Profile Categorization System Based on Feature Reduction

Olfa Mabrouk
olfa.mab@hotmail.fr
MARS Research Laboratory,
LR 17ES05
University of Sousse, Tunisia.

Lobna Hlaoua
lobna1511@yahoo.fr
MARS Research Laboratory,
LR 17ES05
University of Sousse, Tunisia.

Mohamed Nazih Omri
mohamednazih.omri@fsm.rnu.tn
MARS Research Laboratory,
LR 17ES05
University of Sousse, Tunisia.

Abstract

Though the enormous impact of social media on our daily life, we observe a lack of information about those who create the contents. In this regard, author profiling tries to determine the profile category of authors by analysing their published texts. This paper presents a profile categorization system to solve the multi-class categorization problem. The system consists of two modules: the processing module and the classifying module. In the first module TF-IDF with threshold filtering method are used to extract the relevant terms. While support vector machine (SVM) is used in the classifying module. Our proposed feature reduction method is tested using the RepLab 2014 Data set and evaluated using performance measures: precision, recall and F-measure. Result obtained show that proposed method improves system performance when solving a multi-class profile categorization problem.

1 Introduction

A microblog is a type of blog in which users can post small pieces of digital content like pictures, video or audio on the Internet. These posts, called micro posts, are immediately available to a small community or public. It differs from a blog due to its smaller content. Microblogging is highly popular among users due to its portability and immediacy. Twitter is a microblogging site which becomes popular daily. It is popular among many types of users from various countries. Thus, many peoples, organizations use Twitter to share their messages and opinions. Because of this reason, Twitter contains a large amount of hidden information. This information can be used for many purposes: opinions extraction sentiments analysis (Liu and Zhang 2012) user interest discovery (Sendi, Omri, and Abed 2017). Thus, by analyzing this messages, one can generate a large database which contains real time information. In this context twitter profile categorization is used to improve results of information retrieval by obtaining a more pertinent response from the influential profiles corresponding to the request domain. The authors or profile categorization is a classification of profile by type of their activity to identify the influential ones in each domain or category.

With millions of users, to perform any analysis or information retrieval, a need was felt to classify users, using machine learning algorithms. With the development of machine learning techniques, now-a-days, many researchers tend to use machine learning techniques how finds application in a wide variety of domain in text mining. News filtering and organization, document organization and retrieval (Chakrabarti & al. 1997), E-mail classification and spam filtering, opinion mining and sentiments analysis (Liu and Zhang 2012). A big challenge in text categorization is the learning from a big data set. This may lead to high computational burden for the learning process. On the other hand, some irrelevant and redundant feature may reduce the performance of text categorization system. To avoid this problem and to speed up the learning process, it is necessary to perform the feature reduction solution to reduce the features vectors size without reducing the system performance (Aggarwal and Zhai 2012). Support Vector Machine(SVM) (Vapnik 1998), developed byV.N.Vapnik, is an important pattern recognition technique based on structural risk minimization(SRM) (Vapnik 1999). It first maps the sample points in to a high-dimensional feature space and aims at seeking for an optimal separating hyperplane that maximizes the margin between two classes in this space, where the margin is defined as the sum of the distances of the hyperplane from the closest point of the two classes. Because of its remarkable characteristics such as global minima, good generalization performance and small size of training data, SVM has been successfully applied in many areas, such as machine fault diagnosis (Widodo and Yang 2007), image identification (Heisele, Ho, and Poggio 2001), text.
The rest of this paper is structured as follow. In section 2 we present the previous work on feature selection techniques and Machine Learning. In section 3 we introduce proposed model and algorithm. Experimental results are given in section 4 along with performance analysis. A conclusion and future work discussion are given in section 5.

2 Literature review

The task of categorizing author profiles has an emerging interest in the scientific community, as can be seen in the number of related tasks around the topic arisen last years such as the shared task on Native Language Identification at BEA-8 Workshop at NAACL-HT 2013, the task on Computational Personality Recognition (WCPR) at ICWSM 2013 and at ACM Multimedia 2014, and the task on Author Profiling (author profile categorization)

at PAN 2013 and PAN 2014. For the task on author profile categorization at PAN 2013 (Rangel & al. 2013), most of the participants used combinations of style-based features(punctuation marks, capital letters, quotations,..) and content-based features(LSI,BOW,TF-IDF,..). The best result using content-based features is obtained with 35% of precision.

The common feature reduction approach for text categorization is the feature selection. In the last decades, a number of feature selection methods have been proposed, and can be categorized into two type of approaches: Filter approach and wrapper approach. Filter approach select some feature basing on general characteristics of data. For each feature we assign a score indicating its importance. This approach selects a number of top ranked features and ignores the rest. On the other hand, the wrapper approach search the better features using some learning algorithms. Although it has been shown that wrapper approach is better than filter approach, it has much more computational cost, which makes it in some cases impractical. On the other hand, the filter approach is predominantly used in text categorization because of its simplicity and efficiency.
Useful features in text classification are simple words from the language vocabulary, user-specified or extracted keywords, multi-words or metadata. In text classification literature, text documents generally use words from a large vocabulary, but not all words occurring in a document are useful for classification. So, researchers have proposed feature reduction techniques like TF-IDF (Sparck Jones 1972), semantic extraction approach (Omri 2004), LSI (Liu & al. 2004) (Zhang, Yoshida, and Tang 2008), multi-word (Zhang, Yoshida, and Tang 2007) etc. or a combination of such techniques. The TF-IDF is a purely statistical technique to evaluate the importance of a word based on its frequency of occurrence in the document and in its relevant corpus(Dalal and Zaveri 2011).

According to the Ziphs Law (Zipf 2016), given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table which means, highly used words(common words) will provide very low information. Thus common words can be removed by removing most frequent words. According to Pang et al(Pang, Lee, and Vaithyanathan 2002) and Joachims (Joachims 1998), the lowest frequency words does not contain much information and can be neglected as noise words. To remove the noise words and common words, a threshold was needed to be define. This can be easy in document as there are large number of paragraphs and words thus, difference between common words and features are highly noticeable. However, not like documents which contain paragraphs, in Twitter short messages, sometimes there wont be much difference in between the frequencies of most important features and stop words. Thus, by defining a threshold, these unwanted words can remove only for some extent because of the word restriction. Removing stop words manually may cause to lose some important information because sometimes, a stop word can be a feature which provide very useful information (Dilrukshi and de Zoysa 2014).

Thus, a proper feature selection method was needed for further dimension reduction. Forward selection and backward elimination (Liu and Motoda 2007) are two popular statistical techniques which can be use for feature selection. Also TF-IDF is widely adopted statistical method for dimensional reduction of feature set. It helps to identify important words in a text document to construct vector space which improves the scalability, efficiency and accuracy of a text categorization system. These feature selection methods are applied into selected classifiers in order to measure the performance. SVM, Naive Bayes classifier and Decision Trees are the common techniques which often use for text classifications

Useful features in text classification are simple words from the language vocabulary, user-specified or extracted keywords, multi-words or metadata. In text classification literature, text documents generally use words from a large vocabulary, but not all words occurring in a document are useful for classification. So, researchers have proposed feature reduction techniques like TF-IDF (Sparck Jones 1972) , LSI (Liu & al. 2004) (Zhang, Yoshida, and Tang 2008), multi-word (Zhang, Yoshida, and Tang 2007) etc. or a combination of such techniques. The TF-IDF is a purely statistical technique to evaluate the importance of a word based on its frequency of occurrence in the document and in its relevant corpus(Dalal and Zaveri 2011). TF-IDF was used by (Cossu & al. 2014) in the task of profile categorization by assigning a weight for each term and then creating a term model for each category. This reduction method cobined with cosine similarity method give a precision of 47%. The LSI and mult -word techniques are semantics-oriented techniques. The LSI technique basically tries to use the semantics in a document structure using SVD (Singular Value Decomposition) matrix manipulations(Dalal and Zaveri 2011). A multi-word is a sequence of consecutive words having a semantic meaning. Multi-word are useful in classification as well as disambiguation. Several methods can be used to extract multi-word from text such as the frequency approach, mutual information approach, etc. By studying literature works we note that TF-IDF is widely adopted for dimensional reduction in text categorization. It helps to identify important words in a text document to construct vector space which improves the scalability, efficiency and accuracy of a text categorization system.

There are several applications for Machine Learning (ML), the most significant of which is predictive data mining. Every instance in any data set used by machine learning algorithms is represented using the same set of features. The features may be continuous, categorical or binary. If instances are given with known labels (the corresponding correct outputs) then the learning is called supervised, in contrast to unsupervised learning, where instances are unlabeled(Kotsiantis, Zaharakis, and Pintelas 2006). Various supervised machine learning techniques have been proposed in literature for the automatic classification

of text documents such as Naïve Bayes (Kim & al. 2006) (Meena and Chandran 2009), Neural Networks (Wang, He, and Jiang 2006), SVM (Support Vector Machine) (Rujiang and Junhua 2009) (Wang & al. 2006) (Zhang and Zhang 2008), Decision Tree and also by combining approaches (Goyal 2007) (Isa & al. 2008) (Yuan & al. 2008). No single method is found to be superior to all others for all types of classification. The Naïve Bayesian classifier is based on the assumption of conditional independence among attributes. It gives a probabilistic classification of a text document provided there are a sufficient number of training instances of each category. Since the Naïve Bayesian approach is purely statistical, its implementation is straightforward and learning time is less; however, its performance is not good for categories defined with very few features(Dalal and Zaveri 2011). A Decision Tree can be generated using algorithms like ID3 (Quinlan 1986) or C4.5 (Changuel, Labroche, and Bouchon-Meunier 2009) (Quinlan 2014). Unlike Naïve Bayesian classification, Decision Tree classification does not assume independence among its features. In a Decision Tree the representation of the relationship between attributes is stored as links. Decision tree can be used as a text classifier when there are relatively fewer number of attributes to consider, however it becomes difficult to manage for large number of attributes.

SVM is found to be very effective for 2-class classification problems (for example, text document belongs/ not belongs to a particular category; opinion is classified as positive/negative) but it is difficult to extend to multi-class classification. A class-incremental SVM classification approach has been proposed in (Zhang, Su, and Xu 2006), the results of the $F-measure$ for a 6 classes classification are 0.87 with one-against-one method and 0.89 with one-against-all method and Researchers have reported improved classification accuracy by combining machine learning methods. In (Goyal 2007), the performance of Neural Network based text classification was improved by assigning the probabilities derived from Naïve Bayesian method as initial weights. In (Isa & al. 2008), Naïve Bayesian method was used as a pre-processor for dimensional reduction followed by the SVM method for text classification, They have obtained a precision of 99.9%. There is a need to experiment with more such hybrid techniques in order to derive the maximum benefits from machine learning algorithms and to achieve better classification results(Dalal and Zaveri 2011).

The major inconvenient of the SVM is their complex training and classification algorithms and also high memory and time consumption during training and classification. But SVM classification method can handle document with high dimensional vector space; this makes feature selection less critical. SVM is also very efficient in binary classification problems and is very successful in real word learning problems. Results obtained in existing works (Amigó & al. 2014), show that SVM is the most adopted learning machine to construct the profiles categorization systems. In order to solve the problem of SVM algorithm complexity and

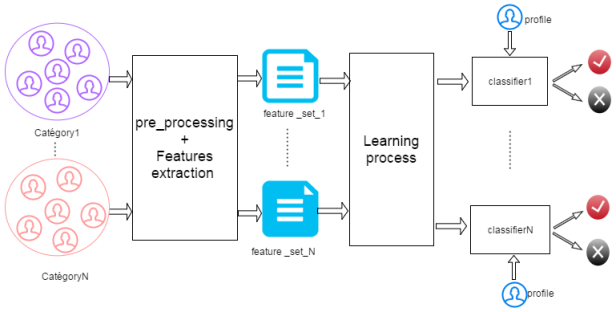

Figure 1: Profile categorization model

improve results for textual classification, we will present in the following section a new approach for profile categorization system using SVM as machine learning and presenting a new feature extraction method.

3 Proposed profile categorization model

In the previous section we have presented the literature review for textual classification. In order to solve the complexity problem and improve the categorization system performance, we present proposed profile categorization model to improve accuracy result comparing to works presented in (Amigó & al. 2014) for the microblog profile categorization task. Proposed model is based on reducing the feature vector size during feature extraction phase in order to eliminate the noisy terms and conserve only the relevant terms. The role of the profile categorization system is to assign Twitter profiles to categories (journalist, professional, sportsmen, public institution, non governmental organization, company and celebrity). An additional class undecidable was proposed to place all those users that did not match any of the proposed categories. Proposed profile categorization model is presented by the figure 1. It is divided into tree principle steps and has as input data the set of learning profiles classified by category. Those set of profile are then processed in the pre-processing step and then a set of feature is extracted for each category. After the feature extraction step we obtain a feature vector for all category that will be used to calculate the input data for the classifier construction step.
The evaluation of proposed model will be done by measuring the global accuracy value of proposed classification system. If we want to determinate the category of a profile we must first represent it using the feature vector then we test it using the seven constructed classifiers one by one. If we have not obtained result then the profile will be classified as undecidable. To ameliorate the accuracy value of proposed model we are going to reduce the size of the feature vector in order to eliminate the noisy feature. In the next section, we will explain proposed approach for features extraction.

3.1 Pre-processing and feature extraction

During the pre-processing phase we normalize some of the most common features of the Twitter jargon, which may have an influence on the performance of proposed system. For each author profile we apply: Tokenization, stop words

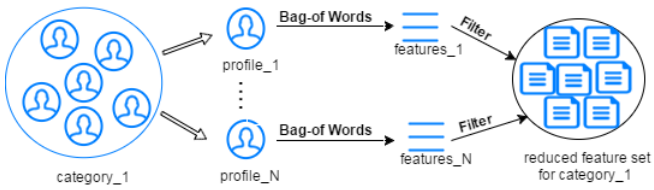

Figure 2: Feature extraction process

elimination and stemming algorithm.
The Idea is to select relevant terms for representing each category. Relevance means that this term is frequent in this category and less frequent in the others category. Proposed feature extraction model is presented by figure 2. After the preprocessing steps we obtain a Bag of words (BOW) for each profile. For each BOW we applied TF-IDF calculation then we are going to use threshold filtering methods to select only the terms having the highest TF-IDF values for each profile. The selected terms of all categories are then combined in unique feature vector. The obtained reduced vector will be used later to train and test proposed profile categorization system.
In the feature reduction step we are going to use two filtering concepts. For the first we are going to use a threshold filtering methods and for the second we are going to use a quantitative filtering method.

TF-IDF Is a feature representation method commonly used in information retrieval field. It removes from the original feature space the rare terms that are considered as non informative for classification. It removes all terms that occur no more than α times. TF-IDF is composed by two measures term frequency and inverse document frequency to calculate the importance and to measure how unique it is. A term which has a good ability to distinguish categories should have a higher wight w measured by equation 1:

$$w = TF * (\frac{1}{DF}) = TF * IDF \tag{1}$$

Where TF is defined by equation 2:

$$TF_{ij} = \frac{freq_{ij}}{maxfreq_j} \tag{2}$$

where $freq_{ij}$ is the number of $i-th$ word in the $j-th$ document, $Maxfreq_j$ is the maximum number of the frequency words in the $j-th$ document.
And IDF is defined by equation 3:

$$IDF_i = \log \frac{N}{n_i} \tag{3}$$

Where N is the total number of documents and n_i is the number of documents contains $i-th$ word.

3.2 SVM Classifier

SVM has attracted a great deal of attention in the machine learning community, the multi-class SVM is still an ongoing research issue. The existing methods can roughly be divided between two different approaches: the single machine approach, which attempts to construct a multi-class SVM by solving a single optimization problem, and the divide and conquer approach, which decomposes the multi-class problem into several binary sub-problems, and builds a standard SVM for each. The most popular decomposing strategy is probably the one against all, which consists of building one SVM per class, trained to distinguish the samples in a single class from the samples in all remaining classes. Another popular strategy is the one against one, which builds one SVM for each pair of classes. A comparison of several multi-class SVM methods has been realized by Hsu Lin (Hsu and Lin 2002). The results observed are very similar; however, the authors conclude that one against one is more practical, because the training process is quicker. Moreover, as to the claim put forward by Allwein et al.(Allwein, Schapire, and Singer 2000) that one against one and other ECOC are more accurate than the one against all strategy, Rifkin Klautau (Rifkin and Klautau 2004) disagree, arguing that the one against all strategy is as accurate as any other approach, assuming that the SVMs are well tuned. Thus, according to the literature review, it seems impossible to conclude which multi-class SVM is better for classification problem. For this reason, we chose to compare the two most popular strategies, which are one against all and one against one(Milgram, Cheriet, and Sabourin 2006).

One-Against-One SVM classifier One-Against-One (OAO) method involves $\frac{N*(N-1)}{2}$ binary SVM classifiers. Each classifier is trained to separate each pair of classes. There are different strategies used to combine these binary classifiers. The main strategies widely used in literature are Pairwise Coupling and a majority voting strategy which is called MaxWins. When classifiers are combined through majority voting scheme, the class with maximal number of votes is the estimation. In pairwise coupling (Hastie and Tibshirani 1998), a pairwise probability p_{ij} , is obtained from each binary SVM output noted as $f_{ij}(x)$.

$$p_{ij} = 1/2 f_{ij}(x) + 0.5 \tag{4}$$

These pairwise probabilities are coupled into a common set of posterior probabilities p_i :

$$P_i = 2/N(N-1) \sum_{j \neq i} p_{ij} \tag{5}$$

The decision function is given by:

$$C(x) = arg \max_{1 \leq i \geq N} P_i \tag{6}$$

One-Against-All SVM classifier One-Against-All (OAA) is the most common and simplest approach (Liu and Zheng 2005). It involves N binary SVM classifiers, one for each class. Each binary SVM is trained to separate one class from the rest. The winning class is the one that corresponds to the SVM with highest output value i.e. the largest decision function value. This approach may suffer from error

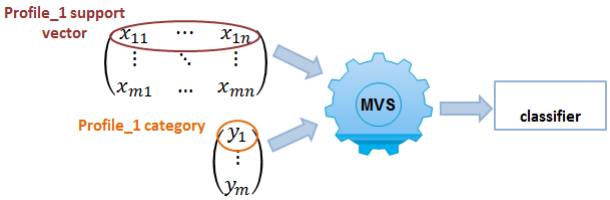

Figure 3: classifier construction model

caused by markedly imbalanced training sets. The decision function for OAA is :

$$C(x) = arg \max_{1 \leq i \geq N} f_i(x) \quad (7)$$

Where $f_i(x)$ is the output of the binary SVM classifier trained for class i against all the other classes.

classifier construction process Proposed profile categorization system is formed by $\frac{N*(N-1)}{2}$ sub classifier when using OAO-SVM classifier and by N sub classifiers when using OAA-SVM classifier we represent in figure 3 an example of classifier construction model of proposed system. This model has as input the training matrix (each line is a profile support vector) and the vector of profile category each y_i is the category value corresponding to the profile support vector ($x_{i1}x_{in}$) and as output the classifier system.

4 Experimentation and results analysis

In this step we start by applying a pre-processing to the learning data set. we obtain a bag of words of all the collection, then for each term in each category we calculate the TF-IDF weight using equations 1, 2 and 3. To improve the performance of proposed categorization system we proposed to reduce the feature set using two methods: Reduction using frequency threshold (RFT) and Reduction using quantitative method (RQM).
The Data set used to construct and test proposed profile categorization system is provided by the organization RepLab2014. It is divided into learning data set and test data set. The learning data set is composed by 1500 English twitter profiles and the test data set is composed by 1350 English twitter profiles. All profiles have at least 1000 followers that represent the automotive and banking domains. Each profile contains (i) screen name; (ii) profile URL, and (iii) the last 600 tweets published by the author at crawling time(Amigó & al. 2014).

The evaluation measures used are precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances, and f-measures is a combination between precision and recall.

Table 1: Precison results of RFT method

Threshold	**0.05**	**0.01**	**0.015**	**0.02**
celebrity	0.2	0.33	0.33	0.33
professional	0.63	0.64	0.69	0.65
journalist	0.27	0.47	0.47	0.47
company	0.15	0.19	0.40	0.18
NGO	0.00	0.00	0.00	0.00
Average precision	0.25	0.33	**0.38**	0.33

Table 2: Recall results for RFT method

Threshold	**0.05**	**0.01**	**0.015**	**0.02**
celebrity	0.2	0.6	0.8	0.6
professional	0.63	0.64	0.67	0.64
journalist	0.18	0.88	0.96	0.88
company	0.21	0.28	0.78	0.28
NGO	0.00	0.00	0.00	0.00
Average recall	0.24	0.48	**0.64**	0.48

Table 3: F-Mesures results for RFT method

Threshold	**0.05**	**0.01**	**0.015**	**0.02**
celebrity	0.2	0.42	0.46	0.42
professional	0.63	0.64	0.64	0.64
journalist	0.21	0.61	0.63	0.61
company	0.17	0.22	0.66	0.22
NGO	0.00	0.00	0.00	0.00
Average F-Measures	0.24	0.37	**0.48**	0.37

The first test using proposed profile categorization system is done by selecting the first 1000 terms having the highest TF-IDF value for the feature vector, but for this test the system doesn't converge. For the RFT approach The threshold is defined first as the average of TF-IDF weights, we retain only terms having TF-IDF higher than this threshold. The rest of test is done by varying the threshold in order to obtain the optimal threshold value which give the better precision value for proposed system. Obtained precision result are presented in table 1, recall results are presented in table 2 and f-measure result are presented in table 3.The best precision and recall result are obtained for value of threshold equal to 0.015 . For the RQM We retain only half of terms having the highest TF-IDF values. The accuracy value obtained for this method is equal to 0.26.

We present in table 5 a comparison between OAA-SVM and OAO-SVM classifiers when using proposed feature extraction method. We assume that filtering threshold is equal to 0.015, we deduce that OAO-

Table 4: RFT vs RQM

	RFT	RQM
size of feature vector	5661	528553
best precision result	0.38	0.26

Table 5: OAA-SVM classifier vs OAO-SVM classifier for profile categorization

method	precision	recall	f-measure
OAO-SVM	0.48	0.64	0.48
OAA-SVM	0.38	0.48	0.47

SVM classifier perform well and improves system performance when solving a multi-class profile categorization problem.

In table 6 we present the rank of proposed approach comparing to other works presented in (Amigó & al. 2014) during RepLab 2014 competition.

5 Conclusion

In this paper we present our approach based on TF-IDF with threshold filtering methods for feature extraction and SVM as learning machine to enhance the precision result of the microblog categorization system. Results obtained validate that our approach perform very well. As future works we can integrate semantic aspect in the feature extraction process like using LSI method, also in learning process using ontology.

References

Aggarwal, C. C., and Zhai, C. 2012. A survey of text classification algorithms. In *Mining text data*. Springer. 163–222.

Allwein, E. L.; Schapire, R. E.; and Singer, Y. 2000. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of machine learning research* 1(Dec):113–141.

Amigó, E.; Carrillo-de Albornoz, J.; Chugur, I.; Corujo, A.; Gonzalo, J.; Meij, E.; de Rijke, M.; and Spina, D. 2014. Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*. Springer. 307–322.

Table 6: precision results for other profile categorization system

Rank	Method	Precision
1	**RFT+OAO-SVM**	**0.48**
2	TF-IDF+cosine similarity+HMM	0.47
3	BOW+SVM	0.46
4	DF(document frequency)+SVM	0.41
5	N-grams+Weka	0.14

Chakrabarti, S.; Dom, B.; Agrawal, R.; and Raghavan, P. 1997. Using taxonomy, discriminants, and signatures for navigating in text databases. In *VLDB*, volume 97, 446–455.

Changuel, S.; Labroche, N.; and Bouchon-Meunier, B. 2009. Automatic web pages author extraction. In *International Conference on Flexible Query Answering Systems*, 300–311. Springer.

Cossu, J.-V.; Janod, K.; Ferreira, E.; Gaillard, J.; and El-Bèze, M. 2014. Lia@ replab 2014: 10 methods for 3 tasks. In *4th International Conference of the CLEF initiative*.

Dalal, M. K., and Zaveri, M. A. 2011. Automatic text classification: a technical review. *International Journal of Computer Applications* 28(2):37–40.

Dilrukshi, I., and de Zoysa, K. 2014. A feature selection method for twitter news classification. *International Journal of Machine Learning and Computing* 4(4):365.

Goyal, R. D. 2007. Knowledge based neural network for text classification. In *Granular Computing, 2007. GRC 2007. IEEE International Conference on*, 542–542. IEEE.

Hastie, T., and Tibshirani, R. 1998. Classification by pairwise coupling. In *Advances in neural information processing systems*, 507–513.

Heisele, B.; Ho, P.; and Poggio, T. 2001. Face recognition with support vector machines: Global versus component-based approach. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, 688–694. IEEE.

Hsu, C.-W., and Lin, C.-J. 2002. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks* 13(2):415–425.

Isa, D.; Lee, L. H.; Kallimani, V.; and Rajkumar, R. 2008. Text document preprocessing with the bayes formula for classification using the support vector machine. *IEEE Transactions on Knowledge and Data engineering* 20(9):1264–1272.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* 137–142.

Kim, S.-B.; Han, K.-S.; Rim, H.-C.; and Myaeng, S. H. 2006. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering* 18(11):1457–1466.

Kotsiantis, S. B.; Zaharakis, I. D.; and Pintelas, P. E. 2006. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review* 26(3):159–190.

Liu, H., and Motoda, H. 2007. *Computational methods of feature selection*. CRC Press.

Liu, B., and Zhang, L. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer. 415–463.

Liu, Y., and Zheng, Y. F. 2005. One-against-all multi-class svm classification using reliability measures. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 2, 849–854. IEEE.

Liu, T.; Chen, Z.; Zhang, B.; Ma, W.-y.; and Wu, G. 2004. Improving text classification using local latent semantic indexing. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, 162–169. IEEE.

Meena, M. J., and Chandran, K. 2009. Naive bayes text classification with positive features selected by statistical method. In *2009 First International Conference on Advanced Computing*, 28–33. IEEE.

Milgram, J.; Cheriet, M.; and Sabourin, R. 2006. one against one or one against all: Which one is better for handwriting recognition with svms? In *Tenth international workshop on frontiers in handwriting recognition*. Suvisoft.

Omri, M. 2004. Possibilistic pertinence feedback and semantic networks for goal extraction. *Asian Journal of Information Technology* 3(4):258–265.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 79–86. Association for Computational Linguistics.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1(1):81–106.

Quinlan, J. R. 2014. *C4. 5: programs for machine learning*. Elsevier.

Rangel, F.; Rosso, P.; Koppel, M. M.; Stamatatos, E.; and Inches, G. 2013. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, 352–365. CELCT.

Rifkin, R., and Klautau, A. 2004. In defense of one-vs-all classification. *Journal of machine learning research* 5(Jan):101–141.

Rujiang, B., and Junhua, L. 2009. A novel conception based texts classification method. In *Advanced Science and Technology, 2009. AST'09. International e-Conference on*, 30–34. IEEE.

Sendi, M.; Omri, M. N.; and Abed, M. 2017. Possibilistic interest discovery from uncertain information in social networks. *Intelligent Data Analysis* 21(6):1425–1442.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28(1):11–21.

Vapnik, V. N. 1998. *Statistical learning theory*, volume 1. Wiley New York.

Vapnik, V. N. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks* 10(5):988–999.

Wang, Z.-Q.; Sun, X.; Zhang, D.-X.; and Li, X. 2006. An optimal svm-based text classification algorithm. In *2006 International Conference on Machine Learning and Cybernetics*.

Wang, Z.; He, Y.; and Jiang, M. 2006. A comparison among three neural networks for text classification. In *2006 8th international Conference on Signal Processing*, volume 3. IEEE.

Widodo, A., and Yang, B.-S. 2007. Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical systems and signal processing* 21(6):2560–2574.

Yuan, P.; Chen, Y.; Jin, H.; and Huang, L. 2008. Msvm-knn: Combining svm and k-nn for multi-class text classification. In *Semantic Computing and Systems, 2008. WSCS'08. IEEE International Workshop on*, 133–140. IEEE.

Zhang, M., and Zhang, D.-x. 2008. Trained svms based rules extraction method for text classification. In *IT in Medicine and Education, 2008. ITME 2008. IEEE International Symposium on*, 16–19. IEEE.

Zhang, B.-f.; Su, J.-s.; and Xu, X. 2006. A class-incremental learning method for multi-class support vector machines in text classification. In *2006 International Conference on Machine Learning and Cybernetics*, 2581–2585. IEEE.

Zhang, W.; Yoshida, T.; and Tang, X. 2007. Text classification using multi-word features. In *2007 IEEE International Conference on Systems, Man and Cybernetics*, 3519–3524. IEEE.

Zhang, W.; Yoshida, T.; and Tang, X. 2008. Tfidf, lsi and multi-word in information retrieval and text categorization. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, 108–113. IEEE.

Zipf, G. K. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.