News articles classification using Random Forests and weighted multimodal features

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

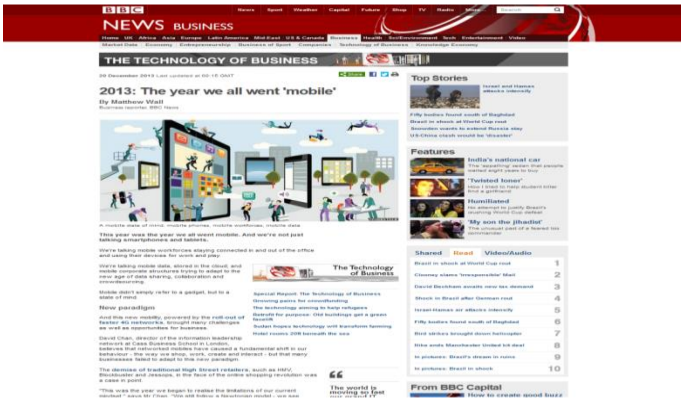
[1] Information Technologies Institute, Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece
{dliparas, moumtzid, stefanos, ikom}@iti.gr
[2] Dept. of Computer Science, Jerusalem College of Technology – Lev Academic Center,
21 Havaad Haleumi St., P.O.B. 16031, 9116001 Jerusalem, Israel
kerner@jct.ac.il

Abstract. This research investigates the problem of news articles classification. The classification is performed using N-gram textual features extracted from text and visual features generated from one representative image. The application domain is 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). Various classification experiments have been performed with the Random Forests machine learning method using N-gram textual features and visual features from a representative image. Using the N-gram textual features alone led to much better accuracy results (84.4%) than using the visual features alone (53%). However, the use of both N-gram textual features and visual features led to slightly better accuracy results (86.2%). The main contribution of this work is the introduction of a news article classification framework based on Random Forests and multimodal features (textual and visual), as well as the late fusion strategy that makes use of Random Forests operational capabilities.

Keywords: Document classification, Supervised learning, Multimodal, News articles, N-gram features, Random Forests, Visual features, Fusion

1 Introduction

The number of news articles published on various websites in general and news websites in particular had a dramatic increase over the last years. These articles contain multimodal information including textual and visual (image and video) descriptions. A visual example of such articles is illustrated in Fig. 1. Nowadays both journalists and media monitoring companies face the problem of mastering large amounts of articles in order to identify important topics and events all around the world. Therefore there is an important need for accurate and rapid clustering and classification of news articles into a set of categories in order to support journalism and media monitoring tasks. Despite of the multimodal nature of the articles posted in the web nowadays, most of the approaches consider only textual data in order to achieve classification (e.g. [1, 2]). Therefore there is an interesting challenge to investigate whether the combination of use of visual features in addition to the textual features will improve the classification accuracy.



20 December 2013 Last updated at 00:15 GMT

2013: The year we all went 'mobile'

By Matthew Wall
Business reporter, BBC News

A mobile state of mind: mobile phones, mobile workforces, mobile data

This year was the year we all went mobile. And we're not just talking smartphones and tablets.

We're talking mobile workforces staying connected in and out of the office and using their devices for work and play.

We're talking mobile data, stored in the cloud; and mobile corporate structures trying to adapt to the new age of data sharing, collaboration and crowdsourcing.

Mobile didn't simply refer to a gadget, but to a state of mind.

New paradigm

And this new mobility, powered by the roll-out of faster 4G networks, brought many challenges as well as opportunities for business.

David Chan, director of the information leadership network at Cass Business School in London, believes that networked mobiles have caused a fundamental shift in our behaviour - the way we shop, work, create and interact - but that many businesses failed to adapt to this new paradigm.

The demise of traditional High Street retailers, such as HMV, Blockbuster and Jessops, in the face of the online shopping revolution was a case in point.

"This was the year we began to realise the limitations of our current mindset," says Mr Chan. "We still follow a Newtonian model - we see

The Technology of Business

Special Report: The Technology of Business
Growing pains for crowdfunding
The technology aiming to help refugees
Retrofit for purpose: Old buildings get a green facelift
Sudan hopes technology will transform farming
Hotel rooms 20ft beneath the sea

The world is moving so fast

Israel and Hamas attacks intensify
Fifty bodies found south of Baghdad
Brazil in shock at World Cup rout
Snowden wants to extend Russia stay
US-China clash would be 'disaster'

India's national car
The 'appalling' sedan that people waited eight years to buy

'Twisted loner'
How I tried to help student killer find a girlfriend

Humiliated
No attempt to justify Brazil's crushing World Cup defeat

'My son the jihadist'
The unusual past of a feared Isis commander

Brazil in shock at World Cup rout	1
Clooney slams 'irresponsible' Mail	2
David Beckham awaits new tax demand	3
Shock in Brazil after German rout	4
Israel-Hamas air attacks intensify	5
Fifty bodies found south of Baghdad	6
Bird strikes brought down helicopter	7
Nike ends Manchester United kit deal	8
In pictures: Brazil's dream in ruins	9
In pictures: Brazil in shock	10

How to create good buzz

Fig. 1. Web-based news article from BBC entitled: 2013: The year we all went “mobile”[1]

News article classification is considered a Document classification (DC) problem. DC means labeling a document with predefined categories. This can be achieved as the supervised learning task of assigning documents to one or more predefined categories [3]. Using machine learning (ML), the goal is to learn classifiers from examples which perform the category classifications automatically. DC is applied in many tasks, such as: clustering, document indexing, document filtering, information retrieval, information extraction and word sense disambiguation. Current-day DC for news articles poses several research challenges, due to the large number of multimodal features present in the document set and their dependencies.

In this research, we investigate the task of category-based classification of news articles using a combination of visual features and textual N-gram features. The textual features are extracted from the textual part of the news article, while the visual features are generated from the biggest image of the article. Specifically we “learn” two Random Forest classifiers with textual and visual features and their results are fused using a late fusion strategy. The main contribution of this work is the usage of visual features in news article classification in order to leverage the text-based results, as well as the late fusion strategy that makes use of Random Forests’ operational capabilities (i.e. out-of-bag (OOB) error estimate and proximity ratios).

The rest of this paper is organized as follows: Section 2 presents the theoretical background and related work. Section 3 describes the textual and visual feature extraction procedure, while section 4 introduces the proposed classification framework. Section 5 presents the results of the experiments and section 6 concludes the paper.

[1] http://www.bbc.com/news/business-25445906

2 Related work and theoretical background

Since in this study we are dealing with supervised machine learning in Document Classification (DC) in general and with news articles classification in particular, we report previous work related to these two fields. Furthermore, since Random Forests (RF) is the machine learning method we use for our proposed classification framework, we provide the theoretical background and related work for this method.

2.1 Document and news articles classification

Several machine learning algorithms have been applied to DC, e.g.: Naïve Bayes [1], Rocchio [2], Logistic regression and Log-linear models [4], SVM [5] and Random Forests [6]. All these studies make use of textual features. The Reuters-21578 dataset, which contains news articles, was for a long time the standard resource dataset for text classification. Various text classification studies on the Reuters-21578 dataset make use of N-grams (N>1) in addition to unigrams. For example, [7] investigate the usefulness of bigrams for document indexing in text categorization (TC), independently of any specific learning algorithm. Their experiments reveal that if the feature evaluation function being used gives rise to a too high bigram penetration level, effectiveness may decrease. This finding is probably due to the elimination of informative unigrams on the part of bigrams that partly duplicate the information carried by existing unigrams. Moreover, [8] investigate text classification using semi-supervised ML methods for unlabeled documents. They apply two semi-supervised algorithms on several text datasets. Their results do not indicate improvement by combining unigrams and bigrams.

Some DC-related studies deal specifically with documents in web page format. For instance, [9] employ Neural Networks and Principal Component Analysis for web page feature selection and classification. Furthermore, in [10] a Random Forests classifier is employed for multi-category web page classification. Again, it is important to mention that the aforementioned studies deal with textual features, while our approach leverages the performance of textual features by using visual features from representative images.

In addition to the aforementioned works, several research attempts dealt specifically with news article classification. In this context, [11] have applied a maximum entropy classifier on unigram features to detect emotional charge around links to news articles in posts from political weblogs. In another work, [12] apply both regression and classification algorithms that make use of social interaction features to the task of predicting the popularity of news items on the social web. They demonstrate that one of the most important predictors of popularity is the source of the article. Finally, [13] utilize news articles streams and Bayesian text classification in order to classify contents of interest related to geographic communities. These works rely mainly upon textual data for classification, while the proposed approach leverages the results of text-based classification by considering visual features.

2.2 Random Forests

Random Forests (RF) is an ensemble learning method for classification and regression [14]. The basic notion of the methodology is the construction of a group of decision trees. RF employs two sources of randomness in its operational procedures:

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.

For an unknown case, the predictions of the trees that are constructed by the RF are aggregated (majority voting for classification / averaging for regression). For a RF consisting of *N* trees, the following equation is used for predicting the class label *l* of a case *y* through majority voting:

$$l(y) = argmax_c(\sum_{n=1}^{N} I_{h_n(y)=c}) \qquad (1)$$

where *I* the indicator function and h_n the *nth* tree of the RF.

RF has an internal mechanism that provides an estimation of its generalization error, called out-of-bag (OOB) error estimate. For the construction of each tree, only 2/3 of the original data's cases are used in that particular bootstrap sample. The rest 1/3 of the instances (OOB data) are classified by the constructed tree and therefore, used for testing its performance. The OOB error estimate is the averaged prediction error for each training case y, using only the trees that do not include y in their bootstrap sample. Furthermore, when a RF is constructed, all the training cases are put down each tree and a proximity matrix between the cases is computed, based on whether pairs of cases end up in the same terminal node of a tree.

Successful applications of RF to an extensive range of disciplines (apart from DC and web page classification) can be found in the relevant literature. Among others, image classification [15], network intrusion detection [16] and neuroimaging [17] can be listed.

In this study, we investigate the application of RF for news articles classification. Although the RF have been successfully applied to several classification problems (as discussed above), to the best of our knowledge they haven't been applied to news article classification problems. Moreover, an important motivation for using RF was the application of late fusion strategies based on the RF operational capabilities.

3 Feature Extraction from News Articles Documents

In this work, we assume that the article has two main parts: a) the textual description and b) the images (e.g. Fig. 1). First we extract N-grams (globally and not per category) from the textual description. N-grams were chosen as our textual features because they were found as relatively easy to compute and effective for various classification tasks (e.g. [7, 8, 18]). Then, we select the biggest image of the article and extract visual features. In this case we assume that the biggest image was the representative one.

3.1 N-gram Textual Features

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

1. All appearances of 421 stopwords for general texts in English are deleted [19].
2. All possible continuous N-gram words (for N =1, 2, 3, 4) 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 corpora is counted.
4. The unigram, bigram, trigram and fourgram (each group alone) features are sorted in descending order.

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 195 of the most frequent N-gram features are selected as follows: a) 100 most frequent unigrams; b) 50 most frequent bigrams; c) 30 most frequent trigrams; d) 15 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. The reduction factor was determined to be approximately 2.

3.2 Visual Features

The low-level visual features that are extracted in order to capture the characteristics of images are the MPEG-7 visual descriptors. The MPEG-7 standard specifies a set of descriptors, each defining the syntax and the semantics of an elementary visual low-level feature. Each descriptor aims at capturing different aspects of human perception (i.e., color, texture and shape). In this work, five MPEG-7 visual descriptors capturing color and texture aspects of human perception are extracted [20]:

1. **Color Layout Descriptor**: captures the spatial distribution of color or an arbitrary-shaped region.
2. **Color Structure Descriptor**: is based on color histograms, but aims at identifying localized color distributions.
3. **Scalable Color Descriptor**: is a Haar-transform based encoding scheme that measures color distribution over an entire image.
4. **Edge Histogram Descriptor**: captures the spatial distribution of edges and it involves division of image into 16 non-overlapping blocks. Edge information is then calculated for each block.
5. **Homogenous Texture Descriptor**: is based on a filter bank approach employing scale and orientation sensitive filters.

Then, we apply an early fusion approach, which involves the concatenation of all the aforementioned descriptors into a single feature vector. In this study, 320 visual features are extracted in total. The number of features/dimensions that are created from each descriptor are the following: a) Color Layout Descriptor: 18 features/dimensions; b) Color Structure Descriptor: 32 features/dimensions; c) Scalable Color Descriptor: 128 features/dimensions; d) Edge Histogram Descriptor: 80 features/dimensions; e) Homogeneous Texture Descriptor: 62 features/dimensions.

4 Proposed Classification Framework

The flowchart of the proposed classification framework (training phase) is depicted in Fig. 2. Next, the different steps of the framework are described in detail.

First, all the necessary **data is collected** in the form of text from news article web pages, as well as images associated to each web page. In the following step, the procedures described in Section 3 are applied to the raw data (**Parsing / Extraction**). In this way, the **visual and textual features are generated**. We note that given the fact that the majority of the web pages contain several images including banners and advertisement logos, it was decided to keep only the biggest image of each site which would most probably be the main image of the article. One other important thing to note is that in this study the features of each modality are treated independently. Hence, two different feature vectors (one for each modality) are formulated.

In the training phase, the feature vectors from each modality are used as input for the **construction of a RF**. From the two constructed RFs (one for the textual and one for the visual features), we **compute the weights for each modality**, in order to **apply a late fusion strategy** and **formulate the final RF predictions**. In this study, two different approaches for the computation of the modality weights are followed:

1. From the OOB error estimate of each modality's RF, the corresponding OOB accuracy values are computed. These values are computed for each class separately. Then, the values are normalized (by dividing them by their sum) and serve as weights for the two modalities.
2. For the second weighting strategy, the same procedure as in 1. is applied. However, instead of employing the OOB accuracy values from each RF, the ratio values between the inner-class and the intra-class proximities (for each class) are used [21]. First, for each RF the proximity matrix between all pairs of data cases $P=\{p_{ij}, i,j=1, \ldots, w\}$ (w=number of data cases) is constructed and then, the aforementioned ratio values are computed as in the following equation:

$$R = \frac{P_{inner}}{P_{intra}} \tag{2}$$

where

$$P_{inner} = \sum_{i,j=1}^{w} p_{ij}\ (if\ l_i = l_j) \tag{3}$$

$$P_{intra} = \sum_{i,j=1}^{w} p_{ij}\ (if\ l_i \neq l_j) \tag{4}$$

and l_i, l_j the class labels of cases i and j, respectively.

During the testing phase, when the RF predicts a case, it outputs probability estimates per class for that case. The probability outputs P_t and P_v from the textual and visual RFs respectively are multiplied by their corresponding modality weights W_t and W_v and summed to produce the final RF predictions as in the following equation:

$$P_{fused} = W_t P_t + W_v P_v \tag{5}$$

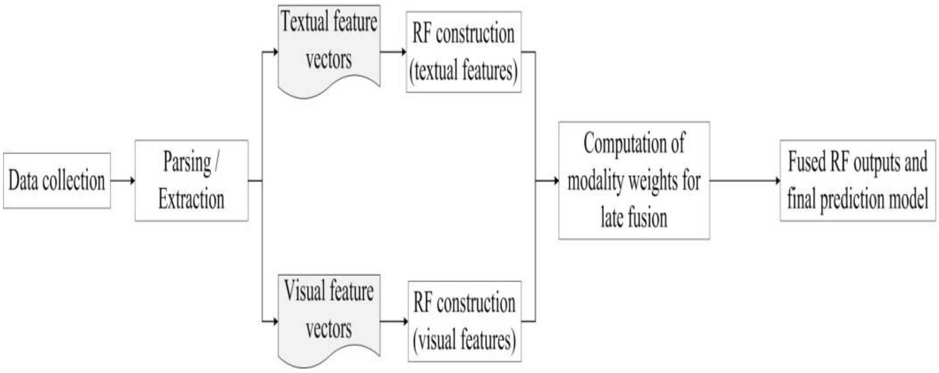

Data collection
Parsing / Extraction
Textual feature vectors
Visual feature vectors
RF construction (textual features)
RF construction (visual features)
Computation of modality weights for late fusion
Fused RF outputs and final prediction model

Fig. 2. Flowchart of proposed classification framework

5 Experimental Results

5.1 Dataset Description

The experiments are realized on a dataset that contains web pages from three well known News Web Sites, namely BBC, The Guardian and Reuter. Overall, 651, 556 and 360 web pages have been retrieved from each site, respectively. At this point it should be noted that the manual annotation of the web pages was necessary, regardless of the fact that in the three News Web Sites descriptions about the topic of each web page are provided, since in many cases the descriptions are inconsistent with the content of the web pages. The manual annotation was realized for a subset of the topics recognized by the IPTC news codes taxonomy[2], which is the global standards body of the news media. Specifically, we selected the most important topics with the guidance of media monitoring experts and journalists. Table 1 contains a detailed description of the final dataset[3] and the topics considered.

Table 1. Details of dataset

Topics / **News Sites**	Business, finance	Lifestyle, leisure	Science, technology	Sports	Num. of documents per site
BBC	102	68	75	202	447
The Guardian	67	59	116	96	338
Reuter	165	7	29	57	258
Num. of documents per topic	334	134	220	355	1043

[2] http://www.iptc.org/site/Home/

[3] The dataset is publicly available at: http://mklab.iti.gr/files/ArticlesNewsSitesData.7z

5.2 Experimental Setup

We randomly split our dataset into training and test sets in order to conduct the experiments. Approximately 2/3 of the cases are kept for training purposes, whereas the rest (1/3) are used as test set, in order to estimate the classification scheme's performance.

As for the RF parameters that we use in the experiments, we opted to apply the following setting: We set the number of trees for the construction of each RF based on the OOB error estimate. After several experiments with different numbers of trees, we noticed that the OOB error estimate was stabilized after using 1000 trees and no longer improved. Hence, the number of trees is set to *N*=1000. For each node split during the growing of a tree, the number of the subset of variables used to determine the best split is set to $m = \sqrt{\mathrm{k}}$ (according to [14]), where k is the total number of features of the dataset.

Finally, for the evaluation of the performance of the proposed methodology, we compute the precision, recall and F-score measures for each category, along with their corresponding macro-averaged values, as well as the accuracy on the entire test set (all categories included).

5.3 Results

The test set results from the application of RF to each modality separately are summarized in Table 2. We are mainly interested in the values of F-score, since it considers both precision and recall. We notice that the textual modality outperforms the visual in all measures, both regarding each topic and the macro-averaged scores. This indicates that textual data is a more reliable and solid source of information, in comparison to the visual data. More specifically:

- The RF trained with the textual data achieves a macro-averaged F-score value of 83.2%, compared to 45.5% for the visual modality
- The accuracy for the textual modality RF is 84.4%, while the visual modality RF achieves only 53%
- The worst results for the visual data RF are attained for the topics "Lifestyle-Leisure" (recall 12% and F-score 20.7%) and "Science-Technology" (precision 45.3%, recall 38.7% and F-score 41.7%). However, the results regarding the topic "Sports" are considered satisfactory. A possible explanation for this is the fact that the images from the "Lifestyle-Leisure" web pages depict diverse topics and therefore their visual appearance strongly varies. On the other hand, the images regarding the topic "Sports" contain rather specific information such as football stadiums (a characteristic example is depicted in Fig. 3).

Table 2. Test set results from the application of RF to each modality

Modality / Topics	Textual			Visual		
	Prec.	**Rec.**	**F-score**	**Prec.**	**Rec.**	**F-score**
Business-Finance	80.0%	87.3%	83.5%	56.3%	57.3%	56.8%
Lifestyle-Leisure	86.7%	78.0%	82.1%	75%	12%	20.7%
Science-Technology	79.1%	70.7%	74.6%	45.3%	38.7%	41.7%
Sports	91.3%	93.8%	92.5%	52.8%	76.8%	62.6%
Macro-average	**84.3%**	**82.5%**	**83.2%**	**57.4%**	**46.2%**	**45.5%**
Accuracy	**84.4%**			**53.0%**		

Fig. 3. Characteristic image from a "Sports" web page (left)[4], along with an image regarding a web page from the "Lifestyle-Leisure" topic (right)[5]

In Table 3 we provide the test set results from the application of the late fusion strategy to RF, using the two different weighting methods described in Section 4 (OOB error/ Proximity ratio). The weighting method regarding the proximity ratio yields better performance results than the corresponding method for the OOB error. More specifically:

- The accuracy of Textual + Visual (Proximity ratio) is slightly better than the corresponding accuracy of Textual + Visual (OOB error) (86.2% compared to 85.9%)
- The two weighted RFs achieve almost equal macro-averaged precision values (86.8% for Proximity ratio and 86.9% for OOB error), while regarding the macro-averaged recall and F-score results, Textual + Visual (Proximity

[4] http://www.bbc.com/sport/0/football/27897075-"_75602744_ochoa.jpg"

[5] http://www.bbc.com/travel/feature/20140710-living-in-istanbul-"p022ktsw.jpg"

ratio) is better (84.2% to 82.9% for the macro-averaged recall and 85.3% to 84.3% for the macro-averaged F-score)

For comparison purposes, we also constructed a fused RF model, where equal weights were assigned to each modality. We notice that after following this weighting approach (i.e. with equal weights), the performance of RF diminishes in all aspects. The superiority of the weighting strategy based on the proximity ratio of each topic is also evident in Fig. 4, where the macro-averaged F-score values of all 5 RF models constructed in this study are sorted in ascending order. We observe that Textual + Visual (Proximity ratio) is the best performing model among all cases.

Table 3. Test set results after the late fusion of RF regarding three different weighting schemes

Weighting method / **Topics**	**Textual + Visual** (Weighting based on OOB error per topic)			**Textual + Visual** (Weighting based on proximity ratio per topic)			**Textual + Visual** (Equal weights per topic)		
	Prec.	**Rec.**	**F-score**	**Prec.**	**Rec.**	**F-score**	**Prec.**	**Rec.**	**F-score**
Business-Finance	80.3%	92.7%	86.1%	82.4%	89.1%	85.6%	71.1%	91.8%	80.1%
Lifestyle-Leisure	92.5%	74.0%	82.2%	92.9%	78.0%	84.8%	91.4%	64%	75.3%
Science-Technology	83.9%	69.3%	75.9%	81.4%	76.0%	78.6%	83.1%	65.3%	73.1%
Sports	90.7%	95.5%	93%	90.5%	93.8%	92.1%	87.4%	86.6%	87%
Macro-average	**86.9%**	**82.9%**	**84.3%**	**86.8%**	**84.2%**	**85.3%**	**83.3%**	**76.9%**	**78.9%**
Accuracy	**85.9%**			**86.2%**			**80.4%**		

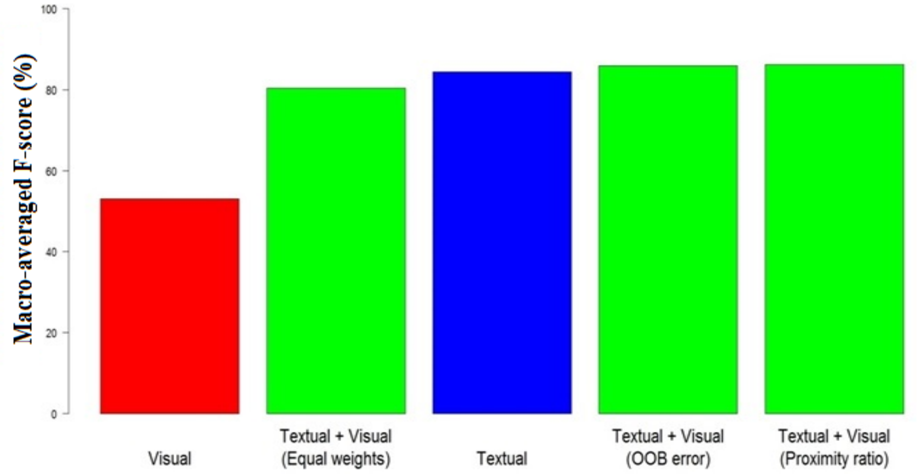


Fig. 4. Macro-averaged F-score values for all RF models sorted in ascending order

6 Summary, Conclusions and Future Work

In this research, we investigate the use of N-gram textual and visual features for classification of news articles that fall into four categories (Business-Finance, Lifestyle-Leisure, Science-Technology, and Sports) downloaded from three news web-sites (BBC, Reuters, and TheGuardian).

Using the N-gram textual features alone led to much better accuracy results (84.4%) than using the visual features alone (53%). However, the use of both N-gram textual features and visual features (weighting based on proximity ratio per category) led to slightly better accuracy results (86.2%).

Future directions for research are: (1) Defining and applying additional various types of features such as: function words, key-phrases, morphological features (e.g.: nouns, verbs and adjectives), quantitative features (various averages such as average number of letters per a word, average number of words per a sentence) and syntactic features (frequencies and distribution of parts of speech tags, such as: noun, verb, adjective, adverb), (2) Applying various kinds of classification models based on textual and visual features for a larger number of documents that belong to more than four categories in the news articles area, as well as in other areas, applications and languages, (3) Selecting a representation for images based on visual concepts. In such a case, the approach could consider more than one image per article. Visual concepts could be extracted from each image and the average score for each visual concept could be calculated, in order to represent the article based on multiple images.

Acknowledgments. The authors would like to acknowledge networking support by the COST Action IC1002: Multilingual and multifaceted interactive information access (MUMIA) and the COST Action IC1307: The European Network on Integrating Vision and Language (iV&L Net). The authors would also like to thank Avi Rosenfeld, Maor Tzidkani, Asaf Sabag and Daniel Nisim Cohen from 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.

References

1. Schneider, K. M.: Techniques for improving the performance of naive Bayes for text classification. In Computational Linguistics and Intelligent Text Processing, pp. 682-693, Springer Berlin Heidelberg (2005)
2. Zeng, A., & Huang, Y.: A text classification algorithm based on rocchio and hierarchical clustering. In Advanced Intelligent Computing (pp. 432-439), Springer Berlin Heidelberg (2012)
3. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys, 34 (1), pp. 1-47 (2002)
4. Toutanova, K.: Competitive generative models with structure learning for NLP classification tasks. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, (pp. 576-584) (2006)
5. Ho, A. K. N., Ragot, N., Ramel, J. Y., Eglin, V., & Sidere, N.: Document Classification in a Non-stationary Environment: A One-Class SVM Approach. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, (ICDAR), (pp. 616-620) (2013)

6. Klassen, M., & Paturi, N.: Web document classification by keywords using random forests. In Networked Digital Technologies, pp. 256-261, Springer Berlin Heidelberg (2010)
7. Caropreso, M. F., Matwin, S., & Sebastiani, F.: Statistical phrases in automated text categorization. Centre National de la Recherche Scientifique, Paris, France (2000)
8. Braga, I., Monard, M., & Matsubara, E.: Combining unigrams and bigrams in semi-supervised text classification. In Proceedings of Progress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence (EPIA 2009), Aveiro (pp. 489-500), (2009)
9. Selamat, A., & Omatu, S.: Web page feature selection and classification using neural networks. Information Sciences, 158, pp. 69-88 (2004)
10. Aung, W. T., & Hla, K. H. M. S.: Random forest classifier for multi-category classification of web pages. In Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific, pp. 372-376, IEEE (2009)
11. Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., & König, A. C.: BLEWS: Using Blogs to Provide Context for News Articles. In ICWSM (2008)
12. Bandari, R., Asur, S., & Huberman, B. A.: The Pulse of News in Social Media: Forecasting Popularity. In ICWSM (2012)
13. Swezey, R. M., Sano, H., Shiramatsu, S., Ozono, T., & Shintani, T.: Automatic detection of news articles of interest to regional communities. IJCSNS, 12(6), 100 (2012)
14. Breiman, L.: Random Forests. In Machine Learning, 45(1), pp. 5-32 (2001)
15. Xu, B., Ye, Y., & Nie, L.: An improved random forest classifier for image classification. In Information and Automation (ICIA), 2012 International Conference on (pp. 795-800), IEEE (2012)
16. Li, W., & Meng, Y.: Improving the performance of neural networks with random forest in detecting network intrusions. In Advances in Neural Networks–ISNN 2013 (pp. 622-629), Springer Berlin Heidelberg (2013)
17. Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., & Rueckert, D.: Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage, 65, pp. 167-175 (2013)
18. HaCohen-Kerner, Y., Mughaz, D., Beck, H., & Yehudai, E.: 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), pp. 213-228 (2008)
19. Fox, C.: A stop list for general text. ACM SIGIR Forum, 24 (1-2), ACM (1989)
20. Sikora, T.: The MPEG-7 visual standard for content description-an overview. IEEE Transactions on Circuits and Systems for Video Technology, 11(6), pp. 696-702 (2001)
21. Zhou, Q., Hong, W., Luo, L., & Yang, F.: Gene selection using random forest and proximity differences criterion on DNA microarray data. Journal of Convergence Information Technology, 5(6), pp. 161-170 (2010)