Discriminating beetwen Brazilian and European Portuguese national varieties on Twitter texts

Dayvid Castro, Ellen Souza
UAST - UFRPE
Serra Talhada, PE - Brazil
{dayvid.welles, ellen.ramos}@ufrpe.br

Adriano L. I. de Oliveira
CIn - UFPE
Recife, PE - Brazil
alio@cin.ufpe.br

***Abstract*—Twitter is one of the most used social media with users generating about 1 million messages per day. As a result of the expansion of this microblog, there is a diversity of languages used by users and many studies aimed at identifying the language of tweets. The third most used language on Twitter is Portuguese, a pluricentric language with two national standard varieties: Brazilian Portuguese and European Portuguese. Identifying a language variety may positively impact various Natural Language Processing tasks, but accomplishing this task is still regarded as one of the bottlenecks in this area, especially when combined with another bottleneck, language identification applied to short texts. Thus, given these challenges, this paper provides a current view on the automatic discrimination of the two main Portuguese language varieties on Twitter texts by using an acknowledged approach with different techniques and features in order to get an optimum configuration to fit our problem. Results reached 0.9271 for accuracy using an ensemble method, which combines character 6-grams and word unigrams and bigrams.**

Keywords—Language Identification; National Language Varieties; Portuguese Language; Twitter; Text Classification

I. INTRODUCTION

The volume of data generated in the digital universe is growing exponentially. It is in this context that the impact of social media is present, generating a massive amount of data, which is complex to analyze as it is represented in an unstructured format (e.g. Twitter feeds or Facebook posts) [1].

Twitter is one of the most used social media currently with 310 million monthly active users, generating about 1 billion messages per day [2]. As a result of this expansion on a global scale, there is a diversity of the languages used by the micro-blog users. Despite the prevalence of the English language, other languages such as Japanese, Spanish and Portuguese already represent 49% of all microblog posts[3]. Due to this evident diversity, several studies have been developed to identify the language of Twitter messages, as discussed in [4].

Identification of the language in which a text is written is a key task that precedes several Natural Language Processing tasks. Some applications, such as Information Retrieval and Text Mining, might not be able to process data efficiently without a basic knowledge of the document's language [5].

Although the language identification (LID) is widely accepted as a solved task in a usual configuration, there is a bottleneck when it comes to detecting the variety of a language such as Portuguese [6], the second most used language in Twitter [7] with two standard national varieties: Brazilian and European Portuguese [8]. A pluricentric language such as Portuguese presents varieties that differ subtly from each other in different linguistic levels such as lexicon, syntax, and orthographic etc. [8], which makes it hard for conventional tools to recognize these tenuous differences [9]. On the other hand, encouragingly, facing this problem we can possibly increase the accuracy of NLP tasks such as POS Tagging, Spell Checking and Machine Translation by using the background information about the language variant to apply more specialized approaches [10]. Apart from this, another bottleneck in the language identification task is dealing with short texts [4], which is exactly the type of text present on Twitter. The reason as described by [10] is that the LID becomes more difficult as we decrease the length of documents.

Given the challenges in the discrimination of language varieties dealing with very short texts, the goal of this study is to provide a current view on the automatic discrimination of Portuguese language varieties in Twitter texts by using an acknowledged approach with different techniques and features in order to get an optimum configuration to fit our problem.

In this work, we aim to detect the language variant from texts retrieved from Twitter, published in 2016 from Brazil and Portugal. We used a statistical method based on smoothed n-gram models and Log-Likelihood Ratio (LLR) [12]. We explore models with different granularities, sizes and smoothing techniques. Moreover, we test a simple and robust ensemble method named Mean Probability Rule [13].

This paper is structured as follows. Section II discusses related work on automatic language detection focusing on language varieties. Section III describes the method used in this study. Section IV presents our experimental setup. In Section V, we report results. Section VI presents the discussion. The final section shows the conclusion.

II. RELATED WORK

One of the first studies that explored the problem of identifying varieties of a language was done by Silva and Lopes [14], which through a statistical approach using Quadratic Discrimination Score and character n-grams they achieved an accuracy of 98.37% in discriminating European and Brazilian documents. In another study, this time dealing with similar languages, the best result reached a recall of 99.01% and

precision of 93.23% in differentiating Croatian from Serbian and Slovenian using a Markov second-order model classification and the rule of Forbidden words[5]. Huang and Lee [15] use a bag-of-words based approach to classify three varieties of Chinese, obtaining accuracy over 92%.

In 2012, Zampieri and Gebre [16] used a statistical approach based on the Log-Likelihood Ratio method and n-gram language models with Laplace Smoothing to identify two varieties of Portuguese (Brazilian and European). In this research, experiments with two journalistic corpora reached 99.8% for accuracy using character 4-grams models. Given these excellent results, the same authors applying this approach in different scenarios, such as varieties of Spanish [17], culminating in the development of *VarClass* [9]. This tool is the first language identification system with a focus on language varieties. *VarClass* report an average accuracy over 90.5% in a real-world setting containing 27 languages in which 10 of them are varieties of a language or similar languages.

Recently, with the growing interest in discriminating similar languages or varieties, the 2014 edition of COLING, the 25th International Conference on Computational Linguistics, held the first edition of the shared task Discriminating between Similar Languages (DSL), organized within the scope of the workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial) [10]. The DSL 2014 provided data from 13 different languages and varieties divided into 6 groups, among these the group D consisted of Brazilian and European Portuguese documents. In this group, the best result achieved a 95.6 performance using two-stage system composed of probabilistic document classifier and linear Support Vector Machines[18]. An overview of the DSL Shared Task 2015 can be found in [19].

To the best of our knowledge, there has been only one study working with discrimination between Portuguese national varieties on Twitter data [20]. On their experiments, for each of the countries (Brazil and Portugal) they selected 1400 users from a corpus containing tweets from 2011. [20] shows that their system is able to detect the nationality of a user with 95% accuracy by using their own methods: Entities, Word tokens, Grammar, and URL features. Unlike what we present here, the work in [20] is focused on identifying the nationality of Twitter users, therefore, their classification method is performed at the user level. Our research tries to pass through the problem of discriminating between Portuguese language varieties on tweet level. Besides, in order to avoid costly and error-prone generation and maintenance manual of features vocabulary-based, we chose to use only n-gram language model based features, since we are motivated by the desire to preserve simplicity and scalability of our system.

III. METHODS

In this section, we describe our data collection method and the techniques we used in our experiments to address the problem of discriminating Portuguese national on Twitter data. Since LID is a special case of text classification problem [6], we apply a supervised learning approach based on a smoothed n-gram model and the LLR estimation method. This approach was previously tested in a scenario of discrimination variants of Portuguese applied to documents [16] and later adapted to a real-world language identification system called *VarClass* – discussed in section II. The algorithm used in our experiments is a modified version of the *VarClass* (which uses Laplace smoothing), built using the programming language Python and the python library NLTK. We perform some variations in classification schema in order to identify the best configuration for our experiments with very short texts. For the n-gram models, we tested several standard smoothing techniques as well as different sizes and granularities. Moreover, we also experiment our approach in an ensemble learning perspective through the Mean Probability Rule method.

A. Dataset

In order to perform the Twitter data extraction, we used the Twitter Search API, which provides samples of the data flowing through the network from the previous 7 days. To access the Twitter API we use a simple python library called Tweepy. We collected Tweets from the first three months of 2016. The search was filtered by country so there were two extraction streams: one for Brazil and another for Portugal.

In sequence, we submitted both collections of tweets at a pre-processing stage with the following tasks:

1. Language detection to filter only Portuguese tweets.
2. Removal of meta-information, URLs and tags.
3. Removal of special words such as "RT" or "via".
4. Removal of punctuation characters as @, dots and commas.
5. Removal of tweets with only three or fewer words, to reduce unmeaningful tweets.

For the first preprocessing step, the language identification system Langid was used [21]. The basic consideration for building the corpora was that a tweet in Portuguese collected from Brazil was written in Brazilian Portuguese and equivalently, a tweet in Portuguese collected from Portugal was written in European Portuguese.

B. N-Gram Language Model

Concerning the linguistic level of the Portuguese Language, in this work, we use three types of n-gram language models:

1) Character n-grams (2 – 7): Models that focus on capturing orthographic differences.

2) Word uni-grams: Models with features that try to capture lexical differences.

3) Word bi-grams: Models focused to capture lexical-syntactical differences.

C. Smoothing Techniques

To improve the estimated probabilities of the language models and therefore increase the accuracy of the models as a whole [22], we applied several smoothing techniques available on NLTK library. We used the following smoothing techniques: Laplace (PL), Lidstone (LD), Witten-Bel (WB), Good-Turing (GT) and Kneser-Ney (KN). Each of these techniques has its own particularity, advantages, and disadvantages, but what they all have in common is that they tend to make more uniform probabilities distributions [22]. Details about these methods can

be found in [22]. In addition, a study of smoothing techniques for LID in the short texts is carried out in [23].

D. Log-Likelihood Ratio

This estimation method was proposed by Dunning[24] and has been widely used in several NLP applications [25][16][26][27] . Given the n-gram language models already built, we used the statistical method Log-Likelihood Ratio (1) to perform classification [12]. Following (1), **N** refers to the amount of n-grams of the text for testing, **L** refers to the language model and $\mathbf{n_i}$ is the **i**th n-gram. Given a test sample, the probability for each of the existing language models is calculated. In the end, the language model which generates the highest probability determines the text language [12]. In our context, the model with the highest probability determines the Portuguese language variety.

$$P(L \mid text) = \underset{L}{argmax} \sum_{i=1}^{N} \log P(n_i \mid L) + \log P(L) \quad (1)$$

E. Ensemble-Based Log-Likelihood Ratio

Based on the classifier ensemble approach [28], we built an ensemble system combining different n-gram language models to be used by our estimation method LLR to perform the classification of Portuguese varieties. The experiments were conducted with the combination method Mean Probability Rule (MPR). Following the idea of the MPR method, on our ensemble-based classifiers for each class the LLR of the models is added together. Then, the class label (Brazilian Portuguese x European Portuguese) with the highest probability average determines the language variety.

F. Evaluation

In an effort to validate the Portuguese language variant classifier, we employed 10-fold cross validation in all the experiments. The metrics accuracy (A), precision (P), recall (R), and F-Measure (F1) were applied to evaluate the results.

IV. Experimental Setup

The experiment sample contains 64,000 tweets divided equally among each national variant to ensure that training and test data are balanced. We compute a word average length of 4.55 for Brazilian variety and 4.49 for European variety. To validate the automatic approach used to built the corpus we manually inspect 10% of the data and as a result we verified the effectiveness of the method in collecting tweets in Portuguese from both regions. The corpus was submitted to four classification schemes using n-gram language models with different configurations to compute the LLR. For each n-gram model we tested all smoothing techniques mentioned in this study. The only exception is the Kneser-Ney smoothing that was applied only on tri-gram models due a limitation of NLTK. It is interesting to point out that the Lidstone is parameterized by a real number λ, which typically ranges from 0 to 1 [28] (when λ = 1 Lidstone is equivalent to Laplace smoothing). Therefore, we use three variations of Lidstone, with λ = 0.1, 0.2 and 0.5.

A. Model 1

In this model, web built the classification method by using n-gram language models from the character-level with size ranging from 2 to 7. As illustrated in Figure 1, here we have 37 possible configurations.

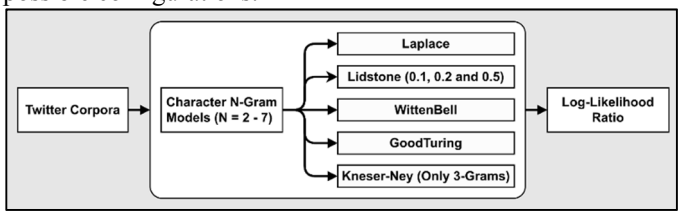

Fig. 1 Character N-Gram Models

B. Model 2

On the second model, the classification approach is built from the word unigrams models. As shown in Figure 2, here we have six possible configurations.

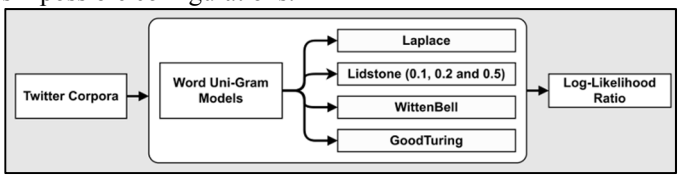

Fig. 2 Word Uni-Gram Models

C. Model 3

On the third model, we use word bigram models. As illustrated in Figure 3, we have six possible configurations.

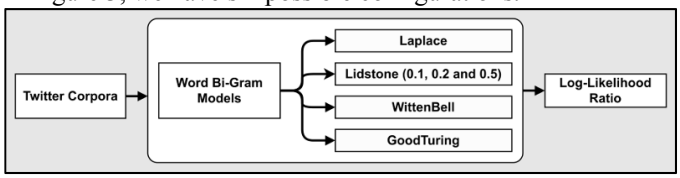

Fig. 3 Word Bi-Gram Models

D. Model 4

In this classification model, we conduct experiments using an ensemble of classifiers based on MPR. The selection of configurations used in combination is defined as the best results from each system, which are then submitted to several possible arrangements: character n-gram and word uni-gram; character n-gram and word bi-gram; word uni-gram and word bi-gram; and lastly, the three types of features together. On figure 4, we illustrate an ensemble learning classification combining three language models, as outlined on subsection III.E, after combine the output of each model using Mean Probability Rule, the decision about the language variety is made.

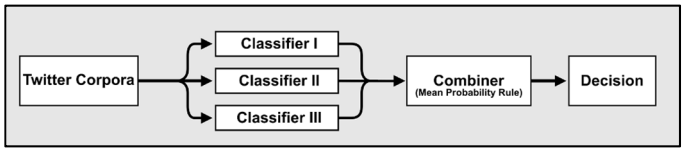

Fig. 4 Ensemble-Based Models

V. RESULTS

In this section, we present the results obtained according to the four models defined in the experimental setup. Using 10-fold-cross validation, we show the results in terms of average accuracy (acc.), precision (pre.), recall (rec.) and f-measure (f1) for all classes and for each class (Brazilian Portuguese x European Portuguese).

Table I presents the results of the experiments of *Model 1*, in which character n-gram models were applied aiming to capture orthographic level differences. The best result was achieved by the character 6-gram model with Lidstone ($\lambda = 0.1$) smoothing (configuration #11), followed respectively by models with the same smoothing technique but with different parameters, configurations #12 and #13. To enable a better view of the character n-gram models, Figure 1 shows the cross-validation accuracy of different classification methods as a function of n.

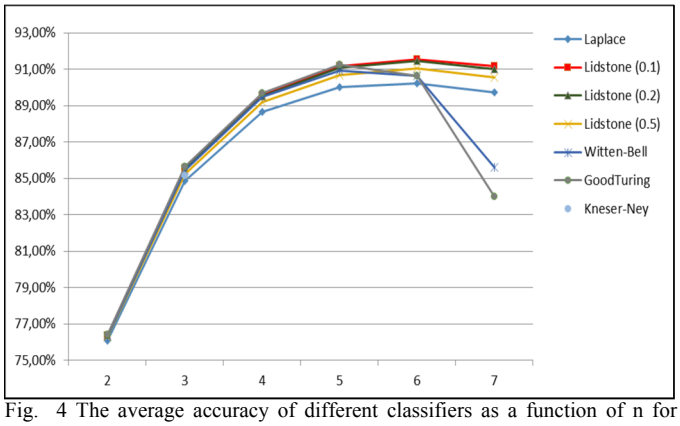

Fig. 4 The average accuracy of different classifiers as a function of n for character n-gram models.

Table II presents the results of *Model 2 and Model 3*, in which experiments with word unigrams and bigrams models were carried out in order to capture lexical and lexical-syntactical differences, respectively. Among the results with word n-gram models, only one model with the Lidstone smoothing is present, given that the other parameter variations of this technique showed no differences. For word uni-gram models (*Model 2*) the best result achieved was 90.43% accuracy when using the Good-Turing smoothing (configuration #44). Regarding the bi-gram models (*Model 3*), the configuration #43 reported the best result with an accuracy of 85.33% using an n-gram model smoothed by Witten-Bell technique.

Given the results obtained by the three first experiments, for *Model 4* we selected: Good-Turing Word Uni-Gram Model, Witten-Bell Word Bi-Gram Model and Lidstone Character 6-Gram Model. These three configurations had the best results in each of the respective models. From the selected configurations, we built different ensemble-based models whose results are expressed in Table III. The best result of the *Model 4*, 92.71% accuracy, was obtained when combining the three language models (configuration #46).

VI. DISCUSSION

In the following section, we present the main findings of our experiments. We evaluated the impact of four models with different configurations on discriminating two national language varieties of the Portuguese language. It seems that in most cases, it is easier to detect the Brazilian variety, as you can see, for example, configuration #5 obtained 93.55% accuracy for Brazilian Portuguese and 87.52% for European variant. This behavior might be related to the fact that European variety presents tweets with a lower average length, resulting in a minor performance. According to our results, the ensemble-based classification system combining three models perform better than all other configurations. As expected, using the ensemble method Mean Probability Rule we are able to improve our results over the single classifier system. This may be explained because the three models combined are able to capture differences in orthographic, lexical and syntactic levels.

Exploring the results of a viewpoint of granularity of models, we observe that higher order character n-grams models perform better than word n-grams models. Between the word-level n-gram models, uni-gram models showed better performance. From examining the findings, this leads us to an intuitive conclusion that the lexical differences proved stronger than the syntactical differences and there are certainly more lexical features in unigram models than in bigram models.

In examining the behavior of character n-gram models, in general, accuracy increase from lower order n-gram models to higher order n-gram models until an optimal parameter is reached. For Witten-Bell and Good-Turing smoothing, their optimum parameter are n-gram models with size of 5 while for Laplace and Lidstone smoothing the best results was achieved by 6-gram models. The models smoothed with Lidstone smoothing presented better results on this character-level. With regard to the variations of Lidstone smoothing ($\lambda = 0.1$, 0.2 and 0.5), a positive association between decreasing the parameter and a better outcome was observed. The difference becomes more significant when we added Laplace Smoothing to our analysis, since this technique is equivalent to Lidstone with $\lambda = 1$. In such a way, comparing Lidstone ($\lambda = 0.1$) 6-gram models with Laplace 6-gram models it is possible to verify an improvement from 0.9025 to 0.9156.

This study validates the fact that when it comes to dealing with very short texts we need to conduct deep investigations based on this domain. To illustrate this, we can observe that the work reported in [16], which uses the same approach but applied to documents, prove that character-based model using 4-grams is the best choice for discriminating Portuguese varieties. On the other hand, in our research with Twitter data, the best performance of our single classifier system was achieved by character 6-grams model. Besides, our investigations using different smoothing techniques, not only the Laplace used in [16], also had an important role in our research. Comparing our result using the equivalent configuration to work in [16] (configuration #3) with the result achieved using the best configuration identified for the single classifier system (configuration #6), we observe an increase of almost 3% by finding the best configuration for our Twitter

domain. Finally, when we compare our system with the best result obtained in [20], which also deals with Portuguese varieties on tweets but with an approach that include features created manually, we verified that our best classifier has only a 2.29% accuracy loss using a feature space built fully automatic.

TABLE I. RESULTS OF CHARACTER N-GRAM MODELS (N = 2 – 7)

Configuration			Overall Results				Brazilian			European		
ID	*N-Gram*	*Smoothing*	*Acc*	*Pre.*	*Rec.*	*F1*	*Pre.*	*Rec.*	*F1*	*Pre.*	*Rec.*	*F1*
#1	2	Laplace	0.7606	0.7683	0.7606	0.7644	0.8135	0.6763	0.7386	0.7231	0.8449	0.7792
#2	3		0.8487	0.8524	0.8487	0.8505	0.8883	0.7977	0.8405	0.8164	0.8997	0.8560
#3	4		0.8868	0.8888	0.8868	0.8878	0.9170	0.8505	0.8825	0.8606	0.9230	0.8907
#4	5		0.9001	0.9021	0.9001	0.9011	0.9307	0.8645	0.8964	0.8736	0.9356	0.9035
#5	6		**0.9025**	**0.9048**	**0.9025**	**0.9037**	**0.9355**	**0.8646**	**0.8986**	**0.8742**	**0.9404**	**0.9060**
#6	7		0.8973	0.9007	0.8973	0.8990	0.9377	0.8511	0.8923	0.8637	0.9435	0.9018
#7	2	Lidstone (0.1)	0.7638	0.7701	0.7638	0.7669	0.8113	0.6876	0.7443	0.7289	0.8400	0.7805
#8	3		0.8558	0.8577	0.8558	0.8568	0.8838	0.8194	0.8504	0.8317	0.8922	0.8609
#9	4		0.8958	0.8963	0.8958	0.8961	0.9094	0.8793	0.8941	0.8832	0.9123	0.8975
#10	5		0.9117	0.9119	0.9117	0.9118	0.9209	0.9008	0.9107	0.9030	0.9226	0.9127
#11	6		**0.9156**	**0.9158**	**0.9156**	**0.9157**	**0.9253**	**0.9042**	**0.9146**	**0.9064**	**0.9270**	**0.9166**
#12	7		0.9120	0.9123	0.9120	0.9121	0.9236	0.8983	0.9107	0.9010	0.9257	0.9131
#13	2	Lidstone (0.2)	0.7634	0.7700	0.7634	0.7667	0.8118	0.6860	0.7436	0.7282	0.8408	0.7804
#14	3		0.8550	0.8572	0.8550	0.8561	0.8848	0.8164	0.8492	0.8296	0.8937	0.8604
#15	4		0.8953	0.8959	0.8953	0.8956	0.9113	0.8759	0.8932	0.8806	0.9147	0.8973
#16	5		0.9109	0.9113	0.9109	0.9111	0.9230	0.8966	0.9096	0.8995	0.9252	0.9122
#17	6		**0.9145**	**0.9149**	**0.9145**	**0.9147**	**0.9275**	**0.8993**	**0.9132**	**0.9023**	**0.9297**	**0.9158**
#18	7		0.9101	0.9107	0.9101	0.9104	0.9262	0.8913	0.9084	0.8953	0.9289	0.9118
#19	2	Lidstone (0.5)	0.7626	0.7697	0.7626	0.7661	0.8131	0.6821	0.7418	0.7262	0.8432	0.7803
#20	3		0.8523	0.8550	0.8523	0.8536	0.8859	0.8087	0.8455	0.8241	0.8958	0.8584
#21	4		0.8918	0.8929	0.8918	0.8924	0.9134	0.8658	0.8889	0.8725	0.9179	0.8946
#22	5		0.9067	0.9075	0.9067	0.9071	0.9261	0.8839	0.9045	0.8890	0.9295	0.9088
#23	6		**0.9103**	**0.9114**	**0.9103**	**0.9109**	**0.9319**	**0.8854**	**0.9080**	**0.8909**	**0.9353**	**0.9125**
#24	7		0.9057	0.9071	0.9057	0.9064	0.9315	0.8758	0.9027	0.8828	0.9356	0.9084
#25	2	Witten-Bell	0.7631	0.7695	0.7631	0.7663	0.8111	0.6859	0.7433	0.7280	0.8402	0.7800
#26	3		0.8545	0.8566	0.8545	0.8556	0.8841	0.8159	0.8487	0.8292	0.8931	0.8599
#27	4		0.8949	0.8956	0.8949	0.8953	0.9114	0.8750	0.8928	0.8798	0.9149	0.8970
#28	5		**0.9091**	**0.9095**	**0.9091**	**0.9093**	**0.9211**	0.8949	**0.9078**	0.8979	0.9233	0.9104
#29	6		0.9065	0.9065	0.9065	0.9065	0.9110	**0.9009**	0.9060	**0.9020**	0.9120	0.9070
#30	7		0.8562	0.8581	0.8562	0.8572	0.8325	0.8920	0.8612	0.8837	0.8205	0.8509
#31	2	Good-Turing	0.7642	0.7699	0.7642	0.7670	0.8088	0.6919	0.7458	0.7309	0.8364	0.7801
#32	3		0.8564	0.8576	0.8564	0.8570	0.8785	0.8273	0.8521	0.8368	0.8856	0.8605
#33	4		0.8969	0.8970	0.8969	0.8970	0.9032	0.8891	0.8961	0.8908	0.9047	0.8977
#34	5		**0.9124**	**0.9124**	**0.9124**	**0.9124**	**0.9117**	**0.9133**	**0.9125**	**0.9132**	**0.9115**	**0.9123**
#35	6		0.9064	0.9066	0.9064	0.9065	0.9187	0.8858	0.9019	0.8897	0.9216	0.9054
#36	7		0.8398	0.8438	0.8398	0.8418	0.8070	0.8934	0.8480	0.8806	0.7862	0.8307
#37	3	Kneser-Ney	0.8516	0.8516	0.8516	0.8516	0.8505	0.8532	0.8518	0.8527	0.8501	0.8514

TABLE II. RESULTS OF WORD N-GRAM MODELS (N = 1 – 2)

Configuration			Average				Brazilian			European		
ID	*N-Gram*	*Smoothing*	*A*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
#38	1	Laplace	**0.8943**	**0.8964**	**0.8943**	**0.8953**	**0.9256**	**0.8574**	**0.8902**	**0.8672**	**0.9311**	**0.8980**
#39	2		0.8258	0.8378	0.8258	0.8318	0.9016	0.7313	0.8076	0.7741	0.9202	0.8408
#40	1	Lidstone (0.1)	**0.9037**	**0.9042**	**0.9037**	**0.9039**	**0.9187**	**0.8858**	**0.9019**	**0.8897**	**0.9216**	**0.9054**
#41	2		0.8381	0.8429	0.8381	0.8405	0.8835	0.7790	0.8280	0.8024	0.8973	0.8472
#42	1	Witten-Bell	**0.9032**	**0.9035**	**0.9032**	**0.9033**	**0.9130**	**0.8914**	**0.9020**	**0.8939**	**0.9150**	**0.9043**
#43	2		0.8533	0.8533	0.8533	0.8533	0.8534	0.8532	0.8533	0.8532	0.8535	0.8533
#44	1	Good-Turing	**0.9043**	**0.9043**	**0.9043**	**0.9043**	**0.9045**	**0.9041**	**0.9043**	**0.9042**	**0.9045**	**0.9043**
#45	2		0.8492	0.8498	0.8492	0.8495	0.8361	0.8688	0.8521	0.8634	0.8296	0.8462

TABLE III. RESULTS OF ENSEMBLE-BASED CLASSIFIERS USING THE AVERAGE PROBABILITY RULE METHOD

Configuration		Results				Brazilian			European		
ID	*Ensemble-Based Models*	*Acc*	*Pre.*	*Rec.*	*F1*	*Pre.*	*Rec.*	*F1*	*Pre.*	*Rec.*	*F1*
#46	Word 1-Gram GT; Word 2-Gram WB; Character 6-Gram LD (0.1)	**0.9271**	**0.9272**	**0.9271**	**0.9272**	**0.933**	0.9203	**0.9266**	0.9214	**0.9339**	**0.9276**
#47	Word 1-Gram GT; Word 2-Gram WB	0.9258	0.9258	0.9258	0.9258	0.9273	**0.9241**	0.9257	**0.9243**	0.9275	0.9259
#48	Word 1-Gram GT; Character 6-Gram LD (0.1)	0.9234	0.9236	0.9234	0.9235	0.9301	0.9157	0.9228	0.917	0.9312	0.924
#49	Word 2-Gram WB; Character 6-Gram LD (0.1)	0.921	0.9212	0.921	0.9211	0.9293	0.9113	0.9202	0.9131	0.9306	0.9217

VII. Conclusion

In this paper, we try to discriminate between the two standard national varieties of Portuguese on Twitter data. We address this problem as a text classification task at tweet level, applying a supervised learning method based on N-Gram Language Models and Log-Likelihood Ratio estimation method. In order to carry our investigation, a balanced corpus was built containing 64 thousand tweets written in Brazilian Portuguese and European Portuguese. We investigated different classification models, including ensemble models based on Mean Probability Rule. After testing several configurations for each model we identified the ensemble classification model combining Good-Turing smoothed word unigram model, Witten-Bell smoothed word bigram model and Lidstone smoothed character 6-gram model as the best performer (configuration #46). These outcomes imply that an arrangement of lexicon, syntax, and orthography features are strong and sufficient for detecting Portuguese varieties.

After we did the experiments on this study and obtained very satisfactory results, we feel encouraged to continue with the evolution of our system in order to use it for real-time applications. With our system we expected to positively affect the performance of various NLP tasks on social media from the use of the national variety information for applying specialized approaches. As a future work, we intend to apply feature selection methods to evaluate whether it can increase performance as well as the efficiency of our system due to reduction in feature space. Besides, we intend to test other advanced smoothing techniques and other ensemble-based methods in order to check if they can produce better results.

Acknowledgment

Ellen Souza is supported by FACEPE (IBPG-0765-1-0311).

References

[1] P. Zikopoulos, Big data beyond the hype. New York [u.a.]: Mc Graw Hill Education, 2015, p. 41.

[2] "Company," in *Twitter*, Twitter, 2016. [Online]. Available: https://about.twitter.com/company. Accessed: Feb. 6, 2016.

[3] A. Seshagiri, "The Languages of Twitter Users," 2014. [Online]. Available: http://bits.blogs.nytimes.com/2014/03/09/the-languages-of-twitter-users/. [Accessed: 03-Mar-2016].

[4] A. Zubiaga, I. S. Vicente, P. Gamallo, J. R. Pichel, I. Alegria, N. Aranberri, A. Ezeiza, and V. Fresno, "Overview of TweetLID: Tweet Language Identification at SEPLN 2014," in *Twitter Language Identification Workshop at SEPLN 2014*, 2014, pp. 1–11.

[5] N. Ljubesic, N. Mikelic and D. Boras, "Language Indentification: How to Distinguish Similar Languages?," in *Proceedings of 29th International Conference on Information Technology Interfaces*, Cavtat, 2007, pp. 541-546.

[6] M. Zampieri, "Using bag-of-words to distinguish similar languages: How efficient are they?," *Computational Intelligence and Informatics (CINTI), 2013 IEEE 14th International Symposium on*, Budapest, 2013, pp. 37-41.

[7] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes, "Do All Birds Tweet the Same?: Characterizing Twitter Around the World," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 2011, pp. 1025–1030.

[8] M. Clyne, *Pluricentric languages*. Berlin u.a.: Mouton de Gruyter, 1992, pp. 11-29.

[9] M. Zampieri and B. Gebre, "VarClass: An Open-source Language Identification Tool for Language Varieties," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

[10] M. Zampieri, L. Tan, N. Ljubešic, and J. Tiedemann, "A report on the DSL shared task 2014," in *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, 2014, pp. 58–67.

[11] T. Baldwin and M. Lui, "Language Identification: The Long and the Short of the Matter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 2010, pp. 229–237.

[12] T. Dunning, "Statistical identification of language." Computing Research Laboratory, New Mexico State University, 1994.

[13] S. Malmasi and M. Dras, "Language identification using classifier ensembles," in *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, 2015, p. 35.

[14] J. F. Da Silva and G. P. Lopes, "Identification of Document Language is Not yet a Completely Solved Problem," in *Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*, Sydney, NSW, 2006, pp. 212-212.

[15] C. R. Huang and L. H. Lee, "Contrastive approach towards text source classification based on top-bag-of-word similarity," *Proc. 22nd Pacific Asia Conf. Lang. Inf. Comput. PACLIC 22*, pp. 404–410, 2008.

[16] M. Zampieri and B. G. Gebre, "Automatic Identification of Language Varieties : The Case of Portuguese," *Proc. KONVENS 2012*, vol. 2012, no. 1994, pp. 233–237, 2012.

[17] M. Zampieri, B. G. Gebre, and S. Diwersy, "N-gram language models and POS distribution for the identification of Spanish varieties," Proc. TALN2013, pp. 580–587, 2013.

[18] C. Goutte, S. Léger, and M. Carpuat, "The NRC system for discriminating similar languages," in *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, 2014, pp. 139–145.

[19] M. Zampieri, L. Tan, and N. Ljubeˇ, "Overview of the DSL Shared Task 2015," no. 2014, 2015.

[20] G. Laboreiro and L. Dei, "Determining language variant in microblog messages," pp. 902–907.

[21] M. Lui and T. Baldwin, "langid.py: An Off-the-shelf Language Identification Tool," *Proc. 50th Annu. Meet. Assoc. Comput. Linguist.*, no. July, pp. 25–30, 2012.

[22] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Comput. Speech Lang.*, vol. 13, no. 4, pp. 359–394, 1999.

[23] T. Vatanen, J. V Jaakko, and S. Virpioja, "Language Identification of Short Text Segments with N-gram Models," *Lr. 2010*, pp. 3423–3430, 2010.

[24] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Comput. Linguist.*, vol. 19, pp. 61–74, 1993.

[25] P. Pantel and D. Lin, "A statistical corpus-based term extractor," in *Advances in Artificial Intelligence*, Springer, 2001, pp. 36–46.

[26] A. Gelbukh, G. Sidorov, E. Lavin-Villa, and L. Chanona-Hernandez, "Automatic term extraction using log-likelihood based comparison with general reference corpus," in *Natural Language Processing and Information Systems*, Springer, 2010, pp. 248–255.

[27] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *J. Informetr.*, vol. 3, no. 2, pp. 143–157, 2009.

[28] R. Polikar, C. Zhang, and Y. Ma, "Ensemble Machine Learning: Methods and Applications." Springer. Chapter: Ensemble Learning, 2012.

[29] "nltk Package — NLTK 3.0 documentation." [Online]. Available: http://www.nltk.org/api/nltk.html. [Accessed: 05-May-2016].