Text Categorization Using n-Gram Based Language Independent Technique

Jelena Graovac

University of Belgrade, Faculty of Mathematics, Department of Computer Science,
Studentski trg 16, 11000 Belgrade, Serbia
jgraovac@matf.bg.ac.rs

Abstract. This paper presents a language and topic independent, byte-level n-gram technique for topic-based text categorization. The technique relies on an n-gram frequency statistics method for document representation, and a variant of k nearest neighbors machine learning algorithm for categorization process. It does not require any morphological analysis of texts, any preprocessing steps, or any prior information about document content or language. For driving experiments, five document collections are used: Ebart-3 in Serbian, Reuters-21578 and 20-Newsgroups in English, Tancorp-12 in Chinese and Maslah-10 in Arabic. Micro- and macro-averaged F1 measures are employed for evaluation process. Comparisons between results obtained by the presented technique and results obtained by other n-gram based and traditional "bag of words" text categorization techniques, demonstrate that this technique is sound and promising.

Keywords: Text Categorization, n-Grams, kNN

1 Introduction

Text Categorization (TC) is the task of classifying unlabeled natural language documents into a predefined set of categories. With the rapid growth of the Internet, TC has become one of the key approaches to organizing and handling data. It has many useful real-world applications, for example, filtering a stream of news for a particular interest group, classification of academic papers by technical domains and sub-domains, organizing patient reports in health-care organizations using multiple aspects such as taxonomies of disease categories, types of surgical procedures, or insurance reimbursement codes. Also, widespread applications of TC are spam filtering, email routing, language identification, genre classification, readability assessment, sentiment polarity detection and many others.

Data mining and machine learning communities have been challenged by many difficult problems presented by TC, mainly due to high dimensionality of text data, and to complex semantics of natural languages. Although most of the research activity has concentrated on English text, the management and study of TC in languages other than English is a growing need. It produces additional difficulties in text analysis due to specific characteristics of different languages.

For example, Serbian language has several properties that significantly influence text analysis: the use of two alphabets (Cyrillic and Latin), phonologically based orthography, rich morphological system, free word order, special placement of enclitics and complex agreement system [13] while Arabic, for instance, is a highly inflectional and derivational language, which makes morphological analysis a very complex task. Regardless of language, an additional difficulty in handling text documents is presence of different kinds of textual errors, such as typing, spelling and grammatical errors. TC must work reliably on all input, and thus must tolerate these kinds of problems to a certain level.

Standard approaches [3] to TC are usually based on traditional word-based vector document representation (Bag of Words, BOW), where each entry of the vector reflects frequency of a word occurrence within the corresponding document (Term Frequency, TF), and within the entire corpus (Inverse Document Frequency, IDF). Although representation of a document at the level of words seems to be intuitive solution, in some languages it could be a particular problem. For example, Chinese does not have explicit whitespace between words so word segmentation itself is a difficult problem in this language. One of the main problems with BOW techniques is that word order information is lost. Traditional preprocessing of documents, such as eliminating stop words, pruning rare words, stemming and normalization, can improve the representation, but it is still limited and become language dependent. Nevertheless, this traditional document representation technique has been used very successfully in conjunction with many learning algorithms, especially with Support Vector Machines (SVM), but also with K Nearest Neighbors (kNN), Decision Trees (DT), Rule-Based Classifiers (RBC) (e.g. RIPPER), Bayesian Classifier, Rocchio's algorithm, Neural Networks, Genetic Algorithms, Latent Semantic Analysis, Centroid Based Classifier, Conditional Random Fields (CRF), Hidden Markov Models (HMM) and others (for references see [1]).

This paper presents quite a different technique for TC that avoid many of difficulties listed above. It is based on byte-level n-gram frequency statistics method for document representation, derived from Kešelj's n-gram based method for authorship attribution [4], and a kind of kNN (for $k = 1$) machine learning algorithm for categorization process. It is fully language and topic independent, and it can be applied to corpora in other languages and domains, without any changes. This technique has been successfully used by the author of this paper in [2] for sentiment polarity detection in movie reviews in English and Spanish. In [9] this technique has been used for topic-based Serbian TC, while in [1] it is used for English, Chinese, and Serbian TC. Here we will additionally test this technique for topic-based Arabic TC, and we will employ new dissimilarity measure.

The paper is organized as follows. Section 2 discusses related work while Section 3 introduces the technique we propose for solving TC problem. Section 4 describes different evaluation measures and datasets used in our experiments. Experimental results are presented and discussed in Section 5. Finally, Section 6 provides conclusions and ideas for future work.

2 Related Work

TC has been extensively studied. This section will focus on well chosen previously defined techniques based on n-gram and BOW document representation models. A particular attention will be paid to those techniques that use the same datasets as we do in this paper. The corresponding results will be used in our comparative study.

N-Gram Based Techniques. N-gram based model for document representation has been used by Cavnar and Trenkle in [15]. They presented a TC technique based on using rank order statistics to compare test document and category document profiles of most frequent character n-grams. Experiments confirmed that this technique works very well for language categorization. Rahmoun and Elberrichi [8] also used character n-grams to represent documents. They conducted several experiments on two benchmark corpora – the Reuters-21578 and the 20-Newsgroups. The results show the effectiveness of this approach compared to the BOW and stem representations. The Reuters-21578 corpus was also used in [10], where Makoto Suzuki with his colleagues test a new mathematical model of TC, based on character n-gram model for document representation. Zhihua Wei [14] and Xi Luo [6] with their colleagues performed Chinese TC on the Tancorp-12 corpus. They compared different feature weights using different combinations of character n-grams and SVM algorithm for categorization.

BOW Techniques. Significant results of the TC on English corpora Reuters-21578 and 20-Newsgroups are achieved by Lan and his colleagues [5]. They studied several widely-used supervised and unsupervised term weighting methods in combination with SVM and kNN algorithms for TC. Tan and others in [11] reported the best performance of several learning algorithms using Tancorp-12 corpus in Chinese and three corpora in English (20-Newsgroups, WebKB and Sector-48). In [7] Mesleh used SVM algorithm for Arabic TC. He presented an empirical comparison of seventeen traditional feature sub-set selection metrics for TC tasks.

3 Byte-n-Gram-Based Categorization Technique

In this paper we present a technique based on byte-level n-gram frequency statistics method for document representation and a kind of kNN (for $k = 1$) machine learning algorithm for categorization process.

Definition 1. *Given a sequence of tokens $S = (s_1, s_2, ..., s_{N+(n-1)})$ over the token alphabet A, where N and n are positive integers, an n-gram of the sequence S is any n-long subsequence of consecutive tokens. The i^{th} n-gram of S is the sequence $(s_i, s_{i+1}, ..., s_{i+n-1})$ [1].*

The term n-gram could be defined on a word, character or byte level. Extracting byte n-grams from a document is like moving an n-byte wide “window” across the document, byte by byte. Each window position covers n bytes, defining a single n-gram. For example, the sentence "No pain, no gain!" is composed of the following n-grams: 1-grams: N; o; _; p; a; i; n; ,; _; n; o; _; g; a; i; n; !; 2-grams: N o; o _; _ p; p a; a i; i n; n ,; , _; _ n; n o; o _; _ g; g a; a i; i n; n !; 3-grams: N o _; o _ p; _ p a; p a i; a i n; i n ,; n , _; , _ n; _ n o; n o _; o _ g; _ g a; g a i; a i n; i n !; end so on. The underscore character ("_") is used here to represent spaces and (";") is used as a delimiter. Since this technique is based on the byte level n-grams, it has many advantages: language and topic independence, no linguistic knowledge is required, independence of encoding and alphabet, relative insensitivity to spelling variations/errors, word stemming is obtained essentially for free, only one pass processing is required (for more details see [1]). The main disadvantage of using n-grams is that it yields a large number of n-grams.

3.1 Categorization Procedure

In general, machine learning approach to TC is composed of two phases: training and testing phase. During the training phase, a classifier is automatically built by learning from a set of previously classified documents (training set). During the testing phase, a new document is assigned to the most suitable categories based on the likelihood suggested by the classifier.

In the TC model that we propose in this paper, training phase consists of the following steps:

– Concatenate all the training documents that belong to the same category into a single document. Each category will be thereby presented by one document only (the "category document").
– For each document (test or category, construct its profile:
  - Select a specific n-gram size n (e.g. 6-gram, 7-gram etc.).
  - Extract the byte-level n-grams for that particular value of n. Calculate the normalized (relative) frequencies, for each n-gram.
  - List the n-grams by descending frequency, so that the most frequent are listed first.
  - Select a specific profile length L at which to cut off all profiles (test document and category).

After the training phase, each test document profile and category profile will be presented as a set of L pairs $\{(x_1, f_1), (x_2, f_2)...(x_L, f_L)\}$ of the most frequent n-grams and their normalized frequencies. Based on these profiles, in testing phase we will determine which categories will be assigned to the certain document. Testing phase consists of the following steps:

– Compute a dissimilarity measure between the test document's profile and each of the category's profiles.
– Select the category (or categories) whose profile has the smallest value of dissimilarity measure with the document's profile.

Note that an important role in the testing phase belongs to a dissimilarity measure.

3.2 Dissimilarity Measures

Definition 2. *Dissimilarity measure d is a function that maps the Cartesian product of a set of profiles into the set of positive real numbers R. Symbolically $d : \Pi \times \Pi \longrightarrow R$. It should reflect the dissimilarity between two profiles and it should meet the following conditions:*

- *$d(P1, P2) \geq 0$*
- *$d(P_1, P_2) = d(P_2, P_1)$*
- *The value $d(P_1, P_2)$ should be small if P_1 and P_2 are similar and should be large if they are not similar.*

where P, P_1 and P_2 are any profiles that belongs to the set Π. The last condition is informal as the notion of similarity (and thus the dissimilarity) is not strictly defined [12].

In this paper we use three dissimilarity measures. Two of them are the same as those used in [1]:

$$dK(\mathcal{P}_1, \mathcal{P}_2) = \sum_{n \in profile} \left(\frac{2 \cdot (f_1(n) - f_2(n))}{f_1(n) + f_2(n)} \right)^2 \tag{1}$$

$$dSD(P_1, P_2) = |P_1 \triangle P_2| \tag{2}$$

where the first measure was introduced by Kešelj [4] and the second one was introduced by the author of this paper [1]. The first measure has the form of relative distance where $f_1(n)$ and $f_2(n)$ are frequencies of an n-gram n in the category profile P_1 and the test document profile P_2, respectively. The second measure represents the number of n-grams that appear in the union of the profiles and not in their intersection. In mathematics, this is known as symmetric difference, so we refer to this measure as dSD.

The third measure is introduced by Cavnar and Trenkle [15] and it represents a simple rank-order statistics. It is calculated as follows: For each n-gram in a test document's profile, its counterpart in a category's profile is located, and then the distance from it is calculated (how far out of place it is). If an n-gram is not in the category's profile, it takes a maximum out-of-place value, which is equal to the number of n-grams in the profile. The dissimilarity measure between the document and the category profiles is calculated as a sum of all of the out-of-place values for all n-grams. We refer to this measure as dOP (Out-of-Place).

3.3 Implementation Details

For producing n-grams and their normalized frequencies, the software package *Ngrams* written by Kešelj [4] is used. For the process of categorization, the software package *NgramsCategorization* developed by the author of this paper is used. Source code can be obtained on request.

4 Experimental Framework

4.1 Evaluation Measures

For evaluating the performance of the technique, we use the typical evaluation metrics that come from information retrieval – precision (P), recall (R), and F1 measure:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = \frac{2PR}{P + R} \tag{3}$$

where TP (True Positives) is defined as the number of documents that were correctly assigned to the considered category, TN (True Negatives) is the number of the assessments where the system and a human expert agree on a negative label, and FN (False Negatives) is the number of negative labels that the system assigned to documents otherwise assessed as positive by the human expert [3].

All presented measures can be aggregated over all categories in two ways: micro-averaging – global calculation of the measure considering all the documents as a single dataset regardless of categories, and macro-averaging – the average on measure scores of all the categories.

4.2 Data Collections

For our experiments, we chose five document collections: Ebart-3 in Serbian, Reuters-21578 and 20-Newsgroups in English, Tancorp-12 in Chinese and Maslah-10 in Arabic.

Ebart-3 – in Serbian. This corpus is selected by the author of this paper as a subset of the Ebart corpus, the largest digital media corpus in Serbia. It consists only of articles from the Serbian daily newspaper "Politika" that belong to columns Sport, Economics and Politics, published from 2003 to 2006. There are 3366 such articles.

Reuters-21578 and 20-Newsgroups – in English. Reuters-21578 is currently the most widely used testing collection for text categorization research. Usually, only the top 10 largest categories are taken into consideration. This corpus is multi-label and it is characterized by skewed category distribution. Contrary to this, 20-Newsgroups corpus is a single-label collection of approximately 20000 newsgroup documents, evenly divided into 20 different newsgroups, each corresponding to a different topic.

Tancorp-12 – in Chinese. Tancorp-12 is a single-label corpus. It consists of 14150 documents unevenly divided into 12 categories.

Mesleh-10 – in Arabic. This corpus is introduced by Mesleh [7] and it consists of 10 categories. It is collected from online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, Al-Dostor and a few other specialized web sites. The corpus contains 78421 documents that vary in length.

For more details about Serbian, English and Chinese corpora see [1].

5 Experiments and Results

In order to test the effectiveness of the proposed n-gram based TC technique, with respect to different n-gram size n and profile length L using different dissimilarity measures, we conducted an extensive set of experiments over the corpora in Serbian (Ebart-3), English (Reuters-21578 and 20-Newsgroups), Chinese (Tancorp-12) and Arabic (Mesleh-10). Results are presented in Fig. 1 and 2. From these figures we conclude that the optimum results for different languages are obtained for different n-gram size n ($n = 7$ for Serbian and English, $n = 6$ for Chinese and $n = 10$ for Arabic) and different profile length L. We also conclude that the choice of a dissimilarity measure does not significantly affect the classification accuracy. Interesting phenomenon in these figures is a sudden drop of performance after L exceeds the maximum possible profile length L for at least one category, for the considered n-gram size n (for more details see [1]). Also, we can see that in the case of Reuters-21578 corpus, values for micro- and macro-averaged F1 measure differ significantly. It can be explained by highly non-uniform distribution of the corpus and very low F1 values for the smallest categories due to the insufficient adjustment of the technique to multi-label feature of the corpus.

5.1 Comparison With Other Related Work

In order to evaluate performance of the technique presented in this paper, we compare the results with the published results obtained by other n-gram based and BOW TC techniques, briefly presented in Section 2. To provide an objective basis for comparison of our results with others, we used the same benchmark corpora and the same test and training split. Figures 3 and 4 present the results of those comparisons. We conclude that in the case of 20-Newsgroups and Tancorp, our technique shows better performance than other n-gram based techniques. Compared to BOW techniques, our technique reaches better accuracy than kNN (with exception of macro-averaged F1 for Reuters-21578 corpus) and worse than SVM (with exception of 20-Newsgroups and Mesleh-10 corpora). We stress out that in the case of Arabic Mesleh-10 corpus, our technique outperforms all previously published results. In the case of Ebart-3 corpus there are no published results, so we can not perform the comparison.

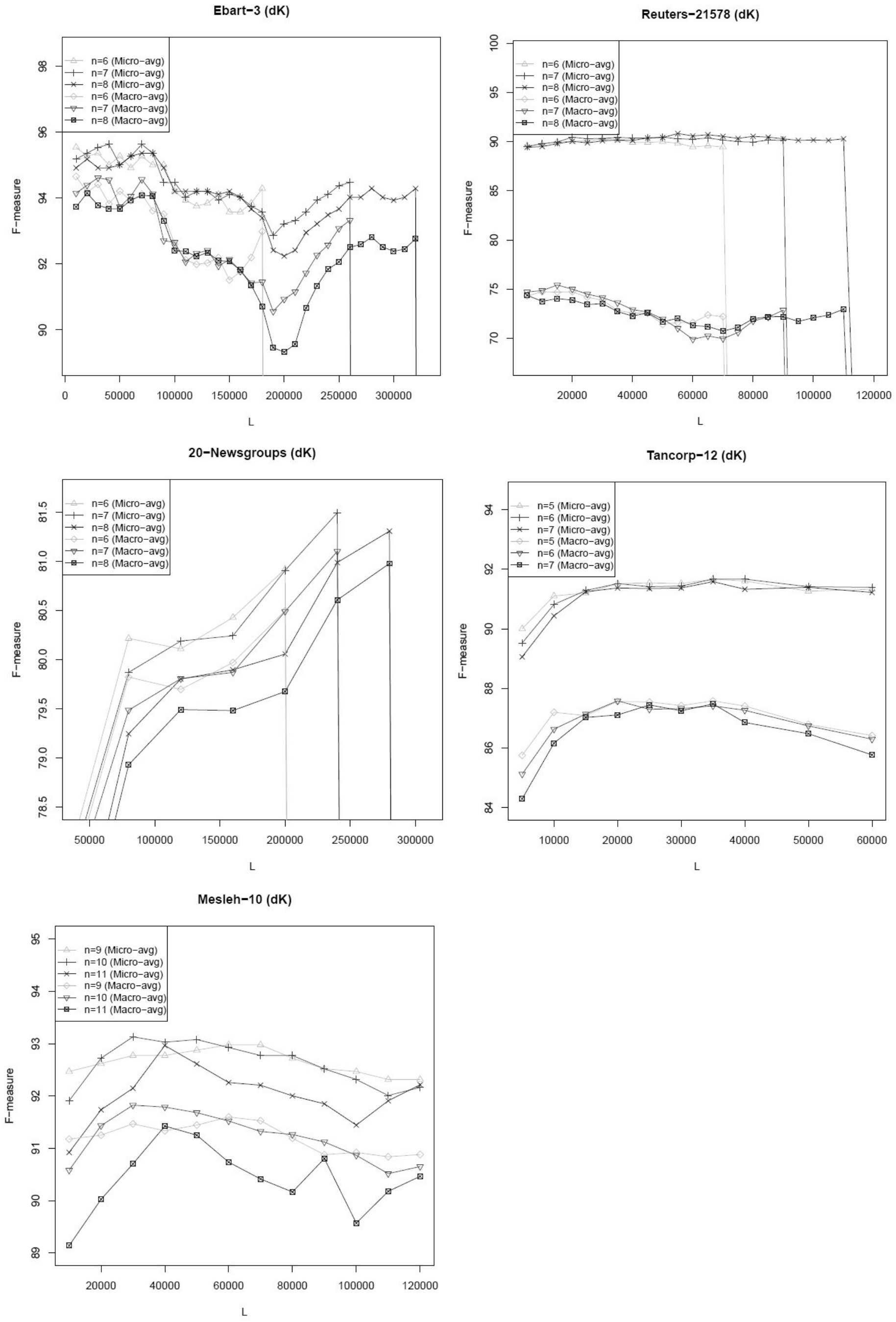

Fig. 1: Micro- and macro-averaged F1 measure for different values of n-gram size n with the dissimilarity measure dK, for corpora in Serbian (Ebart-3), English (Reuters-21578 and 20-Newsgroups), Chinese (Tancorp-12) and Arabic (Mesleh-10).

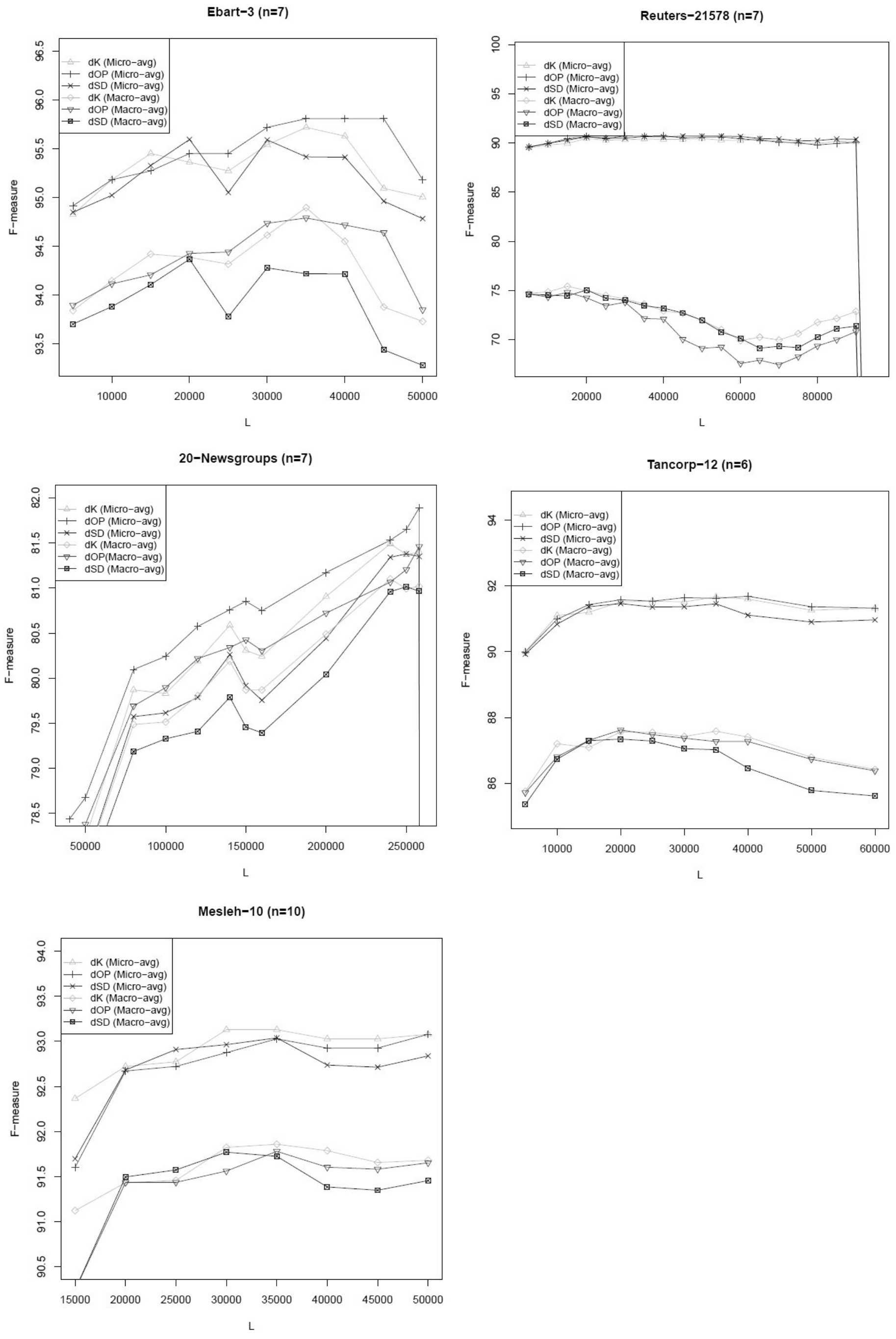

Fig. 2: Micro- and macro-averaged F1 measure for different dissimilarity measures with the chosen value of n-gram size n, for corpora in Serbian (Ebart-3), English (Reuters-21578 and 20-Newsgroups), Chinese (Tancorp-12) and Arabic(Mesleh-10).

6 Conclusion and Future Work

In this paper we presented a language-independent topic-based text categorization technique. It is based on a byte-level n-gram frequency statistics method for document representation and a variant of k nearest neighbors (for $k = 1$) machine learning algorithm for categorization process. For driving experiments we used five document collections in Serbian, English, Chinese and Arabic: Ebart-3, Reuters-21578 and 20-Newsgroups, Tancorp-12 and Mesleh-10, respectively. We obtained results that are comparable (in many cases better) with other n-gram based and "bag of words" techniques. Although optimum results for different languages are obtained for different n-gram size n and different profile length L, the overall success of the proposed technique proves it sound and promising.

The presented technique is fully language and topic independent, so it can be applied to corpora in other languages and domains without any changes, as long as training data exists. We plan to improve this technique by adding n-gram weighting factors schema, that comes from inter-document source. We plan to hybridize our approach by combining it with some other, language-specific and language-independent approaches, in order to improve accuracy.

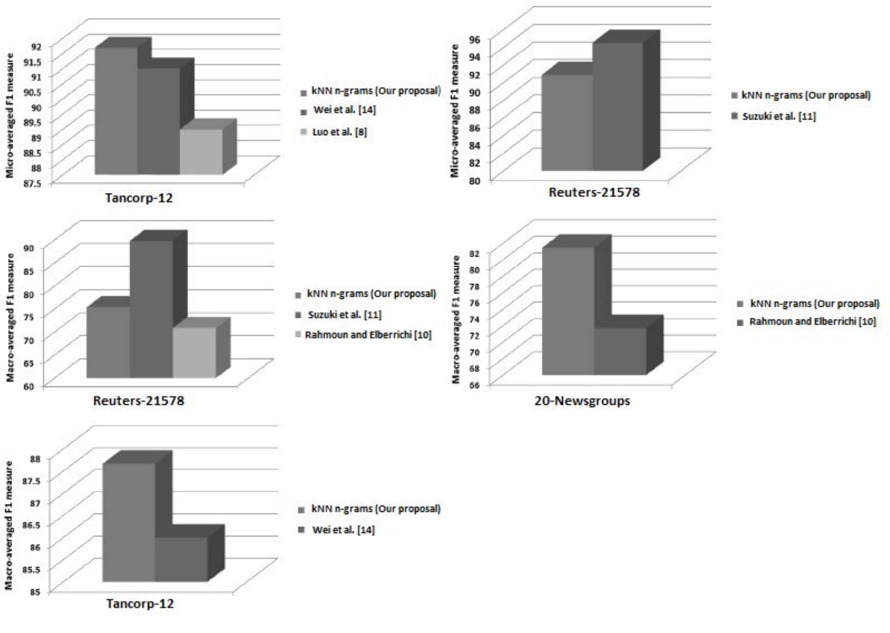

Fig. 3: Micro- and macro-averaged F1 measure comparison of the n-gram based technique proposed in this paper with other n-gram based techniques.

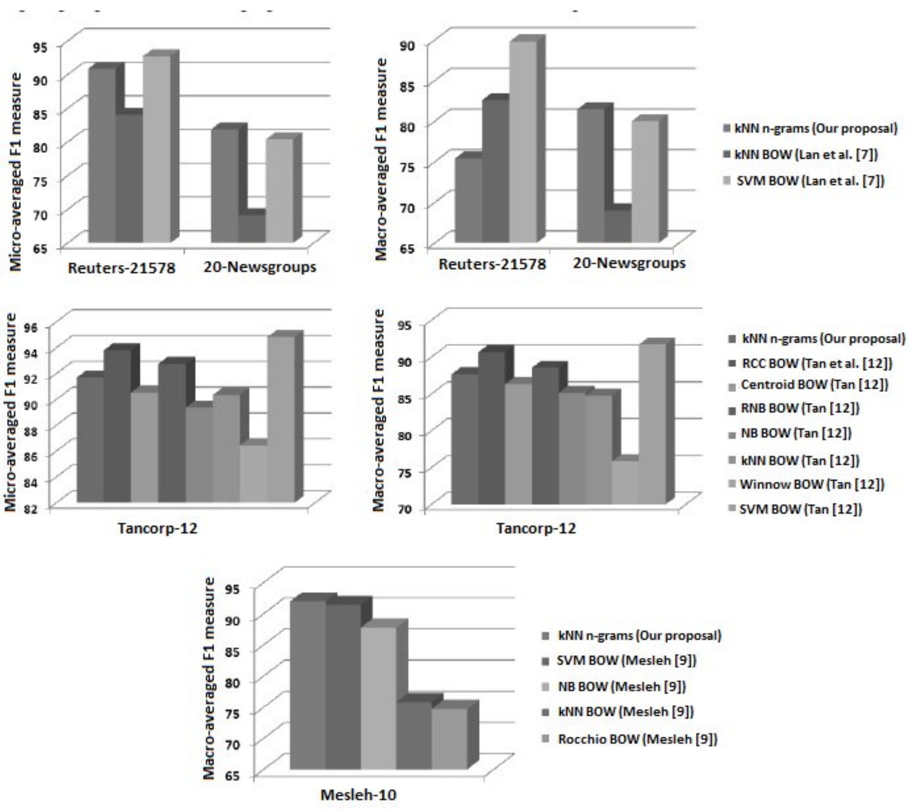

Fig. 4: Micro- and macro-averaged F1 measure comparison of the n-gram based technique proposed in this paper with other BOW techniques.

Acknowledgments. The work presented has been financially supported by the Ministry of Science and Technological Development, Republic of Serbia, through Projects No. III47003 and No. 174021.

References

1. Jelena Graovac. A variant of n-gram based language-independent text categorization. *Intelligent Data Analysis*, 18(4):677–695, 2014.
2. Jelena Graovac and Gordana Pavlović-Lažetić. Language-Independent Sentiment Polarity Detection in Movie Reviews: A Case Study of English and Spanish. *ICT Innovations 2014, Web Proceedings*, pages 13–22, 2014.
3. Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms.* Kluwer Academic Publishers, 2002.
4. Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics*, (3): 255–264, 2003.

5. Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):721–735, 2009.
6. Xi Luo, Wataru Ohyama, Tetsushi Wakabayashi, and Fumitaka Kimura. A study on automatic chinese text classification. In *International Conference on Document Analysis and Recognition (ICDAR), 2011*, pages 920–924. IEEE, 2011.
7. Abdelwadood Moh Mesleh. Feature sub-set selection metrics for Arabic text classification. *Pattern Recognition Letters*, 32(14):1922–1929, 2011.
8. Abdellatif Rahmoun and Zakaria Elberrichi. Experimenting N-Grams in Text Categorization. *Int. Arab J. Inf. Technol.*, 4(4):377–385, 2007.
9. Jelena Graovac. Serbian Text Categorization Using Byte Level n-Grams. In *Workshop on Computational Linguistics and Natural Language Processing of Balkan Languages – CLoBL 2012, First Balkan Conference in Informatics BCI 2012, Novi Sad*, pages 93–96, 2012.
10. Makoto Suzuki, Naohide Yamagishi, Yi-Ching Tsai, Takashi Ishida, and Masayuki Goto. English and Taiwanese text categorization using N-gram based on Vector Space Model. In *Information Theory and its Applications (ISITA), 2010 International Symposium on*, pages 106–111. IEEE, 2010.
11. Songbo Tan, Xueqi Cheng, Moustafa M Ghanem, Bin Wang, and Hongbo Xu. A novel refinement approach for text categorization. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 469–476. ACM, 2005.
12. Andrija Tomović and Predrag Janičić. A variant of n-gram based language classification. In *AI* IA 2007: Artificial Intelligence and Human-Oriented Computing*, pages 410–421. Springer, 2007.
13. Duško Vitas, Cvetana Krstev, Ivan Obradović, Ljubomir Popović, and Gordana Pavlović-Lažetić. Processing Serbian written texts: An overview of resources and basic tools. In *Workshop on Balkan Language Resources and Tools*, (21):97–104, 2003.
14. Zhihua Wei, Duoqian Miao, Jean-Hugues Chauchat, and Caiming Zhong. Feature selection on Chinese text classification using character n-grams. In *Rough Sets and Knowledge Technology*, pages 500–507. Springer, 2008.
15. William B Cavnar, John M Trenkle, et al. N-gram-based Text Categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.