which include Information gain, Term Frequency (TF), Term Frequency-Inverse document frequency(TF-IDF), Mutual Information (MI), and Chi-Square(x^2), Gain Ratio (GR), Symmetrical Uncertainty (SU), Relief-F (RF), One-R (OR), Gini-Index among others(Gurusamy, 2015).

Weights applied to the terms are crucial to the accuracy of the classification process. Some of the commonly used term weighting method of feature selection in text classification include TF, TF-IDF, and Boolean (Xia, Wang, Chen, & Zhai, 2016). Boolean weighting is the simplest way to weighting term of feature vectors by assigning them 0 or 1. It allows to whether or not a query term is present in a document. If the term exists in document, the value equals to 1, in other cases it equals to 0. According to this method, all terms have got uniform importance. Whereas, a term that is mentioned more often in the document should receive a higher weight value (Divya & Kumar, 2015).

Term frequency is another method that assigns to each term in a document a weight which depends on the number of occurrences of the term in the document. Generally, this approach provides more accuracy than the Boolean weighting, but is not adequate. Reason being that the contribution to the classification of a term is directly proportional to the number of encounters in a certain document (Divya & Kumar, 2015). Nevertheless, is inversely proportional to the prevalence in the whole space. Less common attributes are more distinctive than others, therefore the need for a further method- term frequency- inverse document frequency (tf-idf). The tf-idf is frequently used because value increases comparatively to the number of times a term appears in the document but is offset by the frequency of the term in the corpus (Divya & Kumar, 2015).

With some algorithms, feature selection techniques are built-in so that irrelevant columns are excluded, and the best features are automatically discovered (Divya & Kumar, 2015). Each algorithm has its own set of default techniques for intelligently applying feature reduction. However, you can also manually set parameters to influence feature selection behavior. Classification accuracy is a function of the number of features used, therefore, to obtain the highest possible level of accuracy, an optimal feature set is usually required. Since there is no best ranking index for different datasets, the only way to be sure that the highest accuracy is obtained in practical problems is testing a given classifier on a number of feature subsets, obtained from different ranking indices (Karabulut, Özelb, & İbrikçi, 2012).

To get the most value from machine learning, pairing the best algorithms with the right tools and processes is necessary (Masoumeh & Seeja, 2015). Little research has been done on pairing these machine learning algorithms with the feature selection techniques for optimal results (Gurusamy, 2015). This raises a necessity to investigate and analyze the effect of different feature selection techniques on individual supervised machine learning algorithms for best pairing in text classification (Masoumeh & Seeja, 2015). This would be of significant for researchers, practitioners and professionals who seek to create similar predictive models or to improve upon an existing model.

II. RELATED STUDIES

Researchers have studied the various aspects of feature selection. A study on Feature Selection Methods for Text Mining by Divya and Kumar (2015) demonstrate the significant of the Feature selection activity as one of the important and frequently used techniques for data preprocessing for data mining(Divya & Kumar, 2015).A different study on the influence of feature granularity for Chinese short-text classification in the Big Data era conducted by Wang & Deng (2017) demonstrated how text that are short and has redundant attributes affect the

classification accuracy during text classification. They further showed that different feature selection method such as use of low granularity language fragments significantly improve the classification performance (Wang & Deng, 2017).

In their study, Meesad and Li (2013) proposed a hybrid feature selection method to address the issue of sparse features when converting tweets to word vectors in tweets and the unreliability of using average sentiment scores to indicate sentiments. From the findings, the hybrid feature selection method improves the accuracy of the stock trend prediction for the SVM by 2.83%(Meesad & Li, 2014).

A Research on feature-based opinion mining using topic maps conducted by Xia, et al. (2015), highlighted on the complexity of natural language, such as domain dependence of sentiment words and extraction of implicit features. The experimental results revealed that the feature-based method can improve the accuracy of the Opinion mining by up to12% (Xia, Wang, Chen, & Zhai, 2016).

A comparative study conducted by Karabuluta et al (2012) on the effect of feature selection on classification accuracy, stated that Feature selection has become interest to many research areas in, since it helps the classifiers to be fast, cost-effective, and more accurate. They carried out an experiment on different classifiers, using different real datasets which were pre-processed with feature selection methods and up to 15.55% improvement in classification accuracy was observe(Karabulut, Özelb, & İbrikçi, 2012).

Elkhani and Muniyandi (2016) reviewed the effect of feature selection for microarray data on the classification accuracy for cancer data sets. The interest was to address the common difficulty for all techniques is the large number of genes compared to the small sample size which has a negative impact on their speed and accuracy. The study demonstrated that different feature selection techniques has different effects on the classification performance of the various machine learning algorithms. The results indicated that classification accuracy alone can be misleading, and other performance- measures, recall, precision and f-measures should be considered (Elkhani & Muniyandi, 2016).

III. METHODOLOGY

For this research, a text classification experiment was conducted using incident ticketing dataset. Incident ticketing is an ICT service operation process that is responsible of logging all support request raised by service users, categorize and, prioritize them before allocating them to the appropriate support technicians. The incident ticketing dataset was obtained from ServiceNow, which is an incident management system, implemented by the ICT shared services of company X. The dataset was obtained with authority from the data owner.

The incident details collected included the incident subject and the description, which contain the descriptive text about the incident. The experiment involved classifying the incidents into their resolver groups, which formed our classes. The five resolver groups, now the class labels included the administration, customer services, business applications, Infrastructure and Endpoint. 300 incidents were collected for consecutive days and were labeled with the class labels using expertise of the domain knowledge. An ARFF file was created, which is the appropriate model for the data mining environment selected. We used a 10-K cross validation methodology, which is a good method for measuring accuracy as it randomly partitions the original corpus into 10 partitions, and iteratively use 9 partitions as the training set to train the model, and one as the test set to evaluate it, then averages the accuracy for all the iterations.

Four supervised classification algorithms were examined, including support vector machine (SVM), Naïve Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN). The four were selected as they are the widely used data driven supervised machine algorithms.

For feature selection, three techniques, including term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF) and Boolean methods were used, as they are the term weighting methods that have been applied highly in text classification. Feature selection was employed using the filtering approach, where features were selected before the classifiers were trained.

To compare the performance of the classification algorithms with feature selection methods, WEKA data mining tool was used. The performance of the classification algorithms was then evaluated for classification accuracy, which is the number of correctly classified instances in the test set divided by the total number of instances in the test set expressed in percentage.

IV. EXPERIMENTAL RESULTS

The classification accuracy of the classifiers under the different feature selection methods are as shown in table 4.1

	TF	**Boolean**	**TF-IDF**
SVM	88	90	91
NB	70	83	83
DT	79	82	76
KNN	55	75	56

Table 4.1 Classification Accuracy

A graphical representation of the accuracy performance for the classifiers is as shown in figure 5.1

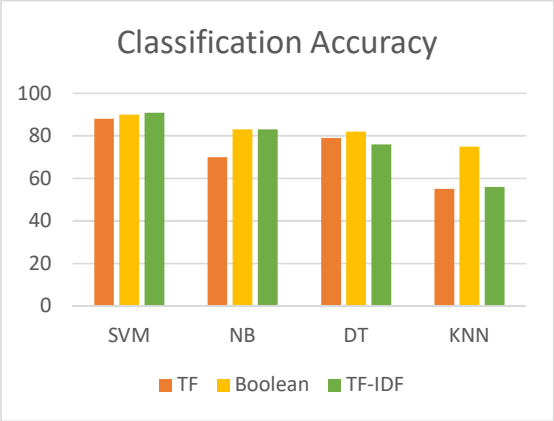

Fig 5.1 Classification Accuracy

V. CONCLUSION

Feature selection is a significant step of data preprocessing in text classification, because it may have a considerable effect on accuracy of the classifier. It reduces the number of dimensions of the dataset, so the processor and memory usage reduce; the data becomes more comprehensible and easier to study on. Presence of noisy data leads to poor learning performance and increases the computational time. Feature selection technique with the best feature subset should be applied for all types of data for accurate classification performance.

In this study, we have investigated the influence of feature selection on three widely used data driven supervised machine classifiers, including support vector machines, Naïve Bayes, decision tree and K Nearest neighbor, using the incident ticketing dataset. The results showed that all the classifiers performed differently depending of the feature selection technique used. In overall performance, SVM performed best, followed by DT, KNN and finally NB. For optimal results, its therefore essential that classification algorithms be paired with the features selection methods that has the best feature subset and minimal effect on classification accuracy.

VI. REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data.* Springer Science & Business Media.

Alpaydin, E. (2016). *Machine Learning: The New AI.* MIT Press.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). *Feature Selection for High-Dimensional Data.* New York: Springer.

Chezian, R. M., & Kanakalakshmi, C. (2015). Perfomance Evaluation for Machine Learning Techniques for Text classification. *Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications*, 53-57.

Divya , P., & Kumar, N. G. (2015). Study on Feature Selection Methods for Text Mining. *International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)* , 11-19.

Elkhani, N., & Muniyandi, R. C. (2016). Review of the effect of feature selection for microarray data on the classification accuracy for cancer data sets. *International Journal of Soft Computing*, 334-342.

Gurusamy, V. (2015). *Preprocessing Techniques for Text Mining.* Retrieved from Researchgate: https://www.researchgate.net/publication/273127322_Preprocessing_Techniques_for_Text_Mining

Karabulut, E. M., Özelb, S. A., & İbrikçi, T. (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, 323-327.

Kogan, J., & Berry, M. (2010). *Text Mining: Applications and Theory.* John Wiley & Sons.

Kumar, E. (2011). *Natural Language Processing.* I. K. International Pvt Ltd,.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.

Marsland, S. (2014). *Machine Learning: An Algorithmic Perspective, Second Edition.* CRC Press.

Masoumeh , Z., & Seeja, K. R. (2015). Feature Extraction or Feature Selection for Text Classification: Case Study on Phishing Email Detection. *I.J. Information Engineering and Electronic Business*, 60-65.

Meesad , P., & Li, J. (2014). Stock Trend Prediction Relying on Text Mining and Sentiment Analysis with Tweets . *Automated news reading*, 257-262.

Radovanovic, M., & Ivanović, M. (2008). *TEXT MINING: APPROACHES AND APPLICATIONS.*

Srivastava, A. N., & Sahami, M. (2009). *Text Mining: Classification, Clustering, and Applications.* CRC Press.

Wang, H., & Deng, S. (2017). Studies on the influence of feature granularity for Chinese short-textclassification in the Big Data era. *A paper-text perspective*, 689-708.

Xia, L., Wang, Z., Chen, C., & Zhai, S. (2016). Research on feature-based opinion mining using topic maps. *The Electronic Library*, 435-456.