Hypatia Digital Library: A novel text classification approach for small text fragments

Ioannis Triantafyllou, Frosso Vorgia, Alexandros Koulouris
Department of Archival, Library & Information Studies, University of West Attica, Athens, Greece
triantafi@uniwa.gr [**ORCID:** 0000-0001-5273-0855], frossovorgia@gmail.com, akoul@uniwa.gr
[**ORCID:** 0000-0002-4011-2678]

Article history:
Received: May 2019
Received in revised form: July 2019
Accepted: Octomber 2019

DOI: *https://doi.org/10.26265/jiim.v4i2.4420*

Abstract:

Purpose - *The purpose of this paper is to further investigate prior work of the authors in text classification in Hypatia, the digital library of University of Western Attica. The main objective is to provide an accurate automated classification tool as an alternative to manual assignments.*

Design/methodology/approach - *The crucial point in text classification is the selection of the most important term-words for document representation. The specific document collection consists of 718 abstracts in Medicine, Tourism and Food Technology. Two weighting methods were investigated: classic TF.IDF and DEVMAX.DF. The last one was proposed by the authors as a more accurate term-word selection tool for smaller text fragments. Classification was conducted by applying 14 classifiers available on WEKA.*

Findings - *Classification process yielded an excellent ~97% precision score and DEVMAX.DF proved to perform better than classic TF.IDF.*

Index Terms — Digital libraries, Statistical natural language processing, Text classification, WEKA, Word stemming.

I. Introduction

Subject classification in libraries is conducted manually with the use of classification systems, subject headings, thesauri and ontologies. This time-consuming process has been adopted for the digital libraries as well **[1].** However, considering the immense and continuous creation of digital objects, a new method of fast classification is required **[2]**.

The purpose of the present work is to employ the text classification method in digital libraries as an alternative solution to the aforementioned problem. Text classification is applied on small text fragments such as the abstracts of the digital objects. Abstracts are considered to be the best option to experiment with as they might be the only available texts which represent the content of resources, since full text is not always available due to copyright constraints. The abstracts are mainly extracted from Hypatia, the digital library of University of Western Attica (former Technological Educational Institute (T.E.I.) of Athens).

In a previous research **[3]** we applied and made measurements of abstract representation by word weighting with TF.IDF. Nevertheless, the results were unsatisfactory and this created the need to reexamine this work with the use of a new weighting method called DEVMAX.DF, which is introduced here. In the final phase, - classification algorithms provided by the open source software WEKA are used **[4, 5]**.

II. Related work

Text classification/categorization (TC) is the task of classifying texts in predefined classes **[6]**. So far TC has been utilized in a machine learning approach, conducted with the use of classifiers (algorithms). The most extensively used ones for TC are Naïve Bayes and Naïve Bayes Multinomial **[7].** However, there are more classifiers, such as Support Vector Machines (SVM), Multilayer Perceptron, IBk, Decision Table, Random Forest etc. which can be exploited **[8]**. Especially in the environment of a digital library which hosts entire collections of documents, scientific papers, dissertations, datasets, images and sounds, TC can be advantageous for browsing and retrieval **[9]**.

Classification techniques have achieved encouraging outcomes in many applications regarding small to medium text fragments, like those already provided by digital libraries (abstracts). One application is the common and ever evolving problem of spam emails. The solution is e-mail filtering to prevent phishing and labeling as spam or ham **[9, 10, 11, 12]**. Likewise, in the field of telecommunications, the approach of TC has been used for SMS labeling similarly **[13, 14, 15]**.

Additionally, microblogging services are valuables sources of small texts. In Twitter, for example, a vast number of Tweets are produced every day. Focused analyses, such as Twitter trending toppings' classification **[16]**, sentiment analysis on financial related Tweets **[17]**, suicidal expressions [18], and recognition of pornographic material **[19]** have produced positive results. In addition, these techniques can overcome language barriers as they can be employed with English, Dutch, Indonesian, or even Chinese **[19, 20, 21]**.

III. Methodology

The initial idea was that TC would be applied on full texts, but inevitably, - some problems due to access limitations were arisen. Therefore, there was made an effort to collect

keywords in order to weigh the words in classes with TF.IDF. However, this approach would produce patently obvious results, so it was abandoned. Eventually, the research team adopted the methodology described in the sections below.

A. Data collection

- 718 abstracts were collected, considering that they are in Greek and already classified either in Medicine or Tourism or Food Technology, as these classes were the most p frequent). Although, Hypatia was the main source of abstracts, it was considered scientifically sound to extract data from more sources. Thus, the research team decided to derive abstracts from other digital libraries aiming to create a balanced corpus for the three classes. Analytically, abstracts were assembled from the following 9 Greek academic digital libraries and repositories.

- Hypatia - University of Western Attica (512),
- The digital repository of Agricultural University of Athens (AUA) (73),
- Eureka! - T.E.I. of Thessaloniki (47),
- Dioni - University of Piraeus (45),
- Psepheda - University of Macedonia (19),
- DSpace - National Technical University of Athens (11),
- Nemertes - University of Patras (9),
- E-Locus - University of Crete (1),
- Anaktisis - T.E.I. Institute of Western Macedonia (1).

However, each digital library applies different subject classification tools to assign the subject categories. In order to ensure uniformity and accordance in the dataset, Dewey Decimal Classification was used as a guide to include or discard the abstracts. The only exception was a set of 22 abstracts from the digital repository of Agricultural University of Athens. These concerned theses from the department of Science and Food Technology, which also included relevant words, so they were considered to have a connection to Food Technology.

The final text corpus consisted of 373 abstracts in Medicine, 223 in Tourism and 122 in Food Technology.

B. Text Handling and Word Stemming

Initially, a basic text pre-processing is necessary to minimize the noise. A system of natural language communication includes nouns, verbs, adverbs, conjunctions, etc. Not every part of speech has useful meaning. For example, the word "και" ("and" in English) has no special meaning, regardless of how many times it appears in a text. These kinds of words are called "stop words" and have to be removed **[22]**.

In addition, it is essential to stem the words of the texts. Greek is a highly inflected language, meaning that almost every word in a sentence has an affix. Stemming, or conflation, is the process of reducing the words to their stem by taking off the affixes **[23]**.

Basic text pre-processing is based on text handler **[8]**, a tool having the responsibility of transforming a text from abstracts into a form suitable for the manipulation required by the application:

- identification of textual units at the level of sentences by using trivial delimiters, such as spaces, stops, question marks, etc.
- identification of extra-linguistic elements, such as dates, abbreviations, acronyms, list enumerators, numbers, etc.

Subsequent to words' identification, the word stemming, or term conflation process is performed. During the latter, the system captures the morphological variations of terms located in the abstracts. Term spotting process is performed in two subsequent phases. The first phase aims at reducing the search space thus improving the performance of the system. During this phase, a small set of candidate similar words, based on statistical information, has been extracted and grouped together under a common representative term.

Consequently, during the second phase a more elaborate procedure occurs, where the system ranks the located terms and produces a complete term "short-list" for each candidate term of the input text. The score mechanism is based on the similarity estimator (Eq. 1), especially designed to assign higher scores to morphological variations of the same root form.

$$\text{Similarity}(W1, W2) = \frac{\text{Common Position Trigrams}(\text{Left}(W1, L), \text{Left}(W2, L))}{L} \quad (1)$$

$$\text{where } L = \frac{\text{Length}(W1) + \text{Length}(W2)}{2}, L \epsilon N$$

Efficient grouping of words in terms has been achieved with a similarity score of 66,6%.

C. Abstract Representation

Special consideration has been granted to the selection of the feature space, a crucial aspect in the performance of any text classification model. Any term-word within the abstracts corpus constitutes a candidate feature with the exception of functional words that are excluded based on stop-lists. Feature selection consists of reducing the vocabulary size of the training corpus by selecting term-words with the highest indicative efficiency over the class variable.

The TF.IDF metric **[23, 24]** is one classic approach to sort the candidates' term-words in a list by scoring their correlation importance to the class variable. In our case TF is the frequency of feature f within the corpus, and IDF is the logarithm of N/Nf, where N is the total number of abstracts and Nf is the number of abstracts containing the feature f. The selected features are the most dominant ones based on that score.

After experimenting with TF.IDF it was observed that a lot of irrelevant term-words, with appearance in all classes, were sorted highly in the importance list. Hence, there was made a decision to introduce a new metric which would promote the term-words appearing mainly in one or more classes but not entirely. The intention was to promote term-words that have the maximum deviation in appearances (in other words the minimum appearances) in other classes from the main (max) class, the class in which they mostly appear. In order to promote high appearance term-words the formula is further regulated with the logarithm of the DF,

the number of abstracts containing the term-word F. The metric with the proposed name DEVMAX.DF is described in the following equation (Eq. 2)."

$$DEVMAX.DF = \sqrt{\frac{\sum_{i=1}^{c}(DF_i/N_i - max)^2}{(c-1)*max^2}} * \log(DF), \quad where \max = maximum_{i=1}^{c} DF_i/N_i \tag{2}$$

DFi is the number of abstracts containing the term-word F in class i, Ni is the number of abstracts in class i and c is the number of classes. The comparison between the two methods is presented in Table 1, where the metric obviously has managed to promote more important term-words for the abstract representation; term-words which are related to one class mainly and therefore provide a good correlation importance for the class.

Table 1. First 10 selected term-words in both metrics and their appearances in the 3 classes.

DEVMAX.DF				**TF.IDF**			
TERM-WORD	*Medicine*	*Tourism*	*Food*	**TERM-WORD**	*Medicine*	*Tourism*	*Food*
ΤΟΥΡΙΣΜΟ *(TOURISM)*	0	187	0	ΤΟΥΡΙΣΜΟ *(TOURISM)*	0	187	0
ΝΟΣΗΛΕΥΤΗΚΑΝ *(HOSPITALISED)*	129	0	0	ΑΣΘΕΝΩΝ *(PATIENTS)*	194	2	9
ΝΟΣΟΚΟΜΕΙΟ *(HOSPITAL)*	101	0	0	ΝΟΣΗΛΕΥΤΗΚΑΝ *(HOSPITALISED)*	129	0	0
ΑΣΘΕΝΩΝ *(PATIENTS)*	194	2	9	ΥΓΕΙΑΣ *(HEALTH)*	147	5	14
ΦΡΟΝΤΙΔΑ *(CARE)*	70	0	0	ΠΑΙΔΙΑ *(CHILDREN)*	49	1	4
ΓΥΝΑΙΚΕΣ *(WOMEN)*	68	1	0	ΑΝΑΠΤΥΣΣΕΙ *(DEVELOPS)*	66	112	48
ΚΛΙΝΙΚΗ *(CLINIC)*	85	1	2	ΠΟΙΟΤΗΤΑΣ *(QUALITY)*	85	39	28
ΤΡΟΦΙΜΩΝ *(FOOD)*	2	1	56	ΜΕΘΟΔΟΥΣ *(METHODS)*	204	34	56
ΘΕΡΑΠΕΙΑΣ *(THERAPY)*	104	3	4	ΑΝΑΓΚΕΣ *(NEEDS)*	151	88	53
ΑΝΑΣΚΟΠΗΣΗ *(REVIEW)*	98	8	1	ΕΚΠΑΙΔΕΥΤΙΚΩΝ *(EDUCATIONAL)*	88	14	2

An additional important issue to consider is the frequency of a term-word when determining the abstract vector. There are cases where a term-word is more indicative to the relevance of the abstract when it appears several times. However, this is not always true since long abstracts usually introduce a lot of noise. The research team experimented with two alternatives concerning the strength of the selected features: the binary (boolean) appearance (0 or 1), and the actual value of the term frequency in the abstract. Thus, the experimental methods consist of four possible combinations based on two axes, the importance metric on which the selection of feature space is based and the strength of the representative feature: TF.IDF-bin, TF.IDF-tf, DEVMAX.DF-bin and DEVMAX.DF-tf.

D. Text Classification with WEKA

Following the extraction of the most important words in the corpus, the abstract representation sampling consisted of 10, 15, 20, 25, 50, 75, 100, 150, 200, 300, 500 and 750 term-words. In order to achieve accurate estimation, a 10-fold cross-validation method was used. Precision Recall and F-score were the evaluation metrics applied for comparing and evaluating the performance of classifiers.

The tool that was used to apply the classifiers was WEKA. It gathers together algorithms for classification, regression, clustering, association rules, visualization and algorithm development. The program is written in Java and it was developed at the University of Waikato in New Zealand **[4, 6]**.

The classifiers were chosen from version 3.7.12 of WEKA for developers. These were:

- Two Bayesian classifiers: Naïve Bayes and Naïve Bayes Multinomial,
- Three Function classifiers: Multilayer Perceptron, Simple Logistic, and SMO(SVM),
- Two Lazy classifiers: IBk and Kstar,
- Two Metalearning classifiers: Classification Via Regression and Logit Boost,
- Three Rule classifiers: Decision Table, JRip, and PART,
- Two Tree classifiers: LMT and Random Forest.

Table 2. F-score (%) with words from DEVMAX.DF.

Vector Size	10W	15W	20W	25W	50W	75W	100W	150W	200W	300W	500W	750W

	Classifier												
BIN	*NaïveBayes (NB)*	89,9	92,5	92,8	93,0	93,6	94,1	94,4	95,1	95,5	95,3	95,6	95,8
	NBMultinomial	87,0	89,3	91,8	94,2	95,1	95,9	96,0	96,1	96,2	95,8	95,9	96,3
	MLP	89,5	92,9	92,3	92,4	93,8	94,2	94,5	96,1	95,5	96,6	fail	fail
	SimpleLogistic	90,0	92,9	91,8	94,3	96,1	96,4	97,0	97,1	96,4	96,9	96,5	96,0
	SMO	89,0	92,9	91,8	93,3	95,3	96,4	96,0	96,2	96,1	97,2	97,1	96,7
	IBk	89,9	92,6	92,7	93,5	92,4	91,9	92,3	90,5	85,8	82,1	73,0	71,1
	Kstar	90,2	92,9	92,5	92,2	92,4	91,9	92,2	90,6	87,2	84,2	76,4	73,2
	Class.ViaRegress.	86,2	86,4	88,7	90,5	93,5	94,2	94,6	94,5	93,7	95,2	95,2	95,2
	LogitBoost	87,2	90,7	91,8	94,3	94,6	96,2	95,5	96,3	96,3	96,0	96,1	96,1
	DecisionTable	86,8	89,0	90,8	91,6	91,5	91,0	91,8	91,0	91,0	92,0	91,7	91,4
	JRip	86,5	92,4	90,9	92,0	92,2	93,1	94,1	93,7	93,3	93,6	93,0	93,5
	PART	89,8	92,0	92,7	92,5	92,3	93,5	94,9	94,1	94,0	93,2	93,6	94,3
	LMT	90,0	92,9	93,2	94,3	96,1	96,4	96,8	96,9	96,2	96,9	96,2	96,0
	RandomForest	90,0	92,8	93,1	93,0	93,8	94,6	95,8	96,4	97,1	97,5	96,9	97,2
TF	*NB*	79,9	87,7	87,2	90,5	90,4	91,1	91,4	91,9	92,8	94,2	94,8	95,0
	NBMultinomial	87,2	89,4	92,6	94,8	95,7	95,8	95,9	96,5	96,3	96,5	96,1	97,2
	MLP	88,4	91,7	91,7	93,0	93,6	93,7	92,6	93,0	92,5	91,3	fail	fail
	SimpleLogistic	87,2	92,7	92,8	95,2	95,4	96,2	96,0	96,0	96,0	94,6	95,4	94,6
	SMO	80,8	83,9	89,2	91,3	95,1	95,2	94,9	93,4	94,7	94,8	95,6	95,1
	IBk	89,5	91,8	92,1	93,3	88,1	87,0	85,9	86,3	85,8	80,0	77,5	73,2
	Kstar	90,2	92,6	91,6	92,3	90,6	89,7	88,9	87,7	83,8	79,9	74,4	72,0
	Class.ViaRegress.	87,3	87,0	88,5	89,6	93,1	92,9	93,2	93,3	93,6	93,8	93,8	94,0
	LogitBoost	87,2	90,7	91,8	94,2	94,9	95,6	95,6	96,3	96,3	95,7	95,3	95,3
	DecisionTable	86,8	89,0	90,8	91,2	91,3	90,9	91,8	91,0	91,0	92,0	91,7	91,4
	JRip	86,7	92,3	90,7	90,8	92,2	93,3	93,0	93,7	93,6	94,3	93,7	93,4
	PART	89,4	91,9	92,9	93,0	93,4	93,3	93,8	94,2	94,2	93,8	94,2	92,9
	LMT	89,6	92,3	93,2	95,2	95,4	96,2	95,8	95,1	96,0	94,6	95,1	94,3
	RandomForest	90,0	92,6	92,3	93,0	93,6	94,3	96,4	96,6	96,4	97,6	97,6	96,6

Table 3. F-score (%) with words from TF.IDF.

	Vector Size	10W	15W	20W	25W	50W	75W	100W	150W	200W	300W	500W	750W
	Classifier												
BIN	*NaïveBayes(NB)*	83,9	83,5	84,6	86,9	92,3	93,0	93,3	93,1	93,3	94,8	93,3	95,8
	NBMultinomial	77,3	82,3	85,5	88,8	93,8	94,9	94,7	94,8	93,2	95,5	95,1	96,3
	MLP	81,9	82,6	83,9	87,5	92,9	95,1	95,1	95,2	95,6	96,3	fail	fail
	SimpleLogistic	80,4	83,2	86,1	87,7	93,5	94,9	95,6	95,9	96,7	95,7	96,4	96,0
	SMO	84,7	83,5	86,0	87,5	92,2	93,3	93,6	94,6	95,7	95,9	95,8	96,7
	IBk	81,6	80,6	80,6	85,6	86,0	86,5	87,1	83,2	80,2	79,3	67,8	71,1
	Kstar	81,7	81,0	82,8	86,4	87,0	88,5	87,7	84,5	82,0	80,7	70,4	73,2

	Class.ViaRegress.	81,7	84,6	86,3	87,0	91,9	93,7	93,8	93,4	93,6	94,0	93,7	95,2
	LogitBoost	81,7	82,4	84,8	88,3	92,4	94,0	94,5	94,7	94,4	96,0	95,8	96,1
	DecisionTable	82,3	81,5	83,3	81,6	89,0	92,5	92,0	92,0	91,7	92,0	92,1	91,4
	JRip	79,5	81,6	83,7	83,3	90,2	91,3	93,2	92,0	92,7	90,4	92,0	93,5
	PART	82,2	81,9	84,2	86,7	90,0	92,0	92,1	92,3	93,0	92,6	93,1	94,3
	LMT	80,8	82,8	86,3	87,7	93,5	94,9	96,0	95,9	96,5	95,7	96,4	96,0
	RandomForest	82,2	82,4	86,1	89,2	93,6	95,8	96,7	96,3	96,7	97,4	96,6	97,2
	NB	74,0	75,9	77,7	80,2	85,7	87,9	89,2	90,1	91,0	92,8	93,0	93,0
	NBMultinomial	81,3	83,3	86,0	87,1	92,5	94,8	94,5	95,2	95,8	97,3	96,7	96,6
	MLP	80,8	81,8	84,1	87,9	91,6	94,8	92,9	93,4	91,7	84,8	fail	fail
	SimpleLogistic	82,1	84,5	86,9	87,9	93,7	94,4	95,2	94,2	94,7	95,0	95,3	95,0
	SMO	76,9	78,9	81,0	83,8	90,2	93,1	92,6	93,0	93,4	94,3	92,9	94,1
	IBk	75,7	75,9	76,2	80,0	79,4	82,5	79,6	78,2	75,8	75,9	72,0	66,0
TF	*Kstar*	79,6	77,6	79,8	80,4	80,1	80,7	77,1	73,4	72,2	70,1	60,5	57,9
	Class.ViaRegress.	81,3	84,6	86,1	87,2	90,1	92,8	93,0	91,6	92,3	92,3	92,3	92,3
	LogitBoost	80,8	83,8	85,8	87,9	92,6	94,7	94,0	94,3	94,4	96,0	95,7	95,3
	DecisionTable	82,0	83,0	81,7	81,9	89,5	92,5	91,9	91,5	91,5	91,8	91,9	92,0
	JRip	80,8	81,1	81,7	83,2	90,3	92,0	92,1	92,7	92,0	91,4	91,6	91,6
	PART	80,9	81,9	83,4	83,9	90,9	92,3	91,7	92,2	92,1	91,5	91,4	90,8
	LMT	82,1	84,2	86,9	87,9	93,7	94,7	95,0	94,3	94,7	95,0	95,3	95,0
	RandomForest	81,0	85,5	87,6	89,7	93,2	95,4	96,8	96,2	96,6	96,3	97,0	97,4

Table 4. Results (%) of the best classifiers.

Classifier	***Method***	***Vector***	***F-score***	***Precision***	***Recall***
RandomForest	DEVMAX.DF-tf	500W	97,60	97,60	97,60
RandomForest	DEVMAX.DF-bin	300W	97,50	97,50	97,50
RandomForest	TF.IDF-bin	300W	97,40	97,40	97,40
RandomForest	TF.IDF-tf	750W	97,40	97,40	97,40
NBMultinomial	TF.IDF-tf	300W	97,25	97,30	97,20
NBMultinomial	DEVMAX.DF-tf	750W	97,20	97,20	97,20
SMO	DEVMAX.DF-bin	300W	97,20	97,20	97,20
SimpleLogistic	DEVMAX.DF-bin	150W	97,10	97,10	97,10

Nevertheless, as Table 4 shows, the best classifier was Random Forest which achieved the highest Precision (P), Recall (R) and F-score (F1) rates in all four methods: DEVMAX.DF-bin (binary appearance), DEVMAX.DF-tf (frequency appearance), TF.IDF-bin and TF.IDF-tf. It yielded up to F1=97,6% in DEVMAX.DF-tf and did not fall under F1=97,4% in TF.IDF-bin and TF.IDF-tf. Naïve Bayes Multinomial, SMO (SVM) and Simple Logistic were also achieved F-scores greater than 97%. Naïve Bayes Multinominal performed better with tf and yielded F1=97,25% in TF.IDF and F1=97,2% in DEVMAX.DF. SMO (SVM) and Simple Logistic achieved an F-score of 97,2% and 97,1% respectively in DEVMAX.DF-bin. The excellent results of each classifier were produced from 150 to 750 vector size in word-terms.

Regardless of the method, Random Forest yielded the highest scores. This is no surprise as it is considered one of the most powerful and successful algorithms, with many applications in real life (banking, medicine, stock market, e-commerce, etc.), which can handle very large numbers of

input attributes **[25, 26]**. The specific method, DEVMAX.DF, boosted the algorithm even more.

Moreover, DEVMAX.DF performed better than classic TF.IDF with all the algorithms. This is especially noticeable with smaller vector size, since it manages to correctly detect the best words for document representation earlier than classic TF.IDF. It is also illustrated in Fig. 1 where the average performance of all classifiers is shown for each method individually. DEVMAX.DF has apparently better average performance than TF.IDF especially in small size vectors.

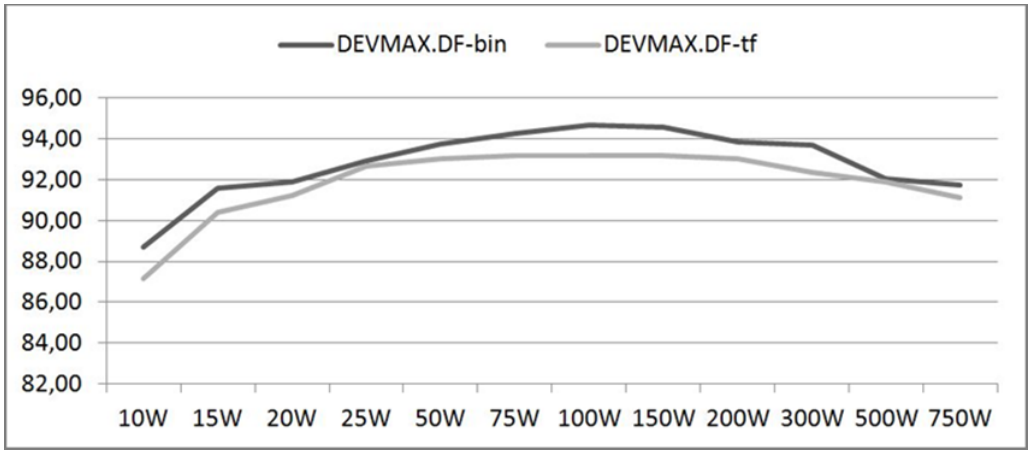

Fig 1. Average F-score (%) performance of all classifiers.

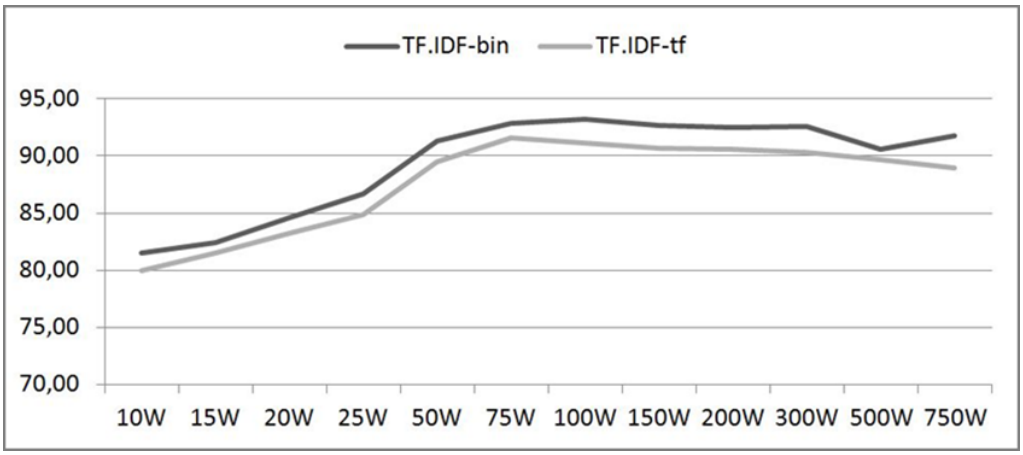

Fig 2. Average F-score (%) performance for all classifiers of binary (bin) and term frequency (tf) representations for DEVMAX.DF.

Another significant observation is that binary representation of document vectors acts in a more beneficiary way than frequency representation in the performance of the examined classifiers. This is illustrated in Fig. 2 and Fig. 3 where the dark gray lines correspond to binary representations while light ones indicate term frequency representations.

IV. Conclusions

An assessment of the use of text classification in digital libraries took place. During the pre-processing, two weighting methods, TF.IDF and DEVMAX.DF with binary and term frequency appearance, were used. The software used to apply classification algorithms was WEKA. Overall, this research indicated that digital libraries could substitute manual classification with the proposed approach. DEVMAX.DF, which proved to be more effective than TF.IDF, produced an F-score greater than 97% in some classifiers. In addition, this method, unlike TF.IDF, yielded adequate results with a small amount of words. However, this raises the question whether the same approach can be exploited with the use of smaller texts.

Hence, in the future the aim is to experiment with titles instead of abstracts. Another important future aspect is to apply clustering techniques to encourage and identify classes and topic fusion.

V. REFERENCES

[1] A. Joorabchi, A. Mahdi, An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata, Journal of Information Science, 37(5) (2011) 499-514.

[2] I. Triantafyllou, A. Koulouris, S. Zervos, M. Dendrinos, D. Kyriaki-Manessi, G. Giannakopoulos, Significance of clustering and classification applications in digital and physical libraries, In: Proceedings of 4th International Conference IC-ININFO, Madrid, Spain, 2014.

[3] F. Vorgia, I. Triantafyllou, A. Koulouris., Hypatia Digital Library: A text classification approach based on abstracts, Strategic Innovative Marketing, Springer International Publishing, (2017), 727-733.

[4] R.R. Bouckaert, E. Frank, M.A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, WEKA- experiences with a Java open-source project, Journal of Machine Learning Research 11 (2010) 2533-2541.

[5] Machine Learning Group at the University of Waikato, WEKA 3- data mining with open source machine learning software in Java, 2015.

[6] F. Sebastiani, Machine learning in automated text categorization, ACM computing surveys (CSUR) 34 (2002) 1-47.

[7] I. H. Witten, E. Frank, M.A. Hall, Data mining: practical machine learning tools and techniques, Morgan Kaufmann, 2011.

[8] I. Triantafyllou, I. Demiros, S. Piperidis, Two level self-organizing approach to text classification, In: Proceedings of RANLP-2001: Recent Advances in NLP, 2001.

[9] C. C. Aggarwal, C. Zhai, A survey of text classification algorithms, Mining text data (2012) 163-222.

[10] T. S. Guzella, W. M. Caminhas, A review of machine learning approaches to Spam filtering, Expert Systems with Applications 36 (2009) 10206–10222.

[11] L. Huan, Y. Lei, Toward integrating feature selection Algorithms for Classification and Clustering, IEEE Transaction on Knowledge and Data Engineering 17(4) (2005).

[12] R. Islam, J. Abawajy, A multi-tier phishing detection and filtering approach, Journal of Network and Computer Applications 36 (2013) 324–335.

[13] I. Ahmed, R. Ali, D. Guan, Y. K. Lee, S. Lee, T. C. Chung, Semi-supervised learning using frequent itemset and ensemble learning for SMS classification, Expert Systems with Applications, 42(3) (2015) 1065-1073.

[14] S. J. Delany, M. Buckley, D. Greene, SMS spam filtering: Methods and data, Expert Systems with Applications 39 (2012) 9899–9908.

[15] W. Liu, T. Wang, Index-based Online Text Classification for SMS Spam Filtering, Journal of Computers 5(6) (2010).

[16] D. Irani, S. Webb, C. Pu, K. Li, Study of trend-stuffing on twitter through text classification, In: Proceedings of Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), 2010.

[17] M. Daniela, R. F. Nevesa, N. Horta, Company event popularity for financial markets using Twitter and sentiment analysis, Expert Systems with Applications 71(1) (2017) 111-124.

[18] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, H. Christensen, Detecting suicidality on Twitter, Internet Interventions 2(2) (2015) 183-188.

[19] E. Barfian, B. H. Iswanto, S. M. Isa, Twitter Pornography Multilingual Content Identification Based on Machine Learning, Procedia Computer Science 116 (2017) 129-136.

[20] B. Desmet, V. Hoste, Online suicide prevention through optimised text classification, Information Sciences 439–440 (2018) 61–78.

[21] L. Li, Y. G. Huang, Z. W. Liu, Chinese text classification for small sample set, The Journal of China Universities of Posts and Telecommunications 18 (2011) 83–89.

[22] W. J. Wilbur, K. Sirotkin, The automatic identification of stop words, Journal of Information Science 18 (1992) 45-55.

[23] W. B. Croft, D. Metzler, T. Strohman, Search engines: information retrieval in practice, Addison-Wesley, 2010.

[24] K. S. Jones, A statistical interpretation of term frequency and its application in retrieval, Journal of Documentation 28 1972 11-21.

[25] K. Fawagreh, M. Medhat Gaber, E. Elyan, Random forests: from early developments to recent advancements, Systems Science & Control Engineering 2(1) (2014) 602-609.

[26] A. J. Wyner, M. Olson, J. Bleich, Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers, Journal of Machine Learning Research 18 (2017)

VI. AUTHORS

Ioannis Triantafyllou is Associate Professor in the Department of Archival, Library & Information Studies at the University of West Attica. He received his Ph.D. from National Technical University of Athens, Department of Electrical and Computer Engineering in 2003.He has worked in the past as a scientific associate in many European and Greek research projects at the Institute for Language and Speech Processing (ILSP / RC "Athena"). Since 2016 he is a member of the research team of the CrossCult European project (Horizon2020). The field of scientific interests and publications are: Digital Libraries, Data Mining, Text Mining, Text Classification & Clustering, Ontologies & Metadata, Information Extraction, Information Retrieval, Automated Summary & Text Synthesis and Translation Memories.

Frosso Vorgia received her bachelor's degree in Library and Information Science from T.E.I. of Athens (currently known as University of West Attica) in 2016. She worked at Uni.Systems as a librarian on several projects since 2017. Her research interests are Digital Libraries, Data analysis, Office management, Sustainable fishing and Fisheries and aquaculture in Eastern Mediterranean Subarea.

Alexandros Koulouris is Assistant Professor in the Department of Archival, Library & Information Studies at the University of West Attica. He has been involved in several European and national R&D projects in the field of information management (DELOS, EuropeanaLocal, Europeana, CrossCult). From 2011, he actively participates in Europeana as a member of the Europeana Network Association. He is member of the Information Management laboratory at the University of West Attica. His research interests include information policy, digital libraries, repositories and open access. He has published more than

45 articles in journals and conferences. In the past, he has worked as a librarian for the National Technical University of Athens and for the National Documentation Centre of Greece. He holds a PhD in Information Science from Ionian University, a BA in Library Science from the Technological Educational Institute of Athens and a BA (Hon) in International and European Studies from Panteion University. More information can be found at users.uniwa.gr/akoul.