Discriminating Between Closely Related Languages on Twitter

Nikola Ljubešić
University of Zagreb, Faculty of Humanities and Social Sciences, Ivana Lučića 3
E-mail: nikola.ljubesic@ffzg.hr, http://nlp.ffzg.hr/

Denis Kranjčić
University of Zagreb, Faculty of Humanities and Social Sciences, Ivana Lučića 3
E-mail: dkranjcic@ffzg.hr

Keywords: microblogging, language identification, closely related languages

Received: October 25, 2014

In this paper we tackle the problem of discriminating Twitter users by the language they tweet in, taking into account very similar South-Slavic languages – Bosnian, Croatian, Montenegrin and Serbian. We apply the supervised machine learning approach by annotating a subset of 500 users from an existing Twitter collection by the language the users primarily tweet in. We show that by using a simple bag-of-words model, univariate feature selection, 320 strongest features and a standard classifier, we reach user classification accuracy of ∼98%. Annotating the whole 63,160 users strong Twitter collection with the best performing classifier and visualizing it on a map via tweet geo-information, we produce a Twitter language map which clearly depicts the robustness of the classifier.

Povzetek: V prispevku raziščemo problem ločevanja uporabnikov družabnega omrežja Twitter glede na to, v katerem jeziku tvitajo, pri čemer obravnavamo zelo podobne južnoslovanske jezike: bosanščino, hrvaščino, srbščino in črnogorščino. Uporabimo pristop nadzorovanega strojnega učenja, kjer označimo vsakega uporabnika iz že obstoječe podatkovne množice 500 uporabnikov z jezikom, v katerem največ tvita. Pokažemo, da z uporabo enostavnega modela vreče besed, univariantno izbiro značilk, 320 najbolj pomembnih značilk in standardnim klasifikatorjem, dosežemo ∼97 % točnost klasifikacije posameznega uporabnika. Če uporabimo najboljši razviti klasifikator za označevanje naše celotne zbirke, ki zajema 63.160 uporabnikov, in rezultat prikažemo na zemljevidu z uporabo geografske informacija na tvitih, smo izdelali Twitter zemljevid jezikov, ki jasno pokaže robustnost razvitega pristopa.

1 Introduction

The problem of language identification, which was considered a solved task for some time now, has recently gained in popularity among researchers by identifying more complex subproblems, such as discriminating between language varieties (very similar languages and dialects), identifying languages in multi-language documents, code-switching (alternating between two or more languages) and identifying language in non-standard user-generated content which often tends to be very short (such as tweets).

In this paper we address the first and the last problem, namely discriminating between very similar languages in Twitter posts, with the relaxation that we do not identify language on the tweet level, but the user level.

The four languages we focus on here, namely Bosnian, Croatian, Montenegrin and Serbian, belong to the South Slavic group of languages and are all very similar to each other.

All the languages, except Montenegrin, use the same phonemic inventory, and they are all based on the write-as-you-speak principle. Croatian is slightly different in this respect, because it does not transcribe foreign words and proper nouns, as the others do. Moreover, due to the fairly recent standardization of Montenegrin, its additional phonemes are extremely rarely represented in writing, especially in informal usage. The Serbian language is the only one where both Ekavian and Ijekavian pronunciation and writing are standardized and widely used, while all the other languages use Ijekavian variants as a standard. The languages share a great deal of the same vocabulary, and some words differ only in a single phoneme / grapheme, because of phonological, morphological and etymological circumstances. There are some grammatical differences regarding phonology, morphology and syntax, but they are arguably scarce and they barely influence mutual intelligibility. The distinction between the four languages is based on the grounds of establishing a national identity, rather than on prominently different linguistic features.

2 Related work

One of the first studies incorporating similar languages in a language identification setting was that of [9] who, among

others, discriminate between Spanish and Catalan with the accuracy of up to 99% by using second order character-level Markov models. In [11] a semi-supervised model is presented to distinguish between Indonesian and Malay by using frequency and rank of character trigrams derived from the most frequent words in each language, lists of exclusive words, and the format of numbers. [3] use a bag-of-words approach to classify Chinese texts from the mainland and Taiwan with results of up to 92% accuracy. [13] propose a log-likelihood estimation method along with Laplace smoothing to identify two varieties of Portuguese (Brazilian and European) obtaining 99.5% accuracy.

In the first attempt at discriminating between the two most distant out of the four languages of interest, namely Croatian and Serbian, [6] have shown that by using a second-order character Markov chain and a list of forbidden words, the two languages can be differentiated with a very high accuracy of $\sim$ 99%. As a follow-up, [12] add Bosnian to the language list showing that most off-the-shelf tools are in no way capable of solving this problem, while their approach by identifying blacklisted words reaches the accuracy of $\sim$97%. [11] have worked with the same three languages as a subtask of producing web corpora of these languages. They have managed to outperform the best-performing classifier from [12] by training unigram language models on the entire content of the collected web corpora, decreasing the error related to the Croatian–Serbian language pair to a fourth. Recently, as a part of the DSL (Discriminating between Similar Languages) 2014 shared task of discriminating between six groups of similar languages on the sentence level [14], the language group A consisted of Bosnian, Croatian and Serbian and the best result in the group yielded 93.6% accuracy, which is not directly comparable to the aforementioned results because classification was performed on the sentence level, and not on the document level as in previous research.

Language identification on Twitter data has become a popular problem in recent years. [1] use language identification to create language specific Twitter collections of low-resource languages such as Nepali, Urdu, and Ukrainian. [2] use character n-gram distance with additional microblogging characteristics such as the language profile of a user, the content of an attached hyperlink, the language profile of mentioned users and the language profile of a hashtag. [7] review a wide range of off-the-shelf tools for Twitter language identification, and achieve their best results with a simple voting over three systems.

To the best of our knowledge, there has been only two attempts at discriminating between languages of high level of similarity on Twitter data. The first attempt dealt with Croatian and Serbian [4], where word unigram language models built from Croatian and Serbian web corpora were used in an attempt to divide users from a Twitter collection according to the two languages. An analysis of the annotation results showed that there is a substantial Twitter activity of Bosnian and Montenegrin speakers in the collection and that the the collected data cannot be described with a two-language classification schema, but rather with a 4-class schema that includes the remaining two languages. The second attempt focused on Spanish varieties spoken in five different countries [8] using geo-information as a gold standard, obtaining best results with a voting meta-classifier approach that combines the results of four single classifiers.

Our work builds on top of the research presented in [4] by defining a four-language classification schema, inside which Montenegrin, a language that gained official status in 2007, is present for the first time. Additionally, this is the first focused attempt at discriminating between those languages on Twitter data.

3 Dataset

The dataset we run our experiments on consists of tweets produced by 500 randomly picked users from the Twitter collection obtained with the TweetCat tool described in [4]. This Twitter collection consists currently of 63,160 users and 42,744,935 tweets. The collection procedure is still running which opens the possibility of the collection becoming a monitor corpus of user-generated content of the four languages.

For annotating the dataset there was only one annotator available. Annotating a portion of the dataset by multiple users and inspecting inter-annotator agreement is considered to be future work.

Having other languages in the dataset (mostly English) was tolerated as long as more than 50% of the text was written in the annotated language. Among the 500 users there were 10 users who did not comply to any of the four classes and were therefore removed from the dataset. One user, tweeting in Bosnian, had most of the tweets in English, there was one user tweeting in Macedonian and 8 users were tweeting in Serbian, but used the Cyrillic script. The users tweeting in Serbian and using the Cyrillic script were discarded from the dataset because we wanted to focus on discriminating between the four languages based on content and not the script used.

The result of the annotation procedure is summarized in the distribution of users according to their language, presented in Table 1. We can observe that Serbian makes up 77% of the dataset. There is a similar amount, around 9%, of Bosnian and Croatian data, while Montenegrin is least represented with around 5% of the data. These results are somewhat surprising because there is a much higher number of speakers of Croatian (around 5 million) than of Bosnian (around 2 million) or Montenegrin (below 1 million). Additionally, Croatia has the highest GDP of all the countries and one would expect that the adaptation rate of such new technology should be higher and not lower than in the remaining countries.

Because we plan to discriminate between the four languages on the user level, we are naturally interested in the amount of textual data we have at disposal for each in-

language	instance #	percentage
Bosnian (bs)	45	9.18%
Croatian (hr)	42	8.57%
Montenegrin (me)	25	5.10%
Serbian (sr)	378	77.14%

Table 1: Distribution of users by the language they tweet in.

stance, i.e. user. Figure 1 represents the amount of data available per user, measured in the number of words. The plotted distribution has the minimum at 561 words and the maximum at 29,246 words, whereas the arithmetic mean lies on 6,607 words. This distribution shows that we have quite a large amount of textual data available for the majority of users. We will inspect the impact of data available for predicting the language in Section 4.5.

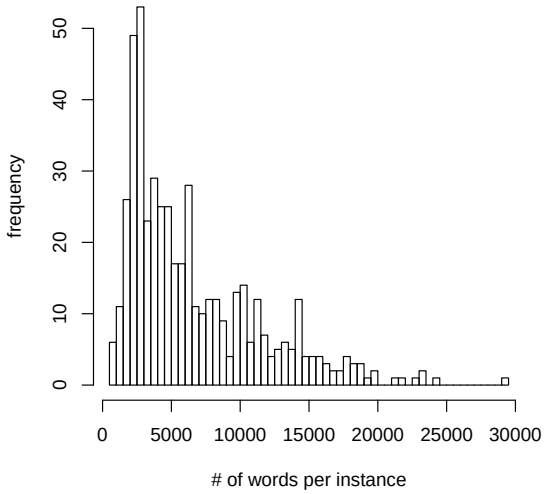

Figure 1: Distribution of dataset instances given the size in number of words.

4 Experiments

We perform data preprocessing, feature extraction and data formatting using simple Python scripts. All the machine learning experiments are carried out with scikit-learn [10]. Our evaluation metric, if not stated otherwise, is accuracy calculated via stratified 10-fold cross-validation.

We extract our features only from the text of the tweets. Using geolocation and user metadata (such as name, bio and location) is considered future work.

We experiment with the following preprocessing procedures:

- no preprocessing
- filtering out mentions, hashtags and URLs (making the data more representative of the user-generated content in general)
- dediacritizing the text (thereby lowering data sparsity)

and the following sets of features:

- words
- character 3-grams
- character 6-grams
- words and character 6-grams

Because no significant difference in accuracy was observed when using either different preprocessing procedures or sets of features (except for a slight drop when using character 3-grams), in the remainder of this section we present the results obtained by filtering out mentions, hashtags and URL-s and using words as features. By skipping dediacritization we keep the preprocessing level to a minimum, while by using words as features we ensure easy understandability of procedures such as feature selection. Finally, by removing textual specificities of Twitter like mentions and hashtags we ensure maximum applicability of the resulting models to other user-generated content besides tweets.

4.1 Initial experiment

The aim of the initial experiment was to get a feeling for the problem at hand by experimenting with various classifiers and features.

We experiment with traditional classifiers, such as the multinomial Naive Bayes (MultinomialNB), K-nearest neighbors (KNeighbors), decision tree (DecisionTree) and linear support-vector machine (LinearSVM). We use the linear SVM because the number of features is much greater than the number of instances. For each classifier we use the default hyperparameter values except for the linear SVM classifier for which we tune the C hyperparameter for highest accuracy.

classifier	accuracy $\pm$ stdev
DecisionTree	0.896 ± 0.026
KNeighbors	0.772 ± 0.040
LinearSVM	0.884 ± 0.034
MultinomialNB	0.806 ± 0.029

Table 2: Accuracy with standard deviation obtained with different classifiers using all words as features.

In the results presented in Table 2 we can observe that the LinearSVM and DecisionTree produce the highest accuracy. The significantly lower accuracy of the MultinomialNB classifier, which normally gives state-of-the-art results on bag-of-words models, but which has no inherent feature selection, provokes us to hypothesize that our results could improve if we applied explicit feature selection on our data. This follows our intuition that similar

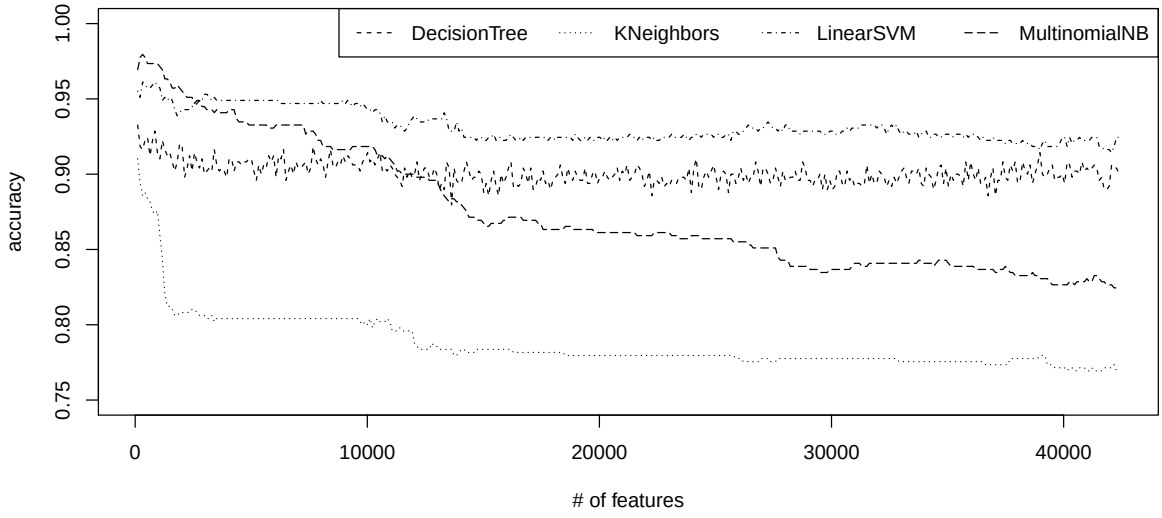

Figure 2: Classification accuracy as a function of number of most informative features used.

languages can be discriminated through a limited number of features, i.e. words, and not through the whole lexicon, which is normally shared to a great extent among such closely related languages.

4.2 Feature selection

Although there are stronger feature selection algorithms, we opt for a simple univariate feature selection algorithm which calculates p-value for each feature regarding the response variable through the F1 ANOVA statistical test. Finally it simply returns the user-specified number (or percentage) of features with lowest p-values. We use this simple feature selection method because we assume independence of our features, i.e. tokens or character n-grams, which is a reasonable assumption for language identification.

classifier	# of feats	acc $\pm$ stdev
DecisionTree	100	0.927 ± 0.019
KNeighbors	100	0.911 ± 0.041
LinearSVM	320	0.961 ± 0.025
MultinomialNB	320	$\mathbf{0.980 \pm 0.016}$

Table 3: Maximum accuracy obtained with each classifier with the number of strongest features used.

During these experiments we calculate accuracy via 10-fold cross-validation, performing feature selection each time on 90% of data used for model estimation.

The results of experimenting with up to 20% (cca. 42,000) of strongest word features are shown in Figure 2. Here we can observe a series of properties of the classifiers used. First of all, LinearSVM and DecisionTree, having implicit feature selection / weighting, operate similarly on the whole scale of number of features available, but still show better performance when using only a few hundred strongest features. On the other hand, MultinomialNB and KNeighbors show significantly better performance when they have to deal with the strongest features only. The best results are obtained with the MultinomialNB classifier at 320 features, reaching the accuracy of 97.97%. A numerical comparison of the best results obtained with the four classifiers is given in Table 3.

We present more detailed results obtained with the best-performing MultinomialNB classifier, trained on 320 features, in Table 4. It contains the confusion matrix of the classification process along with precision, recall and F1 obtained on each class. We can observe that the classification process is most successful on Serbian and Croatian, while the worst results are obtained on Montenegrin, which gets confused with both Bosnian and Serbian.

	bs	hr	me	sr	P	R	F1
bs	42	0	3	0	0.95	0.93	0.94
hr	1	41	0	0	0.98	0.98	0.98
me	0	0	23	2	0.82	0.92	0.87
sr	1	1	2	374	0.99	0.99	0.99

Table 4: Confusion matrix and precision, recall and F1 per class on the best performing classifier.

4.3 Evaluation on the test set

To perform a final test of our best performing classifier we produced an independent test set consisting of 101 annotated users. The MultinomialNB classifier, trained on all 490 users available from our development set, with 320

strongest features identified on that dataset, produces accuracy of 99.0%, having just one Bosnian user identified as Montenegrin. This experiment emphasizes the robustness of our classifier.

4.4 Analysis of the selected features

Using words as features, and not character 6-grams that perform equally well, enables us to easily interpret our final model. In Table 5 we present a systematization of the 320 features selected on the whole development set by language and the linguistic type of feature.

	bs	hr	me	sr
yat reflex	40.8	42.2	44.4	11.3
phonological	6.0	29.3	9.0	2.3
lexical	6.0	48.6	9.8	2.6
orthography	7.5	7.5	2.0	0.0
toponym, cultural	5.0	19.0	26.0	0.0
sum	65.3	146.8	91.3	16.2

Table 5: Feature type distribution across languages.

The features are divided into five categories across the four languages: yat reflex, phonological differences, lexical differences, orthography and toponym or cultural differences. Each feature contributes one point to the table: if a feature is present in more than one language, this point is divided among languages, and if a feature belongs to more than one feature type, the point is divided among those feature types. Almost half of the features belong to the "reflex of yat" category, which is least informative because most of the Ijekavian features are equally present in Croatian, Bosnian and Montenegrin. The exceptions are the words that are distinct both by the "reflex of yat" category and the lexical category, and few examples of Montenegrin-specific reflex of yat in words such as "nijesam" or "đe" (which also belongs to the "phonological differences" category). The "phonological differences" category contactins a lot of words present only in Croatian, such as "itko", "kava" or "večer" ("iko", "kafa" and "veče" in the other three languages). On the other hand, words that differ in only one phoneme and are not specific for Croatian are often spread among the remaining three languages. The category of lexical differences is similar in this respect: more than 70 percent of these features are Croatian. This can be explained by the fact that lexical purism is much more pronounced in Croatian than in the other three languages, which can be observed in the names of the months and some everyday words, such as "obitelj" (family), "glazba" (music), "izbornik" (menu) etc. In place of these words, Bosnian, Montenegrin and Serbian use words with evident foreign origin: "familija" (family), "muzika" (music), "meni" (menu) etc. The category of "orthography" predominantly contains infinitive verb forms without the final "i" letter, which appear in the future tense in Croatian orthography and which are also allowed in Bosnian. Finally, there is the category containing toponyms and culturally-specific items, such as country and city acronyms, names for residents, currency, TV-stations and even some public figures.

Although the features in the table are divided according to their real distribution among the languages, their distribution in the model sometimes differs. The reason for this is a significant difference between Croatian users and their language on the one side, and the rest on the other. Whereas Bosnian, Montenegrin and Serbian users are predominantly young people who use Twitter for chatting and sharing their everyday experiences, Croatian users are frequently news portals, shops, musicians, politicians etc. Consequentially, Croatian language on Twitter is marked by a much more formal register compared to the casual register of the other languages in our model.

4.5 Impact of amount of data available for prediction

Having a test set at our disposal opened the possibility of performing one additional experiment on the impact of the amount of data available for our language predictions. In our test set the user with the least amount of textual material contains 864 words. Therefore we evaluated the classifier trained on the whole development set by using only first N words from each user in the test set, N ranging from 10 to 850.

In Figure 3 we present the obtained results, representing each language with an F1 curve and all the languages with a micro-F1 curve. We can observe that the results peak and stabilize as we have 470 words at disposal for our prediction. This is an interesting result, showing that the large amount of data we have available for each user is actually not necessary. On the other hand, the results show quite clearly that discriminating between these languages on the level of each tweet would, at least with the presented classifier, be impossible given that the average tweet size is 10 words. Having significantly more training data available for each language could make a tweet-level classification possible since for Serbian, which covers 77% of the training data, on 10 words we already obtain a decent F1 of 0.88.

5 Corpus annotation and visualization

To be able to distribute separate Twitter collections of the four languages, we annotated each of the 63,160 users from our Twitter collection of Bosnian, Croatian, Montenegrin and Serbian. The annotation was performed with the MultinomialNB classifier trained on both the 490 development and the 101 testing instances, again selecting the 320 strongest features on that dataset.

Once we had our collection annotated, we decided to present the result of our language discriminator on a map.

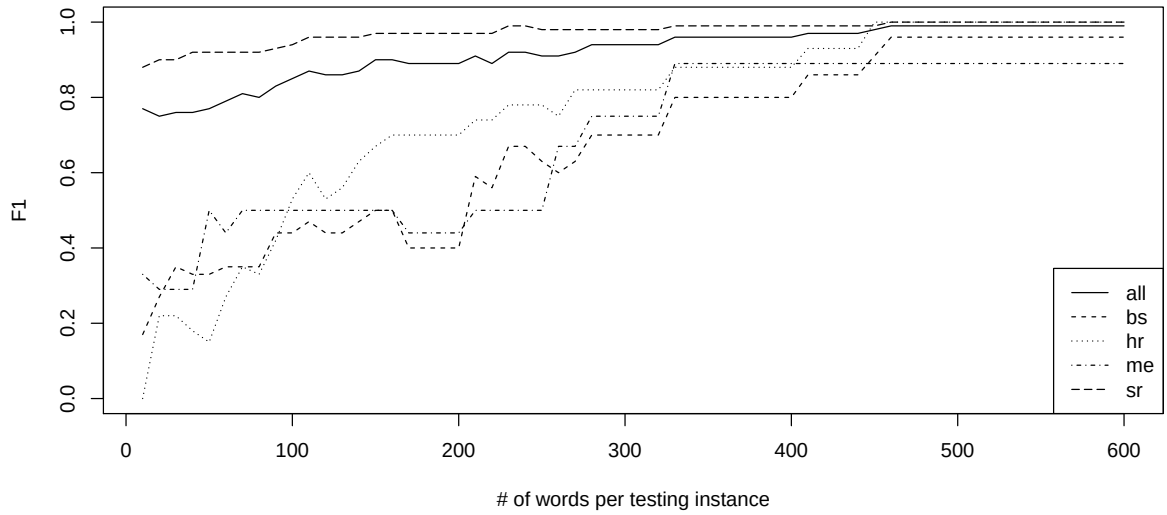

Figure 3: F1 per specific language as a function of the length of the testing instances.

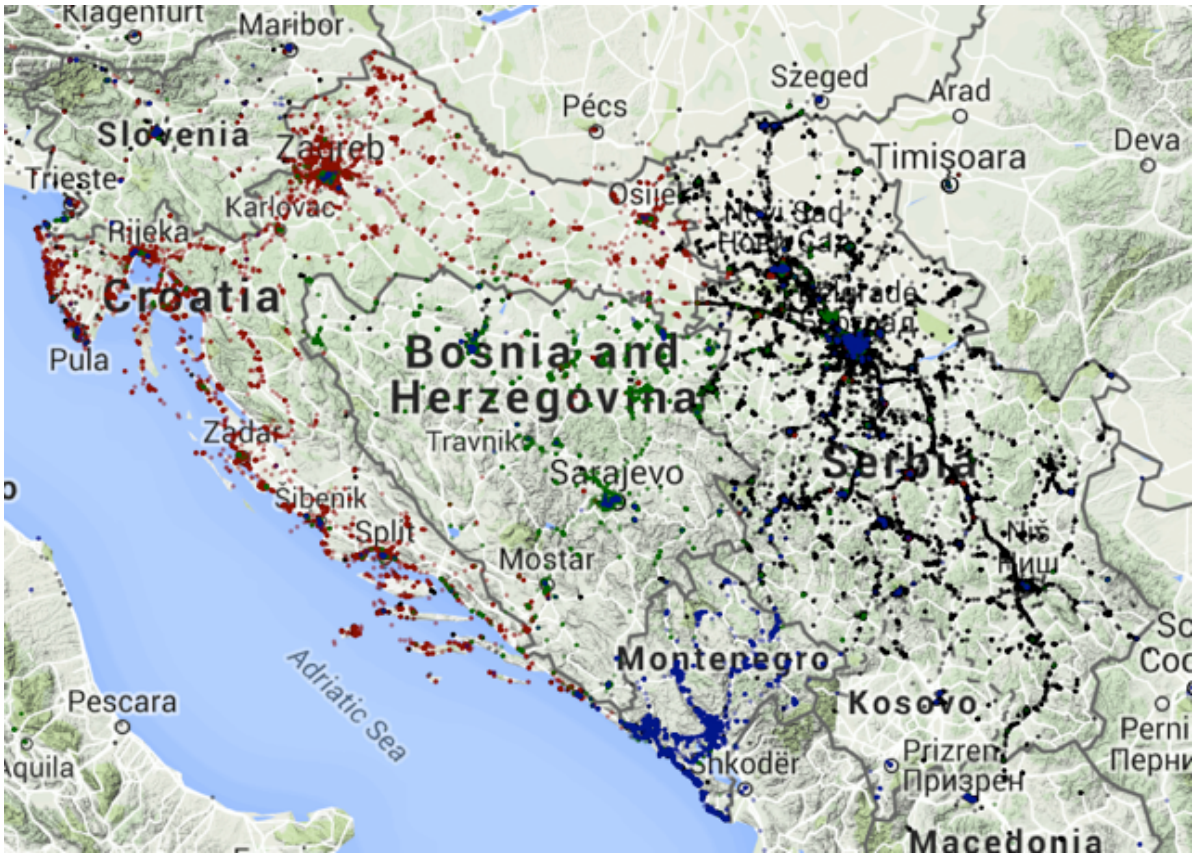

Figure 4: The Twitter language map. Impact of amount of data.

We presented each of the 576,786 tweets having geolocation available as a point on the map, encoding the predicted language of the author of the tweet with a corresponding color. We call the map presented in Figure 4 a "Twitter language map".

The Figure shows the area of the four countries in which the four languages have official status. We can observe that the tweets follow quite consistently the country borders, which is an additional argument that our classifier works properly. From the plot we can also confirm that Twitter is much more popular and widespread in Serbia than in the remaining countries. Mixing of the four languages occurs, as one would expect, mostly in big cities, primarily Belgrade, the capital of Serbia. There we can observe a significant number of Montenegrin speaking Twitter users. To perform a sanity check regarding the correctness of these data, we manually inspected ten random users classified as being Montenegrin and tweeting in the wider Belgrade area. The inspection showed that all ten users actually tweet in Montenegrin.

Overall, we can observe that Croatia and Serbia have a higher amount of foreign-tweeting users which is easily explained by the well-known migrations from Bosnia to both Croatia and Serbia, and from Montenegro primarily to Serbia.

6 Conclusion

In this paper we have presented a straight-forward approach to discriminating between closely related languages of Twitter users by training a classifier on a dataset of 490 manually labeled users. By using the bag-of-words model, 320 strongest features regarding univariate feature selection and the multinomial Naive Bayes classifier, we obtained a very good accuracy of 97.97% on the development set and 99.0% on the test set. Best results were obtained on Croatian and Serbian while most errors occurred when identifying the Montenegrin language.

Analyzing the impact of data available for classification showed that classification accuracy stabilizes at $\sim$470 words per user which still does not enable us to use this classifier on the tweet level.

Finally we annotated the whole 63k-user-strong collection of tweets and presented the collection on a map we call the "Twitter language map". The map shows that the language used on Twitter quite precisely follows the country borders, large cities being an exception to this rule.

Future work includes adding more information to our model besides words from tweets. Strongest candidates are the content to which users link and user meta-information such as username, location and bio. Using the geolocation information from tweets when available is surely a good source of information as well. Additionally, using the geolocation information as our response variable, i.e. redefining our task as predicting the location of a Twitter user is also a very interesting line of research. This surely increases the complexity of the task, but opens the door towards identifying dialects and sociolects.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

References

[1] Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., and Wilson, T. (2012). Language identification for creating language-specific twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 65–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

[2] Carter, S., Weerkamp, W., and Tsagkias, M. (2013). Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

[3] Huang, C.-R. and Lee, L.-H. (2008). Contrastive approach towards text source classification based on top-bag-of-word similarity. In *PACLIC*, pages 404–410. De La Salle University (DLSU), Manila, Philippines.

[4] Ljubešić, N., Fišer, D., and Erjavec, T. (2014). TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[11] Ljubešić, N. and Klubička, F. (2014). {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

[6] Ljubešić, N., Mikelić, N., and Boras, D. (2007). Language identification: How to distinguish similar languages. In Lužar-Stifter, V. and Hljuz Dobrić, V., editors, *Proceedings of the 29th International Conference on Information Technology Interfaces*, pages 541–546, Zagreb. SRCE University Computing Centre.

[7] Lui, M. and Baldwin, T. (2014). Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25. Association for Computational Linguistics.

[8] Maier, W. and Gómez-Rodríguez, C. (2014). Language variety identification in spanish tweets. In *Proceedings*

of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants, pages 25–35, Doha, Qatar. Association for Computational Linguistics.

[9] Padró, L. and Padró, M. (2004). Comparing methods for language identification. *Procesamiento del Lenguaje Natural*, 33:155–162.

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[11] Ranaivo-Malancon, B. (2006). Automatic Identification of Close Languages – Case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology*, 2(2):126–134.

[12] Tiedemann, J. and Ljubešić, N. (2012). Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

[13] Zampieri, M. and Gebre, B. G. (2012). Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012 - The 11th Conference on Natural Language Processing*.

[14] Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A Report on the DSL Shared Task 2014. In *Proceedings of the VARDIAL workshop*.