Text Classification Based on Named Entities

Stefan Anđelić*, Miroslav Kondić*, Ivan Perić*, Marko Jocić*, Aleksandar Kovačević*
* Faculty of Technical Sciences/Department of informatics, Novi Sad, Serbia
1, kondicm@uns.ac.rs, ivanperic@uns.ac.rs, m.jocic@uns.ac.rs, kocha78@uns.ac.rs

***Abstract*—: This work contains an analysis of the impact that named entities have on thematic text classification. The categorization has been conducted on documents acquired from well-known datasets such as the Reuters 21578, 20 Newsgroups and the WebKB set. Multiple classifiers have been trained so as to eliminate the chance of any results being a direct consequence of the chosen algorithm. Results from multiple papers in this field that are based on the same datasets and the same, or similar, data splits, are used to compare the achieved results. Besides that, for each dataset and each classification algorithm a control model, that does not include named entity recognition, is created to confirm the validity of the results. Acquired results do not show any significant improvement when using named entities, and in some cases even show worse performance.**

I. INTRODUCTION

In the early nineties automatic control and document content analysis has gained a lot of attention in the field of information science. This interest has become a consequence of the sudden prevalence and availability of digital knowledge bases. This knowledge is usually contained in an unstructured form of online articles, forum posts or blogs. With the added gain in popularity of online social media, the use of text classification techniques and the need to label and structure the knowledge becomes more necessary.

Text classification represents one of the key tasks during the process of extracting knowledge from the aforementioned sources, as a part of the Information Retrieval field. Text categorization is a process of assigning predetermined labels to documents written in natural language. This kind of analysis can produce information about the sentiment, or in other words the positive or negative context regarding a certain theme, or even the exact theme covered in a specific document.

The impact that automatic text labeling has in the modern age, where information is a prized resource, is substantial. The knowledge about a specific theme that interests a person, whether gained through tracking of that person's favorite websites or social media posts, finds great application in deciding on the best content to serve to someone. It enables the formation of better, more personalized filters and search engines, as well as helping producers to find their market more easily.

Great number of approaches to text classification, that are based on ML (Machine Learning) algorithms, use words contained in the document as it's quantitative features. These methods rely on the intuitive assumption that the frequency of certain words within a text are a good indicator of the general theme. The main assumption that this research is based on is that named entities, such as personal names of people, organizations and locations, might be better suited for classification of text documents.

II. PREVIOUS WORK

The idea of using named entities for classifying documents based on their theme is not a new one, and does occur as the subject of multiple papers in the field such as [1], [2] and [3]. The main accent in most of these papers is on the use of exclusively named entities for classification, disregarding other features. Others use named entities in conjunction with other text features for clustering but not for document classification.

III. REFERENTIAL RESULTS

In her PhD thesis [4], Ana Cardoso-Cachopo provides several methods that combine standard machine learning algorithms, such as Naïve-Bayes, SVM, and k-NN. The author conducts a detailed analysis of the datasets used in the paper as well as the steps that were taken during text processing and transformation prior to training the classifiers. Datasets used are 20 Newsgroups [5] with a standard split to training and test subsets, the Reuters-21578 [6] split to subsets with 8 (R8) and 52 (R52) most prominent categories, the Cade dataset and WebKB [7]. The paper contains a detailed analysis of peroformance observed on the aformentioned methods and their combinations with LSI (*Latent Semantic Indexing*), proving better results with LSI.

For this paper, the most intresting results are specifically the ones observed on standard machine learning algorithms, and they will be used as a reference in further text. Table 1 contains the results provided in [4] for the accuracy measure.

IV. DATASETS

For this paper, classifiers have been trained on four datasets. Namely two subsets of the Reuters-21578 dataset, R8 and R52 the same as the ones proposed in [4]. Besides the Reuters dataset, the two other sets used are 20 Newsgroups and WebKB. While the author in [4] provides already prepared datasets, with all of the text cleaned, tokenized and reduced to a bag of words model, this work requires additional tokenization rules based on POS (*Part of Speech*) analysis for NER (Named Entity

TABLE I.
ACCURACY MEASURE FOR STANDARD ML ALGORITHMS

Classifier	Accuracy			
	R8	R52	20 Ng	WebKB
kNN	0.8524	0.8322	0.7593	0.7256
Naïve Bayes	0.9607	0.8692	0.8103	0.8352
SVM	0.9698	0.9377	0.8284	0.8582

recognition). This imposes a need for recreating the same or at the very least similar text processing steps, as the ones described in [4].

A. Reuters-21578

The collection of news articles contained in the Reuters dataset represents one of the most used sets for development of algorithms in the fields of text processing and information retrieval. The set was initially published in 1987, and it consists of 21578 documents that were manually labeled by the employees of Reuters in collaboration with the Carnegie Group corporation [6].

Most of the documents in this set have multiple labels joined to them, with the total number of distinct labels (categories) being 115. The main issue with the dataset is a very noticeable category imbalance. For example, the "Capacity Utilisation" class is assigned to only 4 documents, while the category with the highest number of documents, "Earnings and Earnings Forecasts" has 3923. It is for this reason that only subsets of this dataset are being used.

For the split to test and training documents, a standard split from the literature for this dataset exists and is called the modApté split. The final number of documents for the R8 set is 5485 for training and 2189 for testing, while the R52 subset contains 6532 training documents and 2568 test files.

B. 20 Newsgroups

This dataset is a collection of around 20000 documents, split into 20 different categories. These categories all refer to a news theme covered in each of the documents. Originally collected by Ken Lang [5], much like the Reuters dataset this set gained a reputation as one of the most used sets for experiments in information retrieval, such as text categorization and clustering.

The text contained in the documents can be found in the form of messages, specifically emails, that contain elements such as signatures, greetings and in some cases HTML, therefore creating a need for additional pruning of the input. The good thing about this set is a great number of named entities contained in the documents, making it a good candidate for this paper.

The dataset is characterized with a good balance in the frequency of used categories. Even so a subset of the 20000 documents is used, reducing it to 18821 text files. The theme with the least assigned documents has a total of 628 documents, while the most prominent category counts 999 documents.

The standard approach to splitting the set to training and test subsets is taken from literature and implies a "by date" split, giving 11293 older documents for training and 7528 newer ones for testing. This kind of a split is intuitively a sound decision providing a great way to test the predictive properties of trained classifiers.

C. WebKB

WebKB, or the "World Wide Knowledge Base" project, for its task had the collection of web pages related to different fields of computer science, hosted by specific universities. The project was started in 1997 and results in a total of 8282 web pages from the mentioned sources. The data gathered from 4 different universities were manually classified in 7 different categories, including a miscellaneous category that adds a fair amount of noise to the dataset.

The set is mostly well balanced, with the exception of two categories "staff" and "department", that have respectively 137 and 182 documents. The very next category with the lowest document count is "project", counting 504 documents. For this reason, following [4], the "staff" and "department" categories, as well as the miscellaneous class are being excluded for analysis.

The literature does not provide a standard train and test split for this dataset. Therefore a typical split of randomly selected documents in a 80 to 20 percent ratio is applied on the WebKB dataset. This split provides a total count of 3357 training and 840 test documents.

V. METHODOLOGY

Standard algorithms for document dimensionality reduction by feature selection are applied before training any classifiers, and frequencies are normalized through the use of tf-idf weighting factors. The evaluation measures used are based on the datasets themselves since they vary in category balance. This chapter contains an overview of methods for text processing that are applied on all of the documents in a dataset.

A. Text processing

Non-structured text obtained from the documents is not suitable for ML algorithms and needs to be transformed. For that reason every document is represented as a n-dimensional array, representing the bag of words model. In this way it is possible to quantify the attributes of each document.

1) Text tokenization

To acquire the attributes of a text, i.e. form the vocabulary for a dataset, each document is split into n-grams [8]. Because this research requires NER, tokenization is conducted in conjunction with POS analysis. This kind of analysis is tasked with finding n-grams, or sequences of words with a specific function in a sentence. As the result, a tree structure is formed that describes the parts of a sentence with a joined function label [9].

After forming the tree, and labeling every part of the sentence as nouns (NNP or NN), verbs (VBD), and similar, the NER system starts labeling the n-grams that correspond to named entities. This kind of labeling is done in steps by traversing the tree structure in a way that allows for complete n-grams to be declared as named entities. If a term is not deemed as an entity, than it is taken as a regular, key word, token.

The described method of tokenization is used only in the cases where named entities are included in classifier training.

2) Lematization

The raw words gained through text tokenization are still not in the best shape for classifier training. They can be found in different tenses, noun cases, and such. The bag of words model, by itself does not hold any knowledge of grammatical constructs, so additional transformations should be applied.

The terms acquired through tokenization, for the purposes of this paper, can be categorized as either key words, which are always made out of single words, and named entities that can be found in the form of n-grams.

Words from the first group are reduced down to their lema [10], or "dictionary" form. For example words "walking", "walks", "walked" are all transformed in to "walk". With this the total vocabulary is reduced greatly, and grammatical constructs are ignored.

3) Stop words removal

Removing the stop words is a usual step in text processing, words that do not hold meaning by themselves and are generally found in any text, such as "a", "the", "again", "even" and many others. Aside from these, since the source for many of the datasets are web sites, the text is cleaned of all HTML tags that might occur.

B. Entity Power Coefficient

This work builds on some of the findings made by Gui, Yaocheng et al [1]. The researchers had great results when applying named entities in conjunction with other document features, on forming a clustering model for the Reuters-21578 dataset. They apply named entities by implementing a coefficient $\alpha \in \mathbb{R} \wedge \alpha \in [0,1]$ that indicates the impact that named entities have on making a decision. Here $\alpha = 0$ indicates the complete absence of named entities and $\alpha = 1$ means that only named entities are used while forming the model. In [1] the authors claim to have gotten the best performance with α close to 1, thereby justifying the use of named entities on this particular problem.

This paper takes the conclusions made in [1] and implements a coefficient $\alpha \in \mathbb{R} \wedge \alpha \in [0, \infty)$ that indicates the impact of named entities, in later text referred to as EPC (*Entity Power Coefficient).* The coefficient is used for creating virtual occurrences of all terms that are found to be named entities (including n-grams). The coefficient multiplies the term count found in all documents thereby modifying the real bag of words model.

VI. Experimental evaluation

The evaluation of the proposed classification method is run in three different settings for each of the 4 described datasets, namely the two subsets of Reuters-21578 (R8 and R52), 20 Newsgroups and WebKB. The classifiers trained consist of the three standard ML algorithms for text classification: Naïve Bayes, kNN (K Nearest Neighbors) and SVM (Support Vector Machine).

Each of these classifiers is trained for every dataset provided, in three different settings. The first setting named "BASE" results in a reference model that is trained without any regard to named entities. The second setting, named "NE" uses the proposed EPC and trains multiple models based on different values for the coefficient. The "NE_ONLY" setting uses only named entities without paying any regard to other tokens during training and testing. Evaluation metrics used are accuracy, recall, precision and f-measure with micro-averaging of the results. For each of the three settings, k best terms are picked through *chi-square* dimensionality reduction.

A. Reuters-21578 (R8)

The R8 subset is processed in the manner described previously and shows results that correspond to the ones provided in [4] in table 1.

1) BASE

Evaluation measures for the BASE setting on R8 dataset are shown on table 2, with a vocabulary totaling 21680 terms.

Table 2 shows the micro-averaged evaluation measure values. The average values for specific categories do not fluctuate greatly in comparison to the global average. Both Naïve Bayes and SVM classifiears see similar performance to the one provided in [4], with slight divergence that can be attributed to certain difference in text processing and tokenization techniques. Also Naïve Bayes and kNN show better performance for cases with a smaller number of attributes. NB even shows degradation in performance when increasing the attribute count. SVM works best when using all of the attributes in the vocabulary.

2) NE

Figure 1 displays the change corresponding to the various values of EPC in the range from 0 to 10. The number of terms in the vocabulary is 29767, and is seeing an increase brought in by the occurrence of named entities.

Table 4 shows the best measures in regard to α. The last column displays the number of terms gained through *chi-square* and the total number of named entities figuring in that subset.

The acquired results show a variation between the algorithms themselves when deciding on the best case for EPC. All algorithms show at best the similar performance as the BASE setting, although for different EPC values. For example the best performance for NB is shown to be for $\alpha \sim 8$ and SVM $\alpha \sim 1.7$. NB shows an oscillation in precision and a noticeable increase in recall that follows EPC.

3) NE ONLY

The last evaluation is performed for R8 is on a bag of words model containing only named entities and no key words.

TABLE II.
R8 BASE RESULTS

Classifier	Accuracy	Precision	Recall	F1	k
NB	0.9123	0.915	0.9123	0.91186	4896
SVM	0.9320	0.9324	0.9320	0.9304	21680
kNN	0.91	0.9084	0.91	0.907	754

TABLE III.
R8 NE ONLY RESULTS

Classifier	Accuracy	Precision	Recall	F1	k
NB	0.6958	0.6939	0.6958	0.6888	6020
SVM	0.6973	0.6975	0.6973	0.6877	8276
kNN	0.5389	0.5520	0.5389	0.4939	241

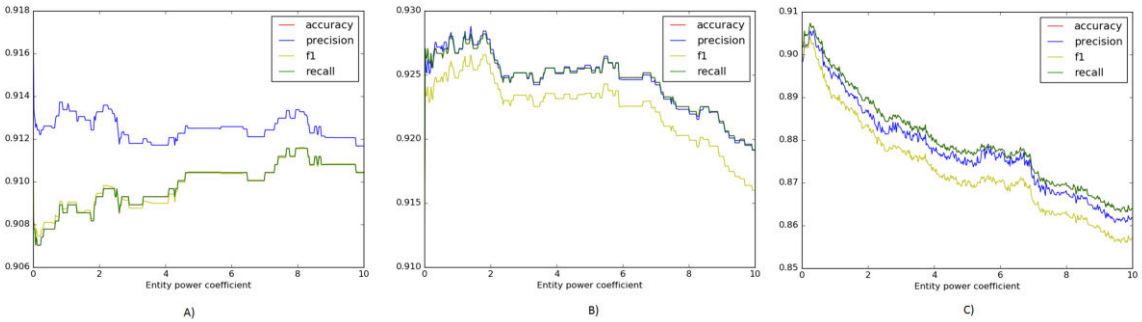

Figure 1 – Results for R8 NE setting A) Naïve Bayes, B) SVM, C) KNN

TABLE IV.
R8 NE RESULTS

Classifier	α	Accuracy	Precision	Recall	F1	k/ent
NB	7.877	0.9121	0.9147	0.9121	0.9115	699/105
SVM	1.7337	0.9263	0.9258	0.92630	0.9243	13885/3507
kNN	0.2506	0.9074	0.9055	0.9074	0.9042	699/105

When taking only named entities into account, there is a significant deterioration in performance. The total number of words in the vocabulary in this case is 14055 which is significantly less than in previous cases. Table 3 displays the results for this setting.

B. *Reuters-21578 (R52)*

The R52 dataset represents an expansion of the R8 subset from previous chapter.

1) BASE

The total number of recognized words in the vocabulary for this subset and the referential setting is 24163. The results are show on Table 5.

TABLE V.
R52 NE ONLY RESULTS

Classifier	Accuracy	Precision	Recall	F1	k
NB	0.808	0.8167	0.808	0.7640	1072
SVM	0.8429	0.8435	0.8429	0.8344	11524
kNN	0.7783	0,7450	0.7783	0.7446	1072

The R52 set shows a certain decrease in performance in brought by added categories.

2) NE

The results gained with classifiers that are aware of named entities and use EPC follow the trends observer in R8. Figure 2 displays how the evaluation measures change based on the EPC in the range of 0 to 10. The best values for accuracy and the corresponding EPC are show in table 6. The number of vocabulary words is 31363 which is significantly more than the last case.

The SVM classifier shows the optimal accuracy for

TABLE VI.
R52 NE RESULTS

Classifier	α	Accuracy	Precision	Recall	F1	k/ent
NB	0.4	0.7823	0.7923	0.7823	0.7656	1363/487
SVM	1.5789	0.8429	0.8413	0.8429	0.8357	17513/6127
kNN	0	0.76	0.7273	0.76	0.7285	1679/594

TABLE VII.
R52 NE ONLY RESULTS

Classifier	Accuracy	Precision	Recall	F1	k
NB	0.6023	0.5797	0.6023	0.5806	7845
SVM	0.598	0.5834	0.598	0.5763	15267
kNN	0.4432	0.3740	0.4432	0.3859	423

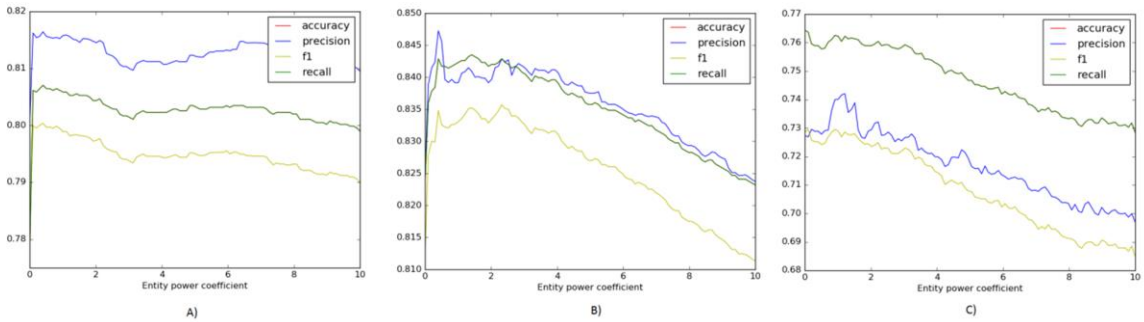

Figure 2 – Results for R52 NE setting A) Naïve Bayes, B) SVM, C) KNN

EPC around 1.5, but precision maximum is achieved for $\alpha = 0.47$. Since R52 shows a greater imbalance in category distribution, precision is taken as a lead measure, which shows that better results are achieved for smaller values of EPC.

3) NE ONLY

In the case where classifiers are trained solely with regard to named entities, 16074 words are recognized and added to the vocabulary. Table 7 displays the results and again shows a noticeable fall in performance.

C. 20 Newsgroups (20_NG)

This dataset id characterized with a greater number of documents that the R8 and R52 sets. Also the Newsgroups set is a collection of e-mail messages as opposed to Reuters' news articles. The same classifiers are trained just like in former cases, starting with the base setting.

1) BASE

The referential classifier for 20_NG dataset shows results very close to the ones provided in [4]. 110120 distinct terms have been found and added to the vocabulary. The results for the referential classifier are shown in table 8.

2) NE

Regarding classifiers that are made aware of named entities, the performance does not trail far from the referential results. On the other hand the value for EPC is higher for better performance. Figure 3 shows the results for the three classifiers trained on a total of 125238 terms including named entities. Table 9 shows the best results gained.

3) NE ONLY

Much like in previous cases with the Reuters dataset, classifiers trained on solely named entities show the worst performance. The total number of vocabulary entries is 40736 for this configuration as is shown in table 10.

D. WebKB

The Worldwide Knowledge Base dataset is the last set on which the evaluation is being performed. The documents in this dataset are web pages from four universities. Standard steps and configurations apply.

1) BASE

The referential Naïve Bayes classifiers is failing to reach the results provided in [4], while SVM remains in the neighborhood of those results. The total number of words without any NER amounts to 35731. Table 11 contains the results of this analysis.

2) NE

TABLE VIII.
20 NEWSGROUPS BASE RESULTS

Classifier	Accuracy	Precision	Recall	F1	k
NB	0.8141	0.8162	0.8141	0.8116	59000
SVM	0.8107	0.8117	0.8107	0.8023	110120
kNN	0.72	0.7388	0.7236	0.7192	74558

TABLE IX.
20 NEWSGROUPS NE RESULTS

Classifier	α	Accuracy	Precision	Recall	F1	k/ent
NB	2.2068	0.8162	0.8194	0.8162	0.8138	125238/32997
SVM	1.9298	0.8095	0.8107	0.8095	0.8082	123974/32828
kNN	1.011	0.7113	0.7280	0.7113	0.7070	80887/20261

TABLE X.
20 NEWSGROUPS NE ONLY RESULTS

Classifier	Accuracy	Precision	Recall	F1	k
NB	0.6341	0.6541	0.6341	0.6331	25032
SVM	0.6195	0.6479	0.6195	0.6273	45732
kNN	0.5312	0.5412	0.5312	0.5124	3590

TABLE XI.
WEBKB BASE RESULTS

Classifier	Accuracy	Precision	Recall	F1	k
NB	0.8583	0.8576	0.8583	0.8574	100
SVM	0.8724	0.885	0.8724	0.8701	100
kNN	0.7476	0.7642	0.7474	0.7339	1540

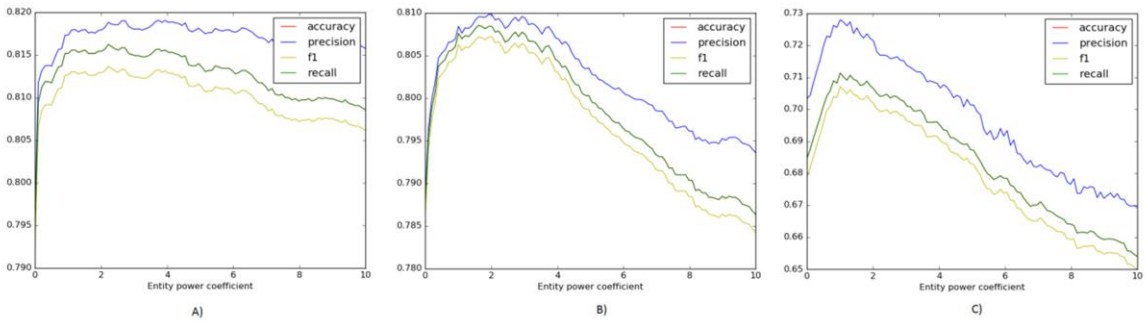

Figure 3 – Results for 20 Newsgroups NE setting A) Naïve Bayes, B) SVM, C) KNN

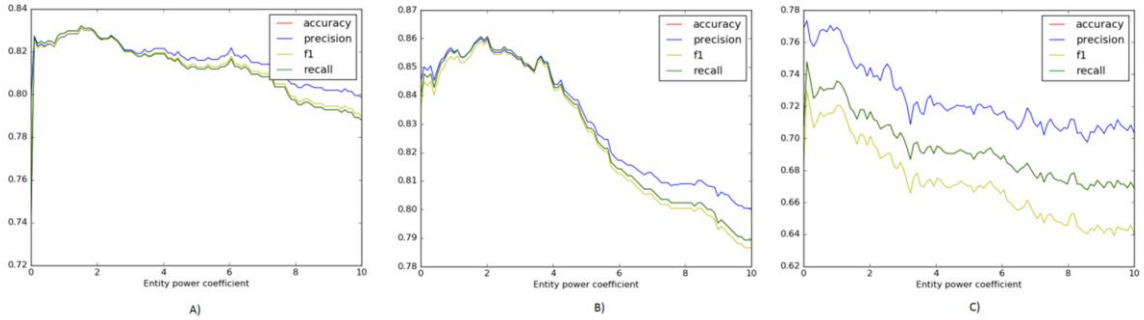

Figure 4 – Results for WebKB NE setting A) Naïve Bayes, B) SVM, C) KNN

TABLE XII.
WEBKB NE RESULTS

Classifier	α	Accuracy	Precision	Recall	F1	k/ent
NB	1.5151	0.8321	0.8310	0.8321	0.8311	100/10
SVM	1.6792	0.8571	0.8572	0.8571	0.8553	51613/25735
kNN	0.1010	0.7476	0.7733	0.7476	0.7298	1162/153

TABLE XIII.
WEBKB NE ONLY RESULTS

Classifier	Accuracy	Precision	Recall	F1	k
NB	0.6357	06454	0.6357	0.5735	26983
SVM	0.6059	06454	0.6059	0.5735	100
kNN	0.4928	0.5894	0.4928	0.4379	100

Classifiers do no show noticeable change in performance in the case of entity aware models. The total number of words in the vocabulary is 52675 and the results show optimal values in table 12. Figure 4 displays the change of evaluation measures following the growth of EPC in the range from 0 to 10.

3) NE ONLY

When taking into account only named entities, the classifiers again show the worst performance, on a vocabulary of 29346 terms. Table 13 shows the acquired results.

WebKB datasets displays the biggest fluctuation in accuracy when focusing on the SVM classifier, and shows the most significant downgrade of performance when regarding only named entities.

VII. CONCLUSION AND FURTHER WORK

After evaluation, a trend can be noticed that follows all four datasets. In general even for the best cases and optimal EPC, the results are at best in the range of the referential results. However we point out that the optimal value for EPC varies from algorithm to algorithm, as well as between datasets.

The results gained in this research do confirm that there is no real benefit in using named entities in any measure, while training classifiers on the listed datasets. However this can be the result of bad NER *(Named Entity Recognition)* as well as ignoring any entity resolution in this work. The work provides a good testing platform for other datasets, which could include social media data such as Twitter or Facebook posts, in later research. The social media could provide beneficial to a better NER that relies on the structured nature of online social media.

REFERENCES

[1] Gui, Yaocheng et al., *Hierarchical Text Classification for News Articles Based on Named Entities*.: Advanced Data Mining and Applications, 2012.

[2] Nick Latourette and Hugh Cunningham, *Classification of News Articles Using Named Entities with Named Entity Recognition by Neural Network.*

[3] S Montalvo, R Martinez, A Casillas, and V Fresno, *Bilingual news clustering using named entities and fuzzy similarity*.: International Conference on Text, Speech and Dialogue, September 2007.

[4] Ana Margarida de Jesus Cardoso Cachopo, "Improving Methods for Single-label Text Categorization," in *PhD diss., Universidade Técnica de Lisboa*, 2007.

[5] Ken Lang, "20 Newsgroups," in *http://qwone.com/~jason/20Newsgroups/*, last accessed. August 2016.

[6] Carnegie Group Inc and Reuters Ltd, "Reuters-21578," in *http://www.daviddlewis.com/resources/testcollections/reuters21578/*, last accessed August 2016.

[7] CMU text learning group, "WebKB - World Wide Knowledge Base," in *http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/*, last accessed August 2016.

[8] Daniel Martin Jurafsky and James H., "Speech and Language Processing," September 2014.

[9] Steven Bird, Ewan Klein, and Edward Loper, "Natural Language Processing with Python," 2014.

[10] Christopher D. Manning, Prabhakar Raghavan, and Heinrich Schütze, "Introduction to Information Retrieval," *Cambridge University Press*, 2008.