Text-based Language Identification for Some of the Under-resourced Languages of South Africa

Tshephisho Joseph Sefara, Madimetja Jonas Manamela and Promise Tshepiso Malatji
Telkom Center of Excellence for Speech Technology
Department of Computer Science, University of
Limpopo Polokwane, South Africa
sefaratj@gmail.com, Jonas.Manamela@ul.ac.za, tshepiso29@gmail.com

***Abstract*— Language identification is the problem of correctlyclassifying a sample of text/documents based on its language. However, much of the research work focused on the English language corpora and little research work focused on other South African official languages. In a multilingual society like South Africa, the use of automatic language identification in any language-specific system would be a vital step in bridging the digital divide between diverse members of the society. Various machine learning algorithms can be used to solve the problem of identifying the natural language of a document/text. This paper presents a text-based language identification using individual proper names, specifically surnames in a South African context. Three supervised machine learning methods are implemented to perform 3-way multiclass classification using support vector machines, and naïve Bayes language models. These algorithms are applied to the language identification task and evaluated in extensive experiments for three official languages of South Africa:**
***Tshivenda, Xitsonga and Sepedi.* All three machine learningmethods achieved remarkable results in a 10-fold cross validation. The results indicate that a multinomial naïve Bayes method achieved better performance than other algorithms.**

Keywords— support vector machines; multinomial naïve Bayes; WEKA; machine learning; text classification; language identification; multiclass classification

I. INTRODUCTION

The goal of language identification is to classify a document/text based on its language. An automatic language identification is a growing application of speech processing technology that has many practical uses. It can be used as a front-end system to a telecommunication company - routing a caller to an appropriate human emergency operator depending on the correct identification of the caller's language. Language identification is an important technology used in various fields of natural language processing (NLP) including machine translation, information extraction and retrieval, spell checkers, and question answering systems [1, 2].

Most language-specific systems such as automatic speech recognition and synthesis systems encounter the problem of accurately recognizing and/or synthesizing a person's proper name in different languages. Hence such systems require the integration of a front-end language identification module to identify the language(s) used [3, 4]. Language identification techniques assume that a given document is written in one of thepre-defined languages (class) for which there is a training data for a selected classifier selecting the most likely language from the class. Most NLP systems assume monolingual input data. The inclusion of new or foreign language(s) text may downgrade the performance of such trained systems.

In a multilingual environment like South Africa, the use of language identification is often needed to analyze speech or text for various NLP tasks including topic labelling, part-of-speech tagging, pronunciation prediction [5, 6], and stemming [1]. Among all spoken languages worldwide, the well-known global lingua franca, English, is frequently used as a universal language for communication. However, the use of under-resourced languages such as the South African official languages is a challenging task in text technology or mining [7]. Hence, there may be serious consequences and implications for not engaging more extensive research in language identification of under-resourced languages. Under-resourced languages are defined as languages that have limited presence on the internet, shortage of linguistic expertise, and little or no information technology available. Sometimes this concept is referred to as low-density languages, scarce-resourced languages, and limited data languages [7]. This research is focused on some under-resourced languages spoken in Limpopo province or South Africa.

Various machine learning methods can be used to solve the problem of identifying the natural language used in a document/text. There are many data analysis tools and techniques that can be used to achieve the required results [1-3]. The Waikato Environment for Knowledge Analysis (WEKA) toolkit [8] is a machine learning toolkit developed to run entirely on Java to facilitate the availability of data mining tools regardless of a computer platform. WEKA contains the collection of algorithms for data analysis, data visualization and predictive modelling, together with application programming interface for easy access from other Java-enabled integrated development environments. These algorithms include perceptron neural networks [9], decision trees [10], naïve Bayes classifiers [10], support vector machines (SVM) [11] and other filters and classification functions. WEKA also provides access to pre-processing scripts for conversion of text to vectors with a range of feature extraction options.

The goal of this paper is to use three supervised machine learning methods that are implemented to perform a 3-way multiclass classification, and to explore the language

978-1-5090-2576-3/16/$31.00 ©2016 IEEE

identification accuracy that can be achieved for selected under-resourced official languages of South Africa (namely, Sepedi, Xitsonga, and Tshivenda) using *n*-gram statistics as features. This accuracy relies on the multiclass classification algorithm employed. Hence, we investigate this factor by developing some classifiers based on *n*-gram statistics for Sepedi, Xitsonga, and Tshivenda. The rest of this paper is structured as follows. Section II presents related work in this area. In Section III, we discuss the equipment and methods used for the development of the proposed system. We discuss and analyse the results in Section IV, and conclude in Section V with remarks.

II. Related Work

Relatively little language technology research has been done on language identification of South African indigenous languages. The first research to investigate under-resourced languages is done in 2006 by Botha *et al.* [12]. Likelihood classifiers and SVMs used to examine the accuracy for all eleven official languages of South Africa using *n*-gram statistics of different units as features. Indhuja *et al*. [13] presented a text based language identification system for five Indian languages (Hindi, Sanskrit, Marathi, Nepali and Bhojpuri) based on Devanagari script. They used n-gram statistics as features. Their system achieved an average accuracy of 80% for 5 pair languages which is much less than other pairs. Hannan and Sarma [14] show the development of a text-based language identification of Assamese and Bodo languages. Their system works well with a small text files containing small number of words. But as the data increases the accuracy decrease for the Bodo language, and the Assamese language accuracy remains constant.

Llitjos and Black [5] showed that language identification could enhance the accuracy of letter-to-phoneme conversion. Language identification has been using character/word *n*-gram language models. *N*-gram approach has proven to be more accurate for identification of language in a script/document classification in [15]. Giwa and Davel [1] discuss factors that influence language identification accuracy of individual words for South African languages. They concluded that SVM with Radial Basis Function (RBF) outperformed naïve Bayes classifiers with Witten-Bell smoothing. Botha and Barnard [16] discuss various factors that affect text-based language identification accuracy. These factors include *n*-gram size, text input size, training data, and machine learning algorithm employed as well as language similarities. Fourie *et al*. [17] compared SVM and multinomial naïve Bayes (MNB) for named entity classification of English and Afrikaans. They used WEKA software to conduct the experiments using MNB and SVM algorithms. The implementation of the baseline SVM classifier was through Platt's Sequential Minimal Optimization algorithm. They converted the data with string-to-word vector filter whereby words in the data defined as classes, and strings converted to decimal arrays. They used 10-fold cross-validation and concluded that SVM performed better than MNB models across all granularity levels and both languages. Language identification experiments have adopted character *n*-gram models and demonstrated good results over a variety of applications [18]. The advantage of using *n*-gram statistic over other algorithms is that no linguistic knowledge needed to construct a classifier.

III. METHODOLOGY

In our experiment, we aim to identify the first-language of a person given the person's last name selected from some South African under-resourced official languages, namely, Sepedi, Xitsonga, and Tshivenda. To achieve this aim, we shall

- collect training text data such as person last names for three languages.
- compare and use the best machine learning algorithm to build a text-based language identification predictor for person names classification.

The language identification predictor will be used to classify an input person's last name by predicting the first-language of that person.

A. Data Pre-processing

The WEKA toolkit is used in conducting the experiments using the supplied implementations of the MNB, a library for support vector machine (LibSVM) [19] and a library for large linear classification (LibLINEAR) [20] were downloaded using WEKA package manager. The text data is obtained from the University of Limpopo students database consisting of person names. The data was converted with the help of WEKA toolkit, using a string-to-word vector filter. Every word is tokenized into character trigrams. The data contains 2400 of last names (instances) and three languages (classes) as shown in Table I.

TABLE I. Language Identification Data

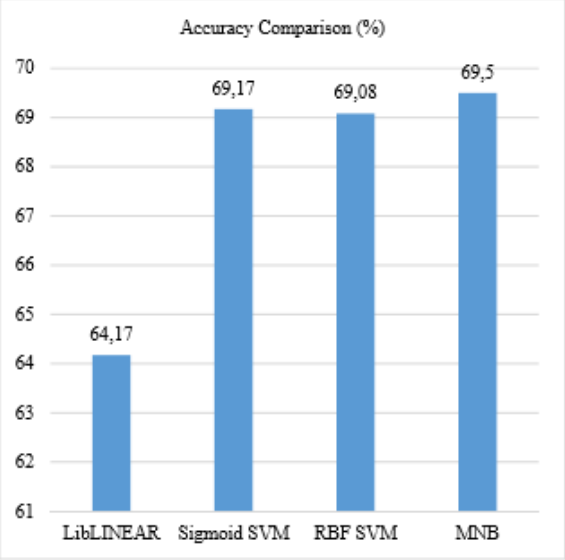
Languages	**Number of names**
Sepedi	800
Tshivenda	800
Xitsonga	800

B. Machine Learning Algorithms

Support vector machines are supervised learning models with associated learning algorithms that recognize patterns and analyse data for regression analysis and classification. The experiment employed MNB and the two SVM libraries namely: LibLINEAR and LibSVM for multiclass classification. LibSVM uses one-against-one classification method where a classifier is created for each pair of classes and classification is performed using a voting strategy. LibLINEAR uses one-against-all classification method where one SVM is trained for each class, and classification is performed using a voting strategy. The training of MNB and SVM is generated by combining trigrams for each language to create the training set. We performed 10-fold cross-validation on our training set. This approach partitions the corpus into ten equal parts and performs training on 9/10 and testing on 1/10 partition. This approach is repeated ten times so that each part is used for training. The SVM and MNB machine learning algorithms have shown to achieve reasonable classification results in [21, 22].

C. Evaluation

We evaluate the accuracy of each classifier model based on certain criteria to assess the performance of the classifier model. Language identification accuracy is measured as the proportion of correctly identified last names in the test set compared to all the person names in the same test set. Each test is characterized by the following parameters: training data size, language, and algorithm employed. The following evaluation measurements are used to evaluate the performance of the results: root mean squared error (RMSE), accuracy (A), used in Table II for precision, recall, accuracy and F1 measurements are shown in equations (1), (2), (3), and (4), where FP = False Positive, TN = True Negative, TP = True Positive and FN = False Negative classifications. Precision is defined as the percentage of applicable last names identified out of all identified last names whereas recall is defined as the percentage of applicable last names retrieved out of all applicable last names in the collection.



Actual or Model Class Ci		Actual class	
		Yes	No
Classifier class	Yes	TP	FP
	No	FN	TN

Accuracy Comparison (%)

Fig. 1: Results of the accuracy in percentage for 10-fold cross validation.\

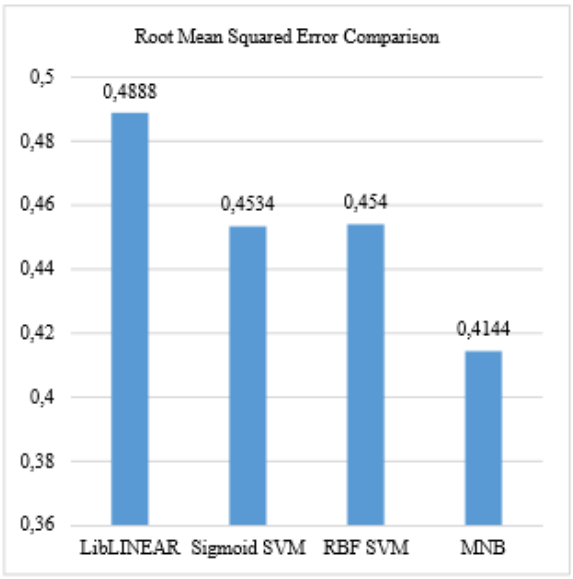


Fig. 2: Results of RMSE for 10-fold cross validation.

$$P = TPi / TPi + FNi \quad (1)$$

$$\Pi = TPi / TPi + FPi\ c \quad (2)$$

$$F1 = 2(P*\Pi) / (P + \Pi) \quad (3)$$

$$A = (TPi + TNi) / (TPi + TNi + FPi + FNi) \quad (4)$$

VI. Preliminary Results And Discussion

The objective of the experiment is to compare the performance of the algorithms using the RMSE, accuracy, recall, precision and F1 measurements. We trained the language identification predictor on 2400 names. The results of the precision (Π), recall (P) and F1 measurements. The formulas language identification accuracy are shown in Fig. 1. The accuracy figures show that Sigmoid SVM, RBF SVM and MNB algorithms have comparable performance results but MNB outperformed Sigmoid SVM, RBF SVM and libLINEAR define algorithm by difference of 0.33%, 0.42%, and 5.33% respectively. Furthermore, the language identification accuracy of LibLINEAR is very low with 64,17% which makes it the least appicable method on such data. LibLINEAR uses one-against-all classification method which compares one class (language) against other classes (languages) to create optimal hyperplanes. However, the creation of hyperplane may be difficult for closely related languages, and this can be solved by introducing kernels. As shown in Fig. 1 the Sigmoid and RBF SVM kernels performed better than LibLINEAR by a difference of 4.91%.

TABLE III. PRECISION, RECALL, AND F-MEASURE RESULTSPER LANGUAGE FORCLASSIFICATION ALGORITHM

	Class	Precision	Recall	F-Measure
SigmoidSVM	Sepedi	0.680	0.753	0.715
	Xitsonga	0.646	0.675	0.660
	Tshivenda	0.763	0.648	0.700
	Average	0.696	0.692	0.692
RBF SVM	Sepedi	0.682	0.746	0.713
	Xitsonga	0.643	0.678	0.660
	Tshivenda	0.761	0.649	0.700
	Average	0.695	0.691	0.691
LibLINEAR	Sepedi	0.658	0.665	0.661
	Xitsonga	0.604	0.615	0.609
	Tshivenda	0.665	0.645	0.655
	Average	0.642	0.642	0.642
MNB	Sepedi	0.679	0.768	0.721
	Xitsonga	0.668	0.665	0.666
	Tshivenda	0.747	0.653	0.696
	Average	0.698	0.695	0.694

Table III shows precision, recall, and F1 measure over 10-fold cross-validation for each language. Also, the table shows precision achieving its highest value (76.3%) for Tshivenda class, and obtained its lowest (60.4) for for Xitsonga class under Sigmoid SVM and Linear SVM algorithms respectively. On the other side, recall achieving its highest value (76.8%) for Sepedi class, and obtained its lowest alue (61.5%) for Xitsonga class under the MNB and Linear SVM algorithms respectively. The F1-measure supply a weighted harmonic mean between the recall and precision. The F1 measure of 0.694 for MNB, 0.692 for Sigmoid SVM, 0.691 for RBF SVM, and 0.642 for linear SVM is achieved. The MNB slightly outperformed other algorithms and achieved a difference of 0.2%. The resulting accuracy was 69% with SVM sigmoid kernel, 69% with SVM radial basis function kernel, and 64% with SVM linear kernel, and 70% with MNB language model as shown in Fig. 1. In this instance, the performance of the MNB algorithm was found to be significantly better than other methods.

V. Concluding Remarks And Future Work

This paper examined a simple language identification model based on character n-gram of three units. The model performed well in identifying unique person's last names. According to the results of the experiments, all algorithms performed better with MNB slightly outperforming other algorithms excluding LibLINEAR with very low classification accuracy. Moreover, the most applicable method is MNB in which high accuracy results were obtained across all languages in all evaluation criteria than Sigmoid SVM, RBF SVM and LibLINEAR SVM algorithms respectively.

This work intends to deliver an accurate language identification system for South African languages. The system may be effectively used to pre-select a language for language-specific systems such as automatic speech recognition and synthesis, and machine translation systems in specific under-resourced environment. The work enhances and elevates the recognition of indigenous South African official languages in the information and communication technologies. Such a development may contribute to minimizing the negative effects of the digital language divide on potential users of digital technology. Future research will try to extend the approach by examining and reducing the RMSEs.

VI. Acknowledgment

Any opinion, findings and conclusions or recommendations expressed in this material are those of the author(s) and therefore the sponsors do not accept any liability on them.

References

[1]. O. Giwa and M. H. Davel, "N-gram based language identification of individual words," in The 24th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), Johannesburg, 2013,15-21.

[2]. W. Pienaar and D. P. Snyman, "Spelling checker-based language identification for the eleven official South African languages," in

[3]. Proceedings of the twenty-first annual symposium of the pattern recognition association of South Africa (PRASA), Stellenbosch, 2010, 213-217.

[4]. K. R. Mabokela and M. J. Manamela, "An integrated language identification for code-switched speech using decoded-phonemes and support vector machine," in Speech Technology and Human - Computer Dialogue (SpeD), 2013 7th Conference on, Cluj-Napoca, 2013, pp. 1-6.

[5]. K. R. Mabokela, M. J. Manamela and M. Manaileng, "Modeling code-switching speech on under-resourced languages for

language identification," in Proc. of SLTU, Petersburg, Russia, 2014, pp. 225-230.
[6]. A. F. Llitjos and A. W. Black, "Knowledge of language origin improves pronunciation accuracy of proper names," in INTERSPEECH, Aalborg, 2001, pp. 1919-1922.
[7]. J. A. Badenhorst, D. R. Van Niekerk and E. Barnard, "Automatic systems for assistance in improving pronunciations," in Proceedings of the Seventeenth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), Parys, 2006, pp. 23-29.
[8]. L. Besacier, E. Barnard, A. Karpov and T. Schultz, "Automatic speech recognition for under-resourced languages: a survey," Speech Communication, vol. 56, pp. 85-100, 2014.
[9]. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update," SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10-18, November 2009.
[10]. A. Nicolaou, A. D. Bagdanov, L. Gomez-Bigorda and D. Karatzas, "Visual script and language identification," in 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, 2016, pp. 393-398.
[11]. M. Lazar, D. Militaru and E. Oancea, "Classifying the lexico-syntactic patterns of semantic relations between two nouns in Romanian language," in Speech Technology and Human - Computer Dialogue (SpeD), 2015 International Conference on, Bucharest, 2015, pp. 1-6.
[12]. M. Heck, S. Stuker and A. Waibel, "A hybrid phonotactic language identification system with an SVM back-end for simultaneous lecture translation," in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 4857-4860.
[13]. G. Botha, V. Zimu and E. Barnard, "Text-based language identification for the South African languages," in Proceedings of the 17th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), Parys, 2006, pp. 7-13.
[14]. K. Indhuja, M. Indu, C. Sreejith, P. C. Reghu Raj and P. C. Raj, "Text based language identification system for Indian languages following devanagiri script," International Journal of Engineering Research & Technology (IJERT), vol. 3, no. 4, pp. 327-331, 2014.
[15]. A. Hannan and S. K. Sarma, "Identification of Assamese and Bodo language from text - an approach," International Journal of Engineering Research & Technology (IJERT), vol. 4, no. 12, pp. 67-70, 2015.
[16]. M. K. Shukla, A. Rana and H. Banka, "Classification of the Bangla script document using SVM," in 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, India, 2016, pp.182-185.
[17]. G. R. Botha and E. Barnard, "Factors that affect the aaccuracy of text-based lagnuage identification," Computer Speech and Language, vol. 26, no. 5, pp. 307-320, 2012.
[18]. W. Fourie, J. V. Du Toit and D. P. Snyman, "Comparing support vector machine and multinomial naïve Bayes for named entity classification of South African languages," in Proceedings of the 2014 PRASA, RobMechand AfLaT International Joint Symposium, Cape Town, 2014, pp. 183-188.
[19]. I. Suzuki, Y. Mikami, A. Ohsato and Y. Chubachi, "A language and character set determination method based on N-gram statistics," ACMTransactions on Asian Language Information Processing (TALIP), vol.1, no. 3, pp. 269-278, 2002.
[20]. C.-C. Chang and C.-J. Lin, "LIBSVM : A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, pp. 1-27, 2011.
[21]. R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang and C. J. Lin, "LIBLINEAR: a library for large linear classification," *Journal of Machine Learning Research,* vol. 9, pp. 1871-1874, 2008.
[22]. T. Vatanen, J. J. Vayrynen and S. Virpioja, "Language identification of short text segments with N-gram models," in *Proc. LREC*, Valletta, Malta, 2010, pp. 3423-3430.
[23]. T. J. Sefara and M. J. Manamela, "The development of local synthetic voices for an automatic pronunciation assistant," in *Southern AfricaTelecommunication Networks and Applications Conference (SATNAC),*Georg e, 2016, pp. 142-146.