Classification Using Various Machine Learning Methods and Combinations of Key-Phrases and Visual Features

Yaakov HaCohen-Kerner[1], Asaf Sabag[1], Dimitris Liparas[2], Anastasia Moumtzidou[2], Stefanos Vrochidis[2], Ioannis Kompatsiaris[2]

[1]Dept. of Computer Science, Jerusalem College of Technology - Lev Academic Center, 9116001 Jerusalem, Israel
kerner@jct.ac.il, sabag.asaf.1@gmail.com
[2]Information Technologies Institute, Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece
{dliparas, moumtzid, stefanos, ikom}@iti.gr

Abstract. In this paper, we present a comparative study of news documents classification using various supervised machine learning methods and different combinations of key-phrases (word N-grams extracted from text) and visual features (extracted from a representative image from each document). The application domain is news documents written in English that belong to four categories: Health, Lifestyle-Leisure, Nature-Environment and Politics. The use of the N-gram textual feature set alone led to an accuracy result of 81.0%, which is much better than the corresponding accuracy result (58.4%) obtained through the use of the visual feature set alone. A competition between three classification methods, a feature selection method, and parameter tuning led to improved accuracy (86.7%), achieved by the Random Forests method.

Keywords: Document classification, Supervised learning, Feature selection, Key-phrases, N-gram features, Visual features

1 Introduction

During the last years, news agencies and newspapers face the challenge of automatically classifying news documents into a set of categories. This challenge becomes even more attractive when the documents contain not only text but also images. One such typical news document is depicted in Fig. 1. Moreover, in light of the explosion in the number of available news documents, the issue of fast and error-free classification of such documents is becoming more critical.

Classification using supervised learning is a task that is supervised by a set of examples with class assignments and the goal is to assign documents to one or more predefined categories [1]. Many supervised machine learning (ML) methods have been applied to document classification. The classification models are automatically built from annotated corpora. Comprehensive overviews of classification are given by [2-4].

Although many news documents include images in addition to text, most of the classification approaches make use of only textual data, in order to build the models. Therefore, it is interesting to perform a comparative study of news documents classification using different ML methods and different combinations of textual and visual feature

sets, in order to see whether the addition of the visual features can improve the classification performance.

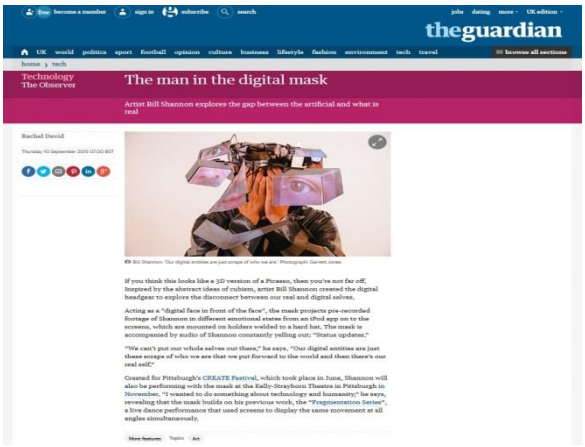

Fig. 1. Web-based news document from The Guardian entitled: The man in the digital mask[1]

In this paper, we explore domain-based classification of news documents using three general types of features: textual N-gram features, visual features and a combination of the above. The classification experiments are performed using three different supervised ML methods, namely J48, Random Forests (RF) and Sequential Minimal Optimization (SMO).

The rest of the paper is structured as follows: Section 2 provides the relevant background concerning document classification. Section 3 describes the textual and visual feature extraction procedures. Section 4 presents the involved classification methods, the experimental results and their analysis. Finally, Section 5 summarizes the main findings, concludes and suggests future directions.

2 Document Classification

Current-day document classification presents challenges due to the large number of training documents, the large number of available features and their dependencies. The document classification task is one of the most fundamental tasks in data mining and ML literature [5]. Document classification has been successfully applied to many fields such as document filtering, information extraction and text mining [6-8].

Document classification can be divided into two main types: according to categories (e.g., fields and topics) and according to stylistic classification. Document classification according to categories is usually based on content word or character N-grams. Some examples of document classification according to categories are the following: detection of author profiles for authorship attribution [8], detection of virus programs [9] and phrase and topic discovery [10].

In contrast, stylistic classification utilizes various linguistic features, e.g., function words, orthographic features, parts of speech (PoS) features, topographic features and

[1] http://www.theguardian.com/technology/2015/sep/10/the-man-in-the-digital-mask-bill-shannon

vocabulary richness features. Examples of document classification according to stylistic classification are: blog classification [11], computer science conference classification [12], ethnicity/time/place classification [13-14], and sentiment classification [15].

The majority of document classification-related studies consider only textual features. For instance, [16] focus on political news by tracking blogs and the articles they cite, tagging each article with the number of blogs citing it. They use a maximum entropy classifier based on unigram features. [17] develop various criteria to predict the popularity of news on Twitter and indicate that traditionally prominent news sources differ from news sources that are popular in social media platforms. Moreover, [18] present an approach for identifying and classifying contents of interest related to geographic communities from news articles streams. Their approach contains two stages: (1) Filtering out contents irrelevant to communities, and (2) Classifying the remaining relevant news articles by means of a Bayesian text classifier. Finally, [19] present learning of sentiment-specific word embedding for twitter sentiment classification. They learn embedding for unigrams, bigrams and trigrams separately using three developed neural networks.

Some examples of document classification studies that use only visual features are the following: [20] classify document pages using various visual features that express "visual similarity" of layout structure, e.g. percentages of text and non-text (graphics, images, tables, and rulings) content regions, column structures, relative point sizes of fonts, density of content area and statistics of features from connected components. They implement their classification scheme using decision tree classifiers and self-organizing maps. [21] explore image clustering as a basis for constructing visual words for representing documents. They apply the bag-of-words representation and standard classification methods to train an image-based classifier. Their main contribution is the exploration of a new space of features, based purely on the clustering of subfigures for document classification.

The task of classifying documents that contain both textual and visual features is a relatively new and interesting challenge. In this context, [22] explore the classification of news articles using both textual and visual features. By using only N-gram textual features they achieve a much better accuracy result than using only visual features. The use of both N-gram textual features and visual features leads to slightly improved accuracy results. Furthermore, [23] classify document images by combining 1000 textual features extracted with the Bag of Words (BoW) technique and 1000 visual features extracted with the Bag of Visual Words (BoVW) technique. Experiments conducted on an industrial document image database reveal that the proposed late fusion scheme significantly improves the classification performance.

3 Feature Extraction from News Documents

In this study, we assume that each news document has two main components: (a) the textual information, and (b) the image(s). Firstly, we extract continuous word N-grams (excluding stopwords) from the textual description of the given document. Secondly, we extract low-level visual features from the biggest image of the document, which is assumed to be the representative one.

3.1 Extraction of Key-Phrases

For the extraction of the textual features from a corpus of news web documents, the following procedure is applied:

1. All appearances of 421stopwords for general texts in English are deleted [24].
2. All possible continuous N-gram words (N = 1, 2, 3, 4) from the examined corpus are created, provided that the all the words in a certain N-gram are in the same sentence.
3. The frequency of each N-gram feature in the corpus is counted.
4. The frequencies of the unigram, bigram, trigram and fourgram (each group alone) features are sorted in descending order.
5. To avoid unnecessarily large number of N-grams, only a subset of the most frequent features from each group is selected. More specifically, in our study, 624 of the most frequent N-gram features are selected as follows: a) 500 most frequent unigrams; b) 100 most frequent bigrams; c) 20 most frequent trigrams and d) 4 most frequent fourgrams. The motivation for these numbers is as follows: The larger the value of N is, the smaller the number of relatively frequent N-grams in the corpus is. According to the abovementioned frequencies of the N-grams, the reduction factor was determined to be 5.

3.2 Extraction of Visual Features

The low-level visual feature that was used for capturing the characteristics of images is the RGB-SIFT visual descriptor [25], which is an extension of SIFT. In general, SIFT descriptors belong to the category of local descriptors that represent local salient points and thus capture the characteristics of the interest points (or keypoint) of images. For each keypoint only the pixel intensity of it is considered, while the color information is dropped. On the other hand, RGB-SIFT considers not only the pixel intensity, but also the color itself in the three channels Red, Green, Blue for each interest point. Thus, it captures more information and is able to better represent the image, compared to SIFT. However, when local descriptors are employed and given that the whole procedure is arduous, a visual word assignment step is applied after the feature extraction step. Specifically, k-means clustering is applied to the produced features vectors, in order to acquire the visual vocabulary. Finally, VLAD encoding is realized for representing images [26]. As a result, a descriptor is produced that gives an overall impression of the visual data. In this case, the dimensionality of the visual feature set is 4000.

4 Application of ML Methods for Classification of News Documents

Three supervised ML methods have been selected for the experiments in this study: J48, Random Forests (RF) and Sequential Minimal Optimization (SMO). Below, a short description of the methods is provided.

J48 is an improved variant of the C4.5 decision tree ML method [27], which is implemented in the WEKA ML platform [28]. J48 generates pruned or unpruned C4.5 decision trees. At each step, the most predictive attribute is determined and a node is split based on this attribute. J48 attempts to account for noise and missing data. It also deals with numeric attributes by determining where thresholds for decision splits should be placed.

Random Forests (RF) is an ensemble learning method for classification and regression [29]. The basic concept of RF is the construction of a set of decision trees. Moreover, two sources of randomness are employed in the operational procedures of RF: (1) Each decision tree is grown on a different bootstrap sample drawn randomly from the training data. (2) At each node split during the construction of a decision tree, a random subset of *m* variables is selected from the original variable set and the best split based on these *m* variables is used.

Sequential Minimal Optimization (SMO) [30-31] is an algorithm for solving the optimization problem that occurs during the training of Support Vector Machines (SVM) [32]. SMO divides this problem into a series of smallest possible sub-problems, which are then resolved analytically.

5 Examined Corpus, Experimental Setup and Results

5.1 Examined corpus

The application domain is news documents written in English that belong to four categories: Health, Lifestyle-Leisure, Nature-Environment and Politics. The news documents were downloaded from a large number of news web-sites (http://www.washingtonpost.com/, http://www.huffingtonpost.com/, etc.) and were annotated manually. The 1237 documents of the corpus contain around one million words and around 6.2 million characters. Table 1 presents some general information about the dataset and its domains including number of documents, number of words, number of characters, average number of words per document and average number of characters per document.

Table 1. General information about the dataset

Domain	# of documents	# of words	# of characters	Avg. # of words per document	Avg. # of characters per document
Health	187	130,157	795,435	696	4253.7
Lifestyle-Leisure	326	297,492	1,712,799	912.6	5254
Nature-Environment	447	250,859	1,543,137	561.2	3452.2
Politics	277	352,177	2,145,257	1271.4	7744.6
Total	1237	1,030,685	6,196,628	833.2	5009.4

5.2 Experimental Setup

The three supervised ML methods were applied using the WEKA platform, along with their default parameter values, which are described below:

- J48: minNumObj = 2 (the minimum number of instances per leaf), confidenceFactor = 0.25 (the confidence factor used for pruning) and seed = 1 (the value used for randomizing the data when reduced-error pruning is used).
- RF: numTrees = 100 (the number of trees to be generated), maxDepth = 0 (unlimited maximum depth of the trees) and seed =1 (the random number seed to be used).
- SMO: The kernel to use is the polynomial kernel (exponent = 1.0 and cachSize = 250007), c = 1.0 (the complexity parameter), toleranceParameter = 0.001 and randomSeed = 1 (the random number seed for the cross-validation).

It should be noted that for all classification experiments, a 10-fold cross-validation scheme was adopted. The measures used for evaluating the performance of the methods are the following: accuracy (test set), precision, recall and F-score for each category.

After determining the two ML methods that gave the best results, we performed additional experiments using only these methods. Non-relevant features were filtered out by means of a filter method for feature selection in WEKA called CfsSubsetEval (Correlation-based Feature Subset Selection) [33]. CfsSubsetEval evaluates a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class, while having low inter-correlation are preferred. All the optimized parameter values were obtained as follows: each parameter was tuned in a hill climbing fashion, changing one parameter at a time (manually) until the best value was obtained.

5.3 Results

Table 2 presents the accuracy results (%) for various combinations of feature sets using 3 ML methods (J48, RF and SMO) with their default parameter values. We relate to the accuracy that were achieved by all 624 N-gram features (81.0%) using the RF

method with default parameter values (the number of trees was set to 100) as the baseline, with which to compare the other results. The performance of the N-gram feature set is superior to that of the visual feature set (58.4%) not only in this study, but in a previous related study as well [22].

Table 2. Accuracy results (%) for various combinations of feature sets using 3 ML methods

Combinations of features	J48	RF	SMO
624 textual features	69.6	81.0	80.8
4000 visual features	55.1	58.4	57.3
59 textual features (Best First – 565 were filtered)	72.9	82.3	79.8
31 visual features (Best First – 3969 were filtered)	54.2	59.4	52.1
624 textual & 4000 visual features	69.7	68.5	81.2
59 textual & 4000 visual features	69.5	68.5	80.8
624 textual & 31 visual features	68.4	85.2	83.9
59 textual & 31 visual features	72.7	**85.9**	82.9

Some general conclusions that can be drawn from Table 2 are the following:

- The results presented in the first two rows of the table show that the basic textual feature set is superior to the corresponding visual feature set for all three ML methods. A possible explanation for this is the following: the textual features describe widespread information about the whole text, while the visual features describe information about only one representative image.
- The best accuracy result (85.9%) is achieved by RF using two reduced feature sets: 59 textual features and 31 visual features. On the other hand, SMO achieves only 82.9% for the same combination of features. Finally, the J48 method obtains the worst accuracy results for almost all the conducted experiments.
- There are two interesting opposing phenomena: SMO beats RF when using 4000 visual features and any set of textual features, while RF is better than SMO for all the other experiments. A possible explanation for this could be the fact that SVM (SMO in our case) is known to perform well in high-dimensional feature spaces [34].
- The improvement rate from the best unique set (59 textual features – 82.3%) to the best combination of feature sets (59 textual features & 31 visual features – 85.9%) is 3.6%.

As previously mentioned, we decided to perform further experiments using only the two best ML methods according to Table 2: RF and SMO. Tables 3 and 4 provide the optimized accuracy results for all combinations of feature sets using RF and SMO, respectively. In both Tables we also present the precision, recall and F-score results for each category and for all feature set combinations presented in Table 2.

The optimized results in Table 3 for the RF model have been achieved with 800 trees (experiments were conducted for various numbers of trees between 100 and 1200) and seed = 3. On the other hand, the optimized results in Table 4 for the SMO model have been achieved with Normalized Polynomial Kernel, toleranceParameter = 0.003, c = 9

and randomSeed = 1. For both ML methods, parameters that are not mentioned here are kept with their default values. Any change in their values did not improve the classification performance results.

Table 3. Classification results (%) for different combinations of feature sets (RF with the best parameter values)

Combinat-ions of features	Acc	Health			Lifestyle-Leisure			Nature-Environment			Politics		
		Pre	Rec	F-sco	Pre	Rec	F-sco	Pre	Rec	F-sco	Pre	Rec	F-sco
624 textual	83.1	93.3	59.4	72.5	79.2	92.0	85.1	81.3	91.7	86.2	88.1	74.7	80.8
4000 visual	58.3	61.1	57.2	59.1	53.6	66.9	59.5	58.3	63.5	60.8	66.7	40.4	50.3
59 textual	82.5	83.1	71.1	76.7	79.4	85.0	82.1	83.1	88.8	85.8	85.2	76.9	80.8
31 visual	59.7	59.8	57.2	58.5	57.1	67.8	62.0	59.2	66.0	62.4	30.2	66.5	41.5
624 textual & 4000 visual	67.8	77.4	56.7	65.4	60.2	77.0	67.6	68.5	75.8	72.0	76.1	51.6	61.5
59 textual & 4000 visual	69.0	83.2	61.0	70.4	58.4	81.3	67.9	72.5	75.4	73.9	76.2	49.8	60.2
624 textual & 31 visual	84.9	87.7	65.1	74.7	82.0	91.1	86.3	83.2	93.1	87.9	91.1	77.6	83.8
59 textual & 31 visual	**86.7**	89.2	75.4	81.7	85.7	90.2	87.9	85.8	91.7	88.6	88.0	81.9	84.8

Below we provide some general conclusions that can be drawn from Table 3:

- The best combination of features (59 textual features and 31 visual features) achieves an accuracy value of 86.7% and F-score values between 81.7% and 88.6% for the four categories.
- The precision values for the textual feature sets (624 and 59 textual features) are significantly higher than the corresponding recall values for the categories “Health” and “Politics”. Higher precision values indicate less false positives, which means that the RF method with the use of the textual features has a high ability to present relevant “Health” and “Politics” documents. Similar observations can be made for the feature combinations that include both textual and visual features, probably due to the decisive impact of the textual features.
- On the other hand, the recall values are significantly higher than the corresponding precision values for the categories “Lifestyle-Leisure” and “Nature-Environment”. Higher recall values indicate less false negatives. This means that the RF method has a high ability to classify “Lifestyle-Leisure” and “Nature-Environment” documents accurately.
- The worst results for all types of feature combinations are obtained for the categories “Health” and “Politics”. A possible explanation is that these categories contain a widespread variety of topics and therefore, their features vary strongly.

Table 4. Classification results (%) for different combinations of feature sets (SMO with the best parameter values)

Combina-tions of features	Acc	Health			Lifestyle-Leisure			Nature-Environment			Politics		
		Pre	Rec	F-s	Pre	Rec	F-s	Pre	Rec	F-s	Pre	Rec	F-s
624 textual	82.3	89.9	66.8	76.6	79.0	86.5	82.6	81.1	90.2	85.4	85.2	75.1	79.8
4000 visual	57.9	59.2	55.1	57.1	55.5	65.3	60.0	57.9	61.3	59.6	61.2	45.5	52.2
59 textual	76.6	72.8	71.7	72.2	71.0	75.8	73.3	80.5	79.4	80.0	79.9	76.2	78.0
31 visual	55.8	52.2	50.3	51.2	52.2	58.6	55.2	57.9	65.1	61.3	60.6	41.2	49.0
624 textual & 4000 visual	71.2	72.4	60.4	65.9	68.9	77.6	73.0	70.9	78.5	74.5	74.9	59.2	66.1
59 textual & 4000 visual	54.7	51.9	42.8	46.9	50.8	65.0	57.0	54.6	64.0	58.9	69.7	35.7	47.2
624 textual & 31 visual	**85.2**	87.9	78.0	82.6	85.3	87.1	86.2	83.5	88.4	85.9	86.4	82.7	84.5
59 textual & 31 visual	83.9	86.5	75.4	80.6	81.4	88.7	84.9	86.3	85.7	86.0	81.8	81.2	81.5

Below we provide some general conclusions that can be drawn from Table 4:

- In contrast to the best RF combination of features (59 textual features and 31 visual features), the best feature combination for SMO uses 31 visual and all 624 textual features. This combination achieves an accuracy value of 85.2% and F-score values between 82.6% and 86.2% for the four categories.
- Similar to the RF results, the SMO precision values regarding the best unique textual feature set (624 textual) are significantly higher than the corresponding recall values for the categories "Health" and "Politics". On the other hand, for the majority of the feature sets, the recall values are significantly higher than the corresponding precision values for the categories "Lifestyle-Leisure" and "Nature-Environment". Finally, just like in the case of the RF model, the SMO worst results are obtained for the "Health" and "Politics" categories for almost all types of feature combinations.

6 Summary, Conclusions and Future Work

In this study, we present a comparative classification study of news documents using three popular ML methods (J48, RF, and SMO), and different combinations of key-phrases (word n-grams excluding stopwords) and visual features. This comparative study is in contrast to two previous studies [22-23] that also perform classification with both textual and visual features but using only one ML method, one combination of textual and visual features and no feature selection.

Using the N-gram textual feature set containing 624 features led to an accuracy result of 81.0%. This result was much better than the accuracy result (58.4%) obtained for the visual feature set containing 4000 low-level features. A possible explanation for this finding is that the textual features describe widespread information about the whole text, while the visual features describe information about only one representative image. The use of the best combination of feature sets (59 textual features and 31 visual features) and the best parameter values for the RF model (800 trees) resulted in an accuracy result of 86.7%. Regarding SMO, the use of the best combination of feature sets (624 textual features and 31 visual features) and the best parameter values led to a small accuracy improvement of 1.3% (from 83.9% to 85.2%).

Suggestions for future research are: (1) Define and implement additional types of features, such as function words, morphological features (e.g. nouns, verbs and adjectives), quantitative features (e.g. average number of letters per word, average number of words per sentence) and web-oriented features, (2) Define and implement high-level visual concepts, in order to employ them in the classification tasks and (3) Apply additional ML methods to larger datasets in the news documents area, as well as in other areas, using various combinations of textual and visual features.

Acknowledgments. This work was supported by MULTISENSOR project, partially funded by the European Commission, under the contract number FP7-610411. The authors would also like to thank Avi Rosenfeld, Maor Tzidkani and Daniel Nissim Cohen from the Jerusalem College of Technology, Lev Academic Center, for their assistance to the authors in providing the software tool to generate the textual features used in this research. The authors would also like to acknowledge the networking support by the COST Action IC1302: semantic KEYword-based Search on sTructured data sOurcEs (KEYSTONE) and the COST Action IC1307: The European Network on Integrating Vision and Language (iV&L Net).

References

1. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34 (1), pp. 1-47 (2002)
2. Ozgür, A.: Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization (Doctoral dissertation, Bogaziçi University) (2004)
3. Kotsiantis, S. B., Zaharakis, I., Pintelas, P.: Supervised Machine learning: A Review of Classification Techniques, Informatica 31, 249-268 (2007)
4. Aggarwal, C. C., Zhai, C.: Mining Text Data. Springer Science & Business Media (2012)
5. Pazienza, M. T. (Ed.): Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology (Vol. 1299), Springer (1997)
6. Sebastiani, F.: Text Categorization. In: Zanasi, Alessandro (Ed.), Text Mining and its Applications to Intelligence. CRM and Knowledge Management, WIT Press, Southampton, UK, pp. 109–129 (2005)
7. Kim, S. M., Hovy, E.: Automatic Identification of Pro and Con Reasons in Online Reviews. In Proceedings of the COLING/ACL on main conference poster sessions, pp. 483-490, Association for Computational Linguistics (2006)

8. Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In Proceedings of the conference pacific association for computational linguistics, PACLING, Vol. 3, pp. 255-264 (2003)
9. Reddy, D. K. S., Pujari, A. K.: N-gram Analysis for Computer Virus Detection, Journal in Computer Virology, 2(3), 231-239 (2006)
10. Wang, X., McCallum, A., Wei, X.: Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. Seventh IEEE International Conference on ICDM, pp. 697-702 (2007)
11. Ikeda, D., Takamura, H., Okumura, M.: Semi-Supervised Learning for Blog Classification, AAAI, pp. 1156-1161 (2008)
12. HaCohen-Kerner, Y., Rosenfeld, A., Tzidkani, M., Cohen, D. N.: Classifying Papers from Different Computer Science Conferences. In Proceedings of the 9th International Conference on Advanced Data Mining and Applications (ADMA), China, Part I, LNAI 8346, Springer, Heidelberg, pp. 529–541 (2013)
13. HaCohen-Kerner, Y., Beck, H., Yehudai, E., Mughaz, D.: Stylistic Feature Sets as Classifiers of Documents According to their Historical Period and Ethnic Origin. Applied Artificial Intelligence, 24(9), pp. 847-862 (2010)
14. HaCohen-Kerner, Y., Beck, H., Yehudai, E., Rosenstein, M., Mughaz, D.: Cuisine: Classification using Stylistic Feature Sets and/or Name-based Feature Sets. Journal of the American Society for Information Science and Technology, 61(8), pp. 1644-1657 (2010)
15. Kennedy, A., Inkpen, D.: Sentiment Classification of Movie Reviews using Contextual Valence Shifters. Computational intelligence, 22(2), 110-125 (2006)
16. Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., König, A. C.: BLEWS: Using Blogs to Provide Context for News Articles. In: Proceedings of the second international AAAI conference on weblogs and social media (ICWSM), Seattle, Washington, March 30 – April 2 (2008)
17. Bandari, R., Asur, S., Huberman, B. A.: The Pulse of News in Social Media: Forecasting Popularity. In Proceedings of the In: Proceedings of the sixth international AAAI conference on weblogs and social media (ICWSM) (Arxiv preprint arXiv), Dublin 4–7 June, vol. 1202, pp. 26–33 (2012)
18. Swezey, R. M. E., Sano, H., Shiramatsu, S., Ozono, T., Shintani, T.: Automatic Detection of News Articles of Interest to Regional Communities. International Journal of Computer Science and Network Security (IJCSNS), 12(6), pp. 99-106 (2012)
19. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning Sentiment-specific Word Embedding for Twitter Sentiment Classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1555-1565 (2014)
20. Shin, C., Doermann, D., Rosenfeld, A.: Classification of document pages using structure-based features. International Journal on Document Analysis and Recognition, 3(4), pp. 232-247 (2001)
21. Chen, N., Shatkay, H., Blostein, D.: Exploring a new space of features for document classification: figure clustering. In Proceedings of the 2006 conference of the Center for Advanced Studies on Collaborative research, pp. 35, IBM Corp (2006)
22. Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., Kompatsiaris, I.: News Articles Classification Using Random Forests and Weighted Multimodal Features. In Proceedings of the 3rd Open Interdisciplinary MUMIA Conference and 7th Information Retrieval Facility Conference (IRFC2014), LNCS 8849, pp. 63-75, Berlin: Springer-Verlag (2014)

23. Augereau, O., Journet, N., Vialard, A., Domenger, J. P.: Improving Classification of an Industrial Document Image Database by Combining Visual and Textual Features. In Document Analysis Systems (DAS), In Proceedings of the11th International Workshop on IAPR, pp. 314-318, IEEE (2014)
24. Fox, C.: A Stop List for General Text. ACM SIGIR Forum, 24 (1-2), pp. 19-35 (1989)
25. Van De Sande, K. E., Gevers, T., Snoek, C. G.: Evaluating Color Descriptors for Object and Scene Recognition, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(9), pp. 1582-1596 (2010)
26. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating Local Descriptors into a Compact Image Representation, In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, pp. 3304-3311 (2010)
27. Quinlan, J. R.: C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann (1993)
28. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA Data Mining Software: an Update. ACM SIGKDD Explorations Newsletter, 11(1), pp. 10–18 (2009)
29. Breiman, L.: Random Forests. Machine Learning, 45(1), pp. 5-32 (2001)
30. Platt J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schoelkopf, B., Burges, C. Smola, A. (eds.), Advances in Kernel Methods – Support Vector Learning. Cambridge MA: MIT Press, pp. 185-208 (1998)
31. Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., Murthy, K. R. K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 13 (3), pp. 637-649 (2001)
32. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning, 20, pp. 273-297 (1995)
33. Hall, M. A.: Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand (1998)
34. Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification. J. of Machine Learning Research, 3:1289–1305 (2003)