Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2018) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Topic-based Classification through Unigram Unmasking

Yaakov HaCohen-Kerner[a, 1], Avi Rosenfeld[b]

Asaf Sabag[a], Maor Tzidkani[a]

[a]*Dept. of Computer Science, Jerusalem College of Technology, 9116001 Jerusalem, Israel*
[b]*Dept. of Industrial Engineering, Jerusalem College of Technology, 9116001 Jerusalem, Israel*

Abstract

Text classification (TC) is the task of automatically assigning documents to a fixed number of categories. TC is an important component in many text applications such as text indexing, information extraction, information retrieval, text mining, and word sense disambiguation. In this paper, we present an alternative method of feature reduction - a concept we call unigram unmasking. Previous text classification approaches have typically focused on a "bag-of-words" vector. We posit that at times some of the most frequent unigrams, which have the greatest weight within these vectors, are not only unnecessary for classification, but can at times even hurt models' accuracy. We present an approach where a percentage of common unigrams are intentionally removed, thus "unmasking" the added value from less popular unigrams. We present results from a topic-based classification task (hundreds of online free text-books belonging to five domains: Career and Study Advice, Economics and Finance, IT Programming, Natural Sciences, Statistics sand Mathematics) and show that unmasking was helpful across several machine learning models with some models even benefiting from removing nearly 50% of the most frequent unigrams from the bag-of-word vectors.

Keywords: Bag of words; Overfitting Features; Supervised machine learning; Textual features; Text classification; Topic-based classification Unmasking; Word unigrams.

1. Introduction

Text classification (TC) is a supervised learning task that assigns natural language text documents to one (the typical case) or more predefined categories (Joachims, 1998)[1]. Classification algorithms typically use a supervised

[1] Corresponding author. Tel.: +972-2-6751018; fax: +972-2-6751046.
E-mail address: kerner@jct.ac.il

machine learning (ML) algorithm or a combination of several ML algorithms (Sebastiani, 2002[2]; Jain and Mandowara, 2016[3]).
TC is an important component in many research domains such as text indexing, information extraction, information retrieval, text mining, and word sense disambiguation (Pazienza, 1997[4]; Knight, 1999[5]; Sriram et al., 2010[6]; Navigli et al., 2011[7]; Zhou et al., 2016[8]). There are two main types of TC: topic-based classification where a given document is ascribed to one of c>=2 categories, and stylistic classification where a document is ascribed to one of c>=2 writing styles. An example of a topic-based classification application is classifying news articles written in English that belong to four categories: Business-Finance, Lifestyle-Leisure, Science-Technology and Sports downloaded from three well-known news web-sites (BBC, Reuters, and TheGuardian) (Liparas et al., 2014[9]). An example of a stylistic classification application is classification based on different literary genres, e.g., action, comedy, crime, fantasy, historical, political, saga, and science fiction (Kessler et al., 1997[10], Gianfortoni et al., 2011[11]). These two classification tasks often require different types of features for best performing the learning. Whereas stylistic classification is typically performed using linguistic features such as quantitative features, orthographic features, part of speech (POS) tags, function words, and vocabulary richness features, topic-based classification is typically performed using unigrams and/or ngrams (for n > 2) (Argamon et al., 2007[12]; HaCohen-Kerner et al., 2008[13], HaCohen-Kerner et al., 2010A[14]; HaCohen-Kerner et al., 2010B[15]).

The traditional model for topic-based TC is based on the bag-of-words (BOW) representation, which associates a text with a vector indicating the number of occurrences of each chosen word in the training corpus (Sebastiani, 2002[2]). In a topic-based classification, Naive Bayes and Maximum Entropy (ME) (e.g., Jaynes, 1990[16], El-Halees, 2015[17]), support vector machines (SVMs) (Cortes and Vapnik, 1995[18]), Naive Bayes (NB) (Heckerman, 1997[19]), and C4.5 decision tree induction (Quinlan, 2014[20]) have been reported to use BOW representation to achieve accuracies of 90% and greater for particular categories (Joachims, 1998[1]).There are two main types of TC: topic-oriented classification and stylistic classification. An example of a topic-oriented classification application is classifying news articles as Business-Finance, Lifestyle-Leisure, and Sports (Liparas et al., 2014[9]). An example of a genre-oriented classification application is classifying between different literary genres, e.g., action, comedy, fantasy, political, and saga (Kessler et al., 1997[10]). While stylistic classification is usually performed using linguistic features such as quantitative features, part of speech (POS) tags, function words and vocabulary richness features, topic-oriented classification is usually performed using unigrams and/or ngrams (for n>2) (Argamon et al, 2007[12]).

This paper claims that classification models using BOWs should at times intentionally remove the most frequent unigrams. Previous work has suggested removing stopword unigrams (e.g., Forman (2003)[26]). We claim that many more frequent unigrams should also be removed – even those that are not stopwords. In contrast to previous research that focused on results from these strong indicators (Schler et al., 2006[21]), we claim that such usage of frequent unigrams is not only at times unnecessary, but can actually produce less accurate models than those where only less common unigrams are used. We refer to the process of intentionally removing a percentage of common unigrams as "unigram unmasking". We borrow the term "unmasking" from Koppel et al.'s (2007)[22] work where they demonstrated that the removal of a small number of frequent unigrams is useful for identifying the author of an anonymous text.

This paper's key contribution is presenting the idea of unigram unmasking and demonstrating its empirical significance. We found that unigram unmasking can be successfully used within content classification problems. We created a dataset of texts from 5 topic categories and classified them using the 5000 most frequent unigrams. We found that removing up to nearly 50% of the most frequent unigrams from the training set often had no effect on models' classification accuracy. In fact, as we detail in the following sections, unigram unmasking often improved models' accuracy.

This paper is organized as follows: Section 2 supplies relevant background about unmasking of word ngrams. Section 3 presents our methods and the examined corpus. Section 4 describes the experimental results and their analysis. Section 5 concludes and suggests ideas for future research.

2. Unmasking of Word Ngrams

To date, many works on stylistic classification have focused on results generated from BOW vectors, and particularly the most popular terms in these vectors (e.g., Pang et al., 2002[23]; Schler et al., 2006[21]; Ng et al., 2006[24]; Abbasi et al., 2008[25]). Schler et al. (2006)[21] performed gender and age classification tasks on a corpus containing

37,478 blogs and 295,526,889 words. They used 502 stylistic features (POS tags, blog words, and hyperlinks), and 1000 unigrams with the highest information gain in the training set. They obtained an accuracy of 80.1% for gender classification and 76.2% for age classification with three categories: 10s (13-17), 20s (23-27), and 30s (33-42) using the Multi-Class Real Winnow algorithm. Among their findings they show that male bloggers use more words related to money, computers, job, sports and television, while female bloggers use more words related to family, friends, sleep, and eating. Forman (2003)[26] proposed a feature selection metric called Bi-Normal Separation. He presented an extensive comparative study of feature selection metrics. In his experiments on various datasets, Forman used from 10 to 2000 frequent word unigrams. He showed that N-gram frequencies provide accuracy rate of about 98% for some of the datasets. Other works that performed TC according to categories (e.g., disciplines, domains, and topics) using various types of word ngrams are (Fürnkranz, 1998[27]; Martins and Silva, 2005[28]; Liparas et al., 2014[9]). We stress that in all of the above works results typically focused on the most frequent words or expressions.

Most similar to our work, Koppel et al. (2007)[22] investigated the identity of the author of an anonymous text using unmasking of artificial writing features. Their work defined a process of "unmasking" whereby the most common writing features were intentionally removed as they hypothesized that writers intentionally hid their identity through adding a set of features they typically did not use. Through using the unmasking process they show that the authors of several previous anonymous writers can be identified based on comparing their unmasked writing features with those of texts from the same area and geographic region whose authors are known.

We stress that while we use the "unmasking" terminology first described by Koppel et al. (2007)[22], our motivation and usage is quite different. Their work aimed to remove a relatively small number of useful features in order to unmask the identity of a previously unknown author. In contrast, our process of unmasking is more similar to feature selection. We acknowledge that proper identification can be accomplished without the unmasking process, as has been previously done by Joachims (1998)[1], but we claim that even higher accuracies can be achieved through this specialized type of feature reduction. Additionally, while Koppel et al. (2007)[22] removed only a few features (3-10), we found that removing even large percentages of unigrams can at times aid in improving classification accuracy. This process is again more akin to the process of classic feature reduction. However, feature selection in text classification typically chooses only those features with the highest value of various scoring functions such as info gain (IG), mutual information (MI), χ2-test (CHI) (Yang and Pedersen, 1997[29]), TF-IDF (Salton et al. 1975[30]) and Principal Component Analysis (PCA) (Lam and Lee, 1999[31]), and various local and global feature reduction functions (Sebastiani, 2002[2]). Feature selection then removes all other features with lower scores. In contrast, and as we now further explain, we found that intentionally removing common features with high frequency in addition to stopwords that have been previously removed often aids in classification accuracy - something that at first glance seems surprising and counterintuitive.

3. Methods and Corpus Description

Our general claim is that even if the first X unigrams are removed from the training set, the overall accuracy will drop by no more than a small amount, and will often even improve. To empirically support this claim, we downloaded hundreds of online free text-books belonging to five domains: Career and Study Advice, Economics and Finance, IT Programming, Natural Sciences, Statistics sand Mathematics. We then used a commercial, off-the-shelf conversion program (www.abbyy.com) to convert the PDF files of these books into text files that could be used for the feature extraction process that will be described later. Table 1 presents general details about the dataset, its domains, # of books, # of words and average # of words per book for each domain.

Table 1. General details about the dataset.

#	Domain	# of books	# of words	Avg. # of words per book
1	Career and Study Advice	46	883,033	19,196
2	Economics and Finance	55	2,041,942	37,126
3	IT Programming	109	3,585,815	32,897
4	Natural Sciences	60	1,925,589	32,093
5	Statistics and Mathematics	115	4,198,052	36,504
	Total	385	12,634,431	32,817

The following process was applied to dataset described in Table 1 in order to identify the 5000 most common unigrams over all classes:

(1) All instances of 421 known stopwords for English text (Fox, 1989[32]) were deleted.
(2) All unigrams were identified and counted.
(3) The frequencies of all unigrams were sorted in descending order to identify the 5000 most common unigrams over all classes.

We applied four supervised ML methods using the WEKA platform with their default parameters (Hall et al., 2009[33]; Witten et al., 2016[34]) and the accuracy of each ML method was estimated by a 10-fold cross-validation testing. We considered the accuracy of models that classify each text by its topic into one of these five categories. We considered four ML algorithms: Reduced error pruning (REP) Tree, J4.8, Random Forest (RF), and BayesNet (BN). A brief description of these algorithms follows:

1. REPTREE is a fast decision tree learner, which builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning with back fitting (Witten et al., 2016[34]). Because the tree grows linearly with the size of the samples presented, no accuracy is gained through the increased tree complexity, and pruning is beneficial and is carefully performed (Elomaa and Kääriäinen, 2001[35]).

2. J48 is an improved variant of the C4.5 decision tree induction (Hormann, 1962[36]), which is implemented in WEKA. J48 uses greedy techniques, determines the most predictive attribute at each step, and splits a node based on this attribute.

3. RF is an ensemble learning method for classification and regression (Breiman, 2001[37]). Ensemble methods use multiple learning algorithms to obtain better predictive performance than what can be obtained from any of the constituent learning algorithms. RF operates by constructing a multitude of decision trees at training time and outputting classification for the case at hand.

4. BN is a variant of a probabilistic statistical classification model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (Pourret et al., 2008[38]).

4. Experimental Results

We began by classifying the documents that are included in the dataset described in Table 1 into the five domains using a BOW based on the 5000 most frequent unigrams. We immediately noticed that certain words (word unigrams) were extremely clearly linked to the given categories in the classification task. For example, the words “career” and “careers” are not stopwords yet were obviously found more with higher frequency in the “Career and Study Advice” set of texts. As further examples, some other obvious unigrams and their relative rank (in parentheses) among the 5000 most frequent words are shown below and in Table 2: Career and Study Advice: career (386), careers (838), study (347), studies (881), and advice (2046); Economics and Finance: rate (22), economics (668), economical (1096), finance (493), money (94), price (76), and prices (494); IT Programming: information (16), program (185), programming (329), programs (1243), technology (59), and technological (2073); Natural Sciences: science (415), natural (593), chemistry (298), chemical (200); physics (1785); and physical (771); and Statistics and Mathematics: statistics (457), statistical (1432), mathematics (499), mathematical (629), stochastic (497), and radius (658).

Table 2. Obvious words in the dataset for the five domains.

Domain	Career and Study Advice	Economics and Finance	IT Programming	Natural Sciences	Statistics and Mathematics
	career (386)	Rate (22)	information (16)	science (415)	statistics (457)
	careers (838)	economics (668)	program (185)	natural (593)	statistical (1432)
	Study (347)	economical (1096)	programming (329)	chemistry (298) chemical (200)	mathematics (499)
	studies (881)	finance (493)	programs (1243)		mathematical (629)
	advice (2046)	Money (94)	technology (59)	physics (1785)	stochastic (497)
		price (76)	technological (2073)	physical (771)	radius (658)
		prices (494)			

Previous works have often focused on such stark differences, which often are not inherently obvious. As previously mentioned, Schler et al. [2006] performed gender and age classification tasks on a blog corpus using 1000 unigrams. They found that male bloggers often used more specific words related to computers (e.g., Linux, Microsoft, programming, software, and Google), while female bloggers use more specific words related to family (e.g., Mom, Mommy, and husband). However, this paper makes what we consider to be a stronger claim that even when removing such common words, which we concede that in our corpus can at times be considered trivial differences, the classification models often improves. We found this finding to be counterintuitive yet indicates the importance of feature reduction of unigrams.

To study this point, we proceeded to systematically remove percentages of the most common words. We iteratively removed 2% of the most common unigrams from the total of 5000. Thus, we first removed the 100 most common unigrams, leaving a BOW with the remaining 4900, we then removed the next 100 most common unigrams leaving 4800, etc. We then observed the relative accuracy of the four ML classification models, which are shown in Fig. 1 with resolution of 2%. Table 3 presents part of the same results with resolution of 10% except the last row with drop of 98% of the unigrams (since a drop of 100% is meaningless).

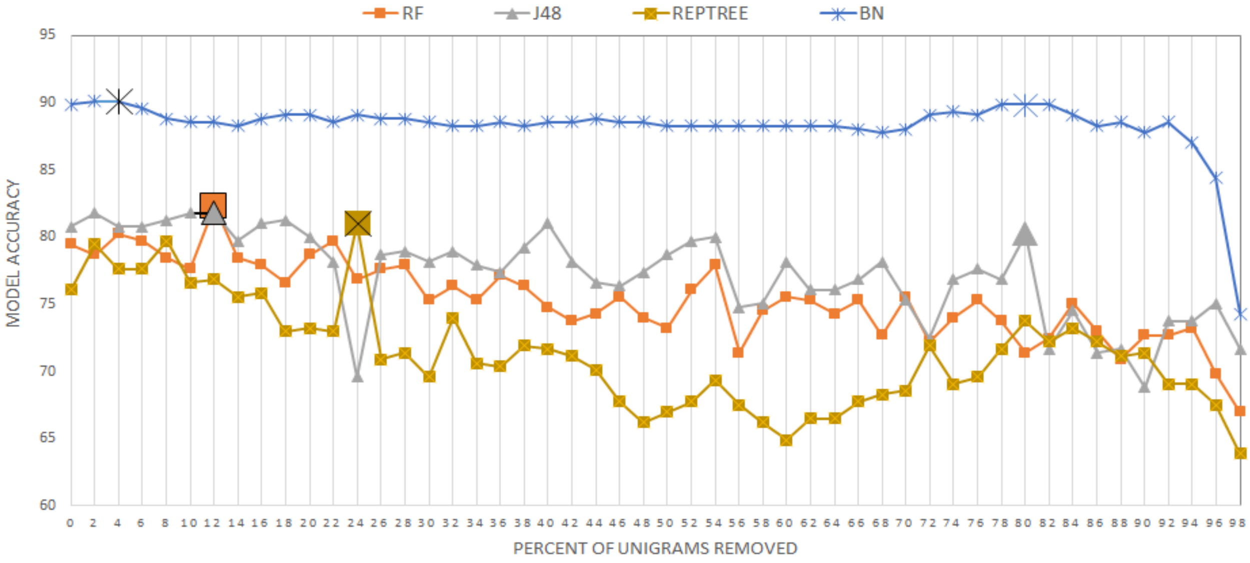

Fig. 1. Accuracy of the four ML classification model as function of the percent of unigrams removed with resolution of 2%.

Table 3. Accuracy of the four ML classification model as function of the percent of unigrams removed with resolution of 10%.

% of dropped features	# of dropped features	BN	REPTREE	J48	RF
No filter	0	89.87	76.1	80.77	79.48
10	500	88.57	76.62	81.81	77.66
20	1000	89.09	73.24	80	78.7
30	1500	88.57	69.61	78.18	75.32
40	2000	88.57	71.68	81.03	74.8
50	2500	88.31	67.01	78.7	73.24
60	3000	88.31	64.93	78.188	75.58
70	3500	88.05	68.57	75.32	75.58
80	4000	**89.87**	73.76	**80.25**	71.42
90	4500	87.79	71.42	68.83	72.72
98	4900	74.28	63.89	71.68	67.01

Several things are interesting about the results presented in Figure 1 and Table 3. First, the highest accuracy overall is found in the BN model. Even within this model, we found that removing unigrams helped up until a certain point. Specifically, we found that the BN accuracy after removing the first (!) 200 most common unigrams (4%) was 90.12%, as opposed to an accuracy of 89.12% with no unigrams removed. We denote this datapoint with a larger marker in Figure 1. For both RF and J48, the best accuracy was achieved after removing 12% of the most common unigrams (600 of 5000 unigrams). RF achieved 82.33% accuracy com-pared to 79.48% with all unigrams, and J48 achieved 81.87% compared to an original value of 80.77%. REP achieved its best result after removing 24% or 1200 of the 5000 unigrams. Here it achieved an accuracy of 81.03% versus 76.1%. A second interesting phenomenon relates to the drop of the models as unigrams are removed. The BN model is the most stable with results around 89.89% still being achieved after 80% (!) of its unigrams were removed (datapoint enlarged in graph) which is slightly better than 89.12% obtained when no unigrams were removed! Similarly J48 achieved 80.25% accuracy at this point as well-similar to the original result of 80.77%. REP and RF seem to be less resilient to unigram removal and have a sharper maximum once a certain number of unigrams are removed.

We believe that one potential explanation for this result is that many of the removed unigram features belong to more than one domain and therefore constitute noise. Examples of such unigrams with their ranks among the 5000 most frequent words in brackets are: download (4), ebooks (5), example (7), data (9), figure (10), value (11), read (14), solution (18), equation (19), table (21), system (23), model (40), create (48), theory (54), method (58), analysis (74), research (157), document (169), approach (361), and tool (709). However, this alone does not seem to fully explain these results and it does seem that text classification often improves as the most common, non-trivial unigrams are re-moved. The reasons for this are something we are currently both empirically and theoretically studying, as we now detail.

5. Conclusions and future work

In this paper, we present feature reduction approach of "unmasking" whereby common unigrams are intentionally removed from training models. We found that the accuracy of the TC models we developed always benefited from this approach up until a certain percentage of com-mon unigrams were removed, after which the models' were negatively impacted by the removal additional terms. This work represents a paradigm shift as previous content classification research has typically focused on the results based on the most frequent unigrams, which we show often actually hurts performance.

For future work various avenues need to be considered. First, we hope to study how to identify and predict the maxima after which removing additional terms is detrimental. Second, we hope to study the theoretical foundations for term unmasking. Third, we hope to study the generality of this approach, not only for additional datasets of TC per topic, but also for other classification tasks e.g., stylistic, gender, or genre classification. Similarly, we hope to study if effect of unmasking of other features such as n-grams of n>1 and stylistic features is equally helpful. We believe that the general approach we present in this paper will lead to significant advances in these areas.

Acknowledgements

This work was partially funded by the Jerusalem College of Technology, Lev Academic Center and we gratefully acknowledge their support. We would also like to thank two anonymous reviewers for their useful and instructive comments.

References

[1] Joachims, Thorsten. (1998) “Text categorization with support vector machines: Learning with many relevant features.” In Proceedings of the European Conference on Machine Learning (ECML). 137–142.

[2] Sebastiani, Fabrizio (2002) “Machine learning in automated text categorization.” ACM Computing Surveys. 34. 1, 1-47.

[3] Jain, A., and Mandowara, J. (2016) "Text classification by combining text classifiers to improve the efficiency of classification." International Journal of Computer Application (2250-1797), 6(2).

[4] Pazienza, M. T. (Ed.). (1997) "Information Extraction: A multidisciplinary approach to an emerging information technology." (Vol. 1299). Heidelberg: Springer.

[5] Knight, Kevin. (1999) “Mining online text.” Communications of the ACM. 42(11), 58-61.

[6] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010) "Short text classification in twitter to improve information filtering." In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 841-842). ACM.

[7] Navigli, R., Faralli, S., Soroa, A., de Lacalle, O., and Agirre, E. (2011) "Two birds with one stone: learning semantic models for text categorization and word sense disambiguation." In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 2317-2320). ACM.

[8] Zhou, Y., Tong, Y., Gu, R., and Gall, H. (2016) “Combining text mining and data mining for bug report classification.” Journal of Software: Evolution and Process, 28(3), 150-176.

[9] Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., and Kompatsiaris, I. (2014) "News articles classification using Random Forests and weighted multimodal features." In Information Retrieval Facility Conference (pp. 63-75). Springer, Cham.

[10] Kessler, Brett, Nunberg, Geoffrey, and Hinrich Schutze. (1997) “Automatic detection of text genre.” In P. R. Cohen & W. Wahlster (Eds.), In Proceedings of the 35th annual meeting of the ACL and eighth conference of the European chapter of the Association for Computational Linguistics. 32-38, Somerset, New Jersey: Association for Computational Linguistics.

[11] Gianfortoni, P., Adamson, D., and Rosé, C. P. (2011). "Modeling of stylistic variation in social media with stretchy patterns." In Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties (pp. 49-59). Association for Computational Linguistics.

[12] Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan R. Hota, Navendu Garg, and Shlomo Levitan. (2007) “Stylistic text classification using functional lexical features: Research articles.” Journal of the American Society for Information Science and Technology. 58, 6, 802-822.

[13] HaCohen-Kerner, Y., Mughaz, D., Beck, H., and Yehudai, E. (2008) "Words as classifiers of documents according to their historical period and the ethnic origin of their authors." Cybernetics and Systems: An International Journal, 39(3), 213-228.

[14] HaCohen-Kerner, Y., Beck, H., Yehudai, E., Rosenstein, and M., Mughaz, D. (2010A) “Cuisine: Classification using stylistic feature sets &/or name-based feature sets.” Journal of the American Society for Information Science and Technology 61 (8), 1644–57.

[15] HaCohen-Kerner, Y., Beck, H., Yehudai, E., and Mughaz, D. (2010B) “Stylistic feature sets as classifiers of documents according to their historical period and ethnic origin.” Applied Artificial Intelligence 24 (9), 847–62.

[16] Jaynes, E. T. (1990) “Notes on Present Status and Future Prospects.” In Maximum Entropy and Bayesian Methods, edited by W. T. Grandy and L. H. Schick. Kluwer, 1-13.

[17] El-Halees, Alaa M. (2015) "Arabic text classification using maximum entropy." IUG Journal of Natural Studies 15.1

[18] Cortes, C., and Vapnik, V. (1995) “Support-vector networks. Machine Learning,” 20 (3), 273–297.

[19] Heckerman, D. (1997) “Bayesian networks for data mining.” Data mining and knowledge discovery, 1(1), 79-119.

[20] Quinlan, J. R. (2014) “C4. 5: programs for machine learning.” Elsevier.

[21] Schler, Jonathan, Koppel, Moshe, Argamon, Shlomo, and Pennebaker, James W. (2006). "Effects of age and gender on blogging." In AAAI spring symposium: Computational approaches to analyzing weblogs (Vol. 6, pp. 199-205).

[22] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. (2007) “Measuring Differentiability: Unmasking Pseudonymous Authors.” Journal of Machine Learning Research, 8(2), 1261-1276.

[23] Pang, Bo, Lee, Lillian, and Vaithyanathain, Shivakumar. (2002) “Thumbs Up? Sentiment Classification Using Machine Learning Techniques,”Proc. Conf.Empirical Methods in Natural Language Processing, pp. 79-86.

[24] Ng, Vincent, Dasgupta, Sajib, and Arifin, S. M. Niaz. (2006) “Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews,” Proc. Conf. Computational Linguistics, Assoc. for Computational Linguistics, pp. 611-618.

[25] Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. (2008) "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums." ACM Transactions on Information Systems (TOIS) 26.3: 12.

[26] Forman, George. (2003) “An extensive empirical study of feature selection metrics for text classification.” Journal of machine learning research, 3(Mar), 1289-1305.

[27] Fürnkranz, Johannes. (1998) “A study using n-gram features for text categorization.” Austrian Research Institute for Artificial Intelligence 3: 1-10.

[28] Martins, Bruno and Silva, Mário J. (2005) “Language identification in web pages.” In Proceedings of the 2005 ACM symposium on applied computing. 764-768. ACM.

[29] Yang, Yiming and Pedersen, Jan O. (1997) “A comparative study on feature selection in text categorization.” In Proceedings of the 14th International Conference on Machine Learning (ICML), 412-420.

[30] Salton, G., Wong, A., Yang, C. S. (1975) "A vector space model for automatic indexing." Communications of the ACM, 18(11), 613-620.

[31] Lam, Savio L. and. Lee, Dik, L. (1999) “Feature reduction for neural network based text categorization.” In Proceedings of DASFAA-99, 6th IEEE International Conference on Database Advanced Systems for Advanced Application (Hsinchu, TW, 1999), 195-202. IEEE.

[32] Fox, Christopher C. (1989) “A Stop List for General Text.” ACM SIGIR Forum, 24, 1-2, 19-35.

[34] Mark Hall. Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten. (2009) “The WEKA data mining software: an update.” ACM SIGKDD explorations newsletter, 11(1), 10-18.

[34] Witten, Ian H., Frank, Eibe, Hall, Mark A., and Pal, Christopher J. (2016) "Data Mining: Practical machine learning tools and techniques." Morgan Kaufmann.

[35] Elomaa, Tapio and Kääriäinen, Matti (2001) “An analysis of reduced error pruning.” J. Artificial Intelligence Res. 15:163–187.

[36] Aiko, M. Hormann. (1962) “Programs for machine learning Part I.” Information and Control, 5(4), 347-367.

[37] Breiman, L. (2001) "Random forests." Machine learning, 45(1), 5-32.

[38] Pourret, Olivier, Naïm, Patrick, and Marcot, Bruce. (2008) “Bayesian Networks: A practical guide to applications.” John Wiley & Sons.