Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 7188 - 7199

Research Article

Automatic Identification of Major Text Language

Dr.M.Indhumathi

Abstract

Text-based language identification is the task of automatically recognizing a language from a given text of document. It is an important research area as large quantities of text are processed automatically for tasks such as spelling and grammar checking, information retrieval, search engines, language translation, and text mining. In this research, an adequate mechanism for efficient text-based language identification is presented with an emphasis on 7 major languages used in Ethiopia and India namely, Afar, Amharic, Nuer, Oromo, Sidamo, Somali and Tigrigna. These languages were chosen because they are spoken by more than 79.3% of the total population of India and Ethiopia. Factors affecting accuracy such as the size and variety of training data and the size of the string to be identified are investigated. Naïve Bayes classifier, SVM classifier and Dictionary Method are used. Naïve Bayes and SVM classifiers are trained by using character n-gram of size 3 as a feature set. The dictionary method uses stopwords. The experiments are conducted on three different character windows that provide an equivalent representation of short, medium and long document size. Overall, the 3-gram Naïve Bayes classifier, the 3-gram SVM classifier and the dictionary method showed an average classification accuracy of 98.37%, 99.53%, and 90.53% respectively. When trained with homogeneously distributed training data per language, the 3-gram Naïve Bayes and SVM classifiers showed an average classification accuracy of 95.16% and 96.2% respectively. To evaluate multilingual identification, an artificial corpus that contains 1050 documents is constructed. 45 out of 1050 documents are wrongly classified which corresponds to 95.71% accuracy. The challenging tasks in the study are: identification of closely related languages that share similar character sequences, identifying the language of short excerpts from texts, and the unavailability of standard corpus. The use of classification approach, combined with linguistically motivated features such as POS tags and morphological information is recommended as a way forward for providing empirical evidence on the convergences and divergences of language varieties in terms of lexicon, orthography, morphology and syntax.

Keywords— Language Identification, Multilingual Identification, N-gram, Feature Set, Naïve Bayes, Support Vector Machine, Dictionary Method, Character Window

I. INTRODUCTION

Ethiopia is a diverse country with various cultural and traditional differences. According to Ethnologue, (2017), there are ninety individual languages spoken in Ethiopia. Most people in the country speak Afro-asiatic languages of the Cushitic or Semitic branches. The Cushitic Languages are mostly spoken in central, southern and eastern Ethiopia (mainly in Afar, Oromia and Somali

[1]Assistant Professor, Department of Computer Science Joseph Arts and Science College Thirunavalur

regions). The Semitic Languages are spoken in northern, central and eastern Ethiopia (mainly in Tigray, Amhara, Harar and northern part of the Southern Peoples' State regions). They use the Ge'ez script that is unique to the country. The Omotic Languages are predominantly spoken between the Lakes of southern Rift Valley and the Omo River. The Nilo-Saharan Languages are largely spoken in the western part of the country along the border with Sudan (mainly in Gambella and Benshangul regions) (Indian Languages, 2008).

These languages use either Geez/Amharic or Latin transcriptions to represent the languages in textual forms. Textual data in some of these languages are getting more and more available on the global network. Among these are Amharic and Tigrigna use Geez/Amharic transcription. Somali, Afaan Oromo, Sidama, Afar and Nuer use Latin transcription (Desta, 2014).

Language identification is the task of automatically detecting the language(s) present in a document based on the content of the document (Lui & Baldwin, 2011). Automatic language identification is an integral part in many monolingual and multilingual language processing systems. Language identification can be divided into two classes: spoken and written language identification. Spoken language identification methods make use of signal processing techniques where language identification from text is a symbolic processing task.

The problem has been researched long both in the text domain and in the speech domain (House & Neuburg, 1977). Computational methods can be applied to determine a document's language before undertaking further processing. State-of-the-art methods of language identification for most European languages present satisfactory results above 95% accuracy (Martins & Silva, 2005).

Research on language identification has seen a variety of approaches. The major approaches include: detection based on stop words usage, detection based on character n-grams frequency, detection based on machine learning (ML) and hybrid methods. Many standard machine learning techniques has been applied to automated text categorization problems, such as Naïve Bayes classifiers, support vector machines, n-gram frequency rank order, and neural networks classifiers (Peng, Schuurmans, & Wang, 2003).

The main challenges in language identification include: Improving the coverage of language identification systems by increasing the number of languages that systems are able to recognize, Improving the robustness of language identification systems by training systems on multiple domains and various text types, Handling non-standard texts (e.g. multilingual texts, computer-mediated communication content, code-switching), and Discriminating between very similar languages, varieties and dialects.

II. RELATED WORK

Kruengkrai, Srichaivattana, Sorlertlamvanich, & Isahara, (2005) compared the performance of two kernel classifiers: the centroid method and an SVM. A string kernel was implemented which computes the inner product in the feature space generated by all subsequences of length k. A subsequence can be of any combination of k characters. The difference between n-grams and these combinations is that the subsequences do not need to follow contiguously. The subsequences are weighed by an exponential decaying factor, thus the subsequences that follow contiguously weighs more. A subsequence of k=5 and a decay factor of 1 gave best results. As baseline in comparisons, an n-gram rank ordering classification system was used and 5-fold cross validation on 17 languages and an average of 578 sentences per language was used to test the performance.

On a test sample (with an average length of 50 character per item) n-gram rank ordering performed with 90.2% accuracy, the centroid method with 95.9% accuracy and the SVM gave an overall best performance accuracy of 99.7%.

L-F.Zhai, (2006) found that a reduction in the feature space of the SVM results in a significant decrease in performance accuracy. They also concluded that the SVM is highly sensitive to prior distributions. This is due to the nature of the SVM classifier that does not compensate for different sizes of training data. Therefore, classification is biased towards the class with the larger training set.

A frequency based n-gram difference based classifier and a support vector machine (SVM) that uses the n-gram frequencies as features are discussed in Botha, Zimu, & Barnard, (2006). Error rates of approximately 0.3% are achieved over large text window sizes. It is also found that the SVM's performance is better than the n-gram based estimator's, but at a much greater computational cost.

H. Lodhi, (2002) showed that an SVM trained with n-gram statistics outperforms an SVM using kernel strings as features. Though it would be assumed that a kernel string would capture more information of a language, it probably introduces complexity into the SVM, which has a negative impact on decision-making.

Padro, (2004) compared the Naïve Bayes method, the dot-product classification and the n-gram rank ordering method to each other. Identification was tested on 6 languages (English, Catalan, Spanish, Italian, German and Dutch). Overall, the Naïve Bayes classifier proved best (significantly so for small test samples), followed by the dot-product classifier and then n-gram rank ordering method.

Truica, Velcin, & Boicea, (2015) presented a statistical method for automatic language identification of written text using dictionaries containing stopwords and diacritics. They proposed different approaches that combine the two dictionaries to accurately determine the language of textual corpora. They tested their method using a Twitter corpus containing 500,000 tweets with 100,000 tweets for each studied language and a news article corpus that contains 250,000 entries with 50,000 articles for each studied language. Their results show that their proposed method has an accuracy of over 90% for small texts and over 99.8% for large texts.

Desta, (2014) compared language identification accuracy by using character n-gram and character n-gram location as a feature set for Naïve Bayes and Frequency rank order models. The results showed that Naïve Bayes classifier achieved highest accuracy for short, medium, and long string test documents. The identification accuracy of Frequency Rank Order is low but showed an improvement for long text test documents. When using character n-gram and its location frequency as feature set, the accuracy of the both models showed an improvement. Although this is an encouraging result, it's difficult to jump to conclusions without considering factors that can determine identification accuracy such as the amount and variety of training data.

From the literature, it is evident that different approaches to solving the language identification problem have been investigated. Naive Bayes classifiers prove to be quite popular and successful with high orders of n for the n-grams used. The use of SVMs also appear to be a popular choice achieving good performance. However, not many studies have been conducted regarding dictionaries containing stopwords and also focusing on major Ethiopian languages. Identification based on SVM, Naïve Bayes and dictionary method were used in this research.

III. METHODOLOGY

Different methods have been used to compute classification accuracies in different studies regarding text-based language identification. Different amounts of textual training data are used to train classifiers where documents spanned different domains. The number of characters and words; the size in bytes, lines and sentences influence the metric for the size of the test string (Botha, Zimu, & Barnard, 2006).

Languages without any family relationships and other within language families were used to perform some of the tests. Some evaluated performance accuracy using only a validation set where others performed a thorough cross-validation test. Classifiers were evaluated under the same conditions in different studies that compared different classifiers against each other so that more reliable conclusions can be made.

This study uses some previous ideas for evaluation process and it is designed to make comparisons more reliable. To assure reliable results, the classifiers were evaluated under the same circumstances.

In this study languages used for the experiment are from different linguistic classification namely, Cushitic and Semitic of the Afro-Asiatic linguistic classification and Nilotic of the Nilo-Saharan linguistic classification. Seven local languages are used in this study namely, Afar, Amharic, Nuer, Oromo, Sidamo, Somali and Tigrigna.

A. Corpora

Seven corpora were collected to perform the experiments described here. All corpora consist of different texts compiled from different sources. The data included text from various sources (such as newspapers, periodicals, books, the Bible and government documents) and therefore, the corpus spans several domains. Documents written in 7 languages and language varieties were compiled and processed resulting in slightly over 11.7 million tokens.

To evaluate multilingual identification, an artificial corpus that contains 1050 documents was constructed. Mixed input texts (3 languages in the same document) are considered in this study and each one of the documents is a concatenation of 3 sections in different languages which are constructed randomly from a collection of medium length texts per language (6 to 50 words).

B. Evaluation and Computational Techniques

We can see that many approaches can be followed in a text-based language identification from the literature study. Using a pure linguistic approach is undeniably the better choice to achieve a high classification accuracy but it requires a large amount of linguistic expertise. Therefore, using statistical approaches is a feasible alternative. Statistics of words, letters or n-grams can be used to build statistical language models. The n-gram based models outperform a word-based model for small text fragments and do equally well for larger fragments (Grefenstette, 1995). In another study n-grams achieved better results than string kernels (Lodhi, Shawe-Taylor, Cristianini, & Watkins, 2002). We can see from the literature that this is by far the most popular choice. That is why the feature sets are restricted to n-gram based features. Naïve Bayes and SVM are trained by using character n-gram as feature set.

To evaluate the performance of the models, windows containing a specified number of words which are known as character windows are used. Character windows were used in the experiments here since they are the most reliable measure of window size. The experiments are conducted on three different character windows. 15-character window which represents a short document size or

2-3 words, 100- character window which represents a medium document size or a long sentence and 300-character window which represents a long document size or a paragraph. The difficulty of correct classification is influenced by the choice of these sizes. Average number of characters per word was calculated for each language. This is helpful to know the number of words required for the character windows. It is possible to generate test documents based on the analysis made on the character windows.

C. Evaluation Metrics

To evaluate the extent to which the methods used in this thesis are suitable for language identification, standard metrics used in Natural Language Processing and Text Classification to report results in terms of accuracy and error rate are used.

D. Smoothing Techniques

In this study, Laplace smoothing is used to avoid zero multiplication for n-gram that was not seen during model training. This is done by adding one to each n-gram frequency. If the n-gram does not exist in training phase it was discarded from the calculation since the size of training corpus is too large. Hence the effect of n-gram with zero frequency is minimal on the classification accuracy. This process continues until probability of language given the character n-gram is calculated for each of the languages. The language with maximum such probability is considered as the language of the unknown text document.

E. Classification Methods

In this study, implementations of Naive Bayes and Support Vector Machines which are popular machine learning classifiers are used. Another technique used in this is dictionaries containing stop words.

Experimenting with all possible combinations was not feasible although a large number of classifiers have been applied to text-based language identification. The classification algorithms selected, were similarly chosen for their proven performance in published studies, as well as their ability to clarify theoretical issues.

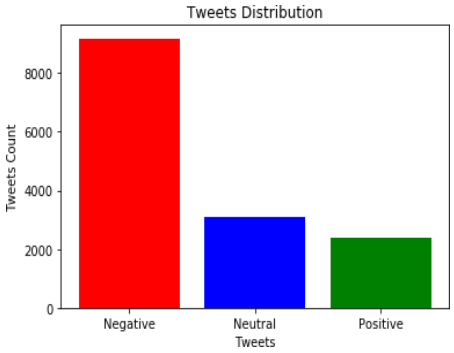

Figure 1. Naïve Bayes and SVM models applied to text languages in this study

IV. EXPERIMENT AND RESULT

The experimentation conducted in this study is separated into two experiments, with each experiment differing with regards to the amount of data used for training. The results of the

experiments are analyzed individually at the end of each experiment and an overall analysis is done to summarize the observations.

The first experiment utilizes the entire data set to train and test the classifiers with seven language classes. This experiment was conducted by randomly taking 90% of the corpus of the seven target languages for training the models and the remaining 10% for testing the models.

Accuracy and error rates were calculated for comparing the classifiers. 500,216 test documents of differing size were used for the seven languages. These were divided into short document size or 2-3 words, medium document size or 6-50 words and long document size or more than 50-word long sentence or a paragraph. The test documents that are divided are equivalent to the three-character windows.

Number of Records	Time taken to execute (In millisecond) K-Mean Algorithms	Time taken to execute (In millisecond) Modified K-Mean Algorithm
300	95240	61613
400	116243	73322
500	135624	103232
600	158333	122429

Table 1. Comparison (Accuracy) on test data for Naïve Bayes Classifier (n=3)

	SVM	**Naïve Bayes**
Accuracy	82.48	76.56
Precision	90.33	89.00
Recall	81.79	83.75
F1 Measure	85.85	86.37

Table 2. Evaluation (Accuracy) on test data for SVM Classifier (n=3)

The second experiment consists of varying the size of the training data used to train the classifiers to simulate scarce data availability. Furthermore, this experiment investigates the effect that homogeneously distributed training data (equal amounts of training data for each language class) has on a classifier when compared to a training set with heterogeneously distributed training data. The models are re-trained and tested with the same data size for the seven languages in this study. The smallest training data size is the data size taken for Nuer language (4,889 training characters). Therefore, similar data sizes of 4,889 training characters were used per language.

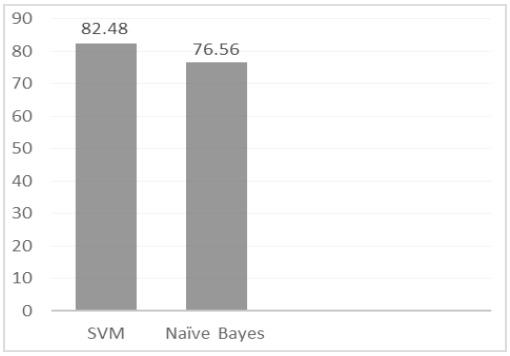

Figure 2 Evaluation (Accuracy) on test data for SVM Classifier (n=3) trained with homogeneously distributed training data per language

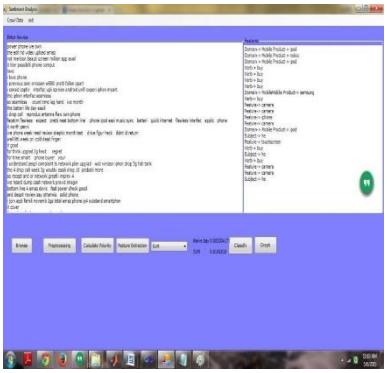

Figure 3. Evaluation (Accuracy) on test data for SVM Classifier (n=3) trained with homogeneously distributed training data per language

Naïve Bayes classifier is used for multilingual language identification. For mixed-language input, the top three languages found is returned. To evaluate multilingual identification, an artificial corpus that contains 1050 documents was constructed. Each one of the documents is a concatenation of 3 sections in different languages which are constructed randomly from a collection of medium length texts per language (6 to 50 words).

V. CONCLUSION AND RECOMMENDATION

A. Conclusion

In this research, an adequate mechanism for efficient language identification of major Ethiopian languages was presented. The languages used for the experiment are from different linguistic classification namely Cushitic and Semitic of the Afro-Asiatic linguistic classification and Nilotic of the Nilo-Saharan linguistic classification. Seven local languages are used in this study namely Afar, Amharic, Nuer, Oromo, Sidamo, Somali and Tigrigna. These languages were chosen because they are spoken by more than 79.3% of the total population of Ethiopia.

Factors affecting accuracy such as the size and variety of training data, the size of the string to be identified, and the type of classifier employed were investigated. Three approaches were considered in this study. The Naïve Bayes, SVM and dictionary method. N-gram statistics were used as features for classification for Naïve Bayes and SVM.

The Naïve Bayes and SVM classifiers were trained by using character n-gram of size 3 as a feature set. The dictionary method uses stopwords. Three different sizes of character windows were used to perform the tests. These sizes were chosen to provide a range of challenges for classification (the larger the test string, the easier the classification task becomes). Text from various domains was collected and was preprocessed before building the models. Naïve Bayes and SVM classifiers were trained with heterogeneously distributed training data at the first experiment and then with homogeneously distributed training data (equal amounts of training data for each language class) at the second experiment. All classifiers were tested under the same conditions.

In the first experiment, which utilizes the entire data set to train and test the classifiers with seven language classes, the 3-gram Naïve Bayes and SVM classifiers showed an average classification accuracy of 98.37% and 99.53% respectively. In the second experiment, which focused on varying the size of the training data used to train the classifiers to simulate scarce data availability and which investigated the effect that homogeneously distributed training data has on a classifier, the 3-gram Naïve Bayes and SVM classifiers showed an average classification accuracy of 95.16% and 96.2% respectively. The dictionary method showed an average classification accuracy of 90.53%.

For multilingual language identification, Naïve Bayes classifier is used. The top three languages found is returned for mixed language input therefore a concatenation of 3 sections in different languages which are constructed randomly from a collection of medium length texts per language (6 to 50 words) is used for evaluating multilingual identification. This led to 95.71% accuracy.

The main challenge in this study is identification of closely related languages that share similar character sequences and lexical units (e.g. Amharic and Tigrigna). Another challenge faced by the researcher is identifying the language of short excerpts from texts particularly those containing non-standard language and the unavailability of standard corpus.

B. Recommendation

The results of this research are beneficial and can be used in any of the areas of language identification applications. Language identification of text is important as large quantities of text are processed or filtered automatically for tasks such as spell checker, information retrieval and machine translation. The results of this research can therefore be used by anyone who is interested in these research areas.

Since Ethiopia is a multilingual country, other languages from different linguistic classifications can be added which improves the coverage of language identification systems by increasing the number of languages that systems are able to recognize.

For further studies, the researcher recommends the use of classification methods combined with linguistically motivated features such as POS tags and morphological information which can provide empirical evidence on the convergences and divergences of language varieties in terms of lexicon, orthography, morphology and syntax.

ACKNOWLEDGMENT

First and foremost, I would like to thank the almighty God who made all things possible. I like to express my sincere thanks to my Secretary, Principal and vice principal for his important comments and encouragement. Most importantly, I thank my parents and my family for giving me the possibility to become who I am today.

REFERENCES

1. Botha, G., Zimu, V., & Barnard, E. (2006). Text-based Language Identification for the South African Languages. *SAIEE Africa Research Journal* .
2. Cavnar, W., & Trenkle, J. (1994). N-gram-Based Text Categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*.
3. Central Intelligence Agency. (n.d.). *Ethiopia*. Retrieved April 13, 2019, from The World Factbook: https://www.cia.gov/library/publications/the-world-factbook/geos/et.html
4. Central Statistical Agency. (2010). *Population and Housing Census 2007 Report*. Retrieved April 13, 2019, from http://catalog.ihsn.org/index.php/catalog/3583/download/50086
5. Chang, C., & Lin, C. (2001). *LIBSVM: a library for support vector machines.* Retrieved August 2, 2019, from http://www.csie.ntu.edu.tw/cjlin/libsvm
6. Collobert, R., & Bengio, S. (2001). SVMTorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research, I*, 143-160.
7. COMESA, R. I. (n.d.). *Regional Somali Language Academy Launched in Djibouti*. Retrieved December 6, 2019, from http://www.comesaria.org/site/en/news_details.php?chaine=regional-somali-language-academy-launched-in-djibouti&id_news=17578&id_article=119
8. Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support Vector Machines: and other kernel-based learning methods.* New York: Cambridge University Press.
9. Desta, L. W. (2014). Modeling Text Language Identification for Ethiopian Cushitic Languages: Masters thesis. HiLCoE School of computer Science.
10. Dunning, T. (1994). Statistical identification of language. *Computing Research Lab, New Mexico State University*.
11. Emerson, G., Tan, L., Fertmann, S., Palmer, A., & Regneri, M. (2014). SeedLing: Building and using a seed corpus for the Human Language Project. 77–85.
12. *Ethiopian Languages*. (2008). Retrieved November 28, 2019, from Dinknesh Ethiopia Tour: http://www.dinkneshethiopiatour.com/index.htm
13. Ethnologue. (2017). *Languages of Ethiopia*. Retrieved April 13, 2019, from https://www.ethnologue.com/country/et/languages
14. Gebremichael, M. (2011). Federalism and conflict management in Ethiopia : case study of Benishangul-Gumuz Regional State. *United Kingdom: University of Bradford.*
15. Good, I. J. (1965). The estimation of probabilities. *An essay on modern Bayesian methods*.
16. Google, T. (2013). *Google Translate - now in 80 languages*. Retrieved December 6, 2019
17. Grefenstette, G. (1995). Comparing two Language Identification Schemes. *Third International Conference on Statistical Analysis of Textual Data*.
18. Gurtong, T. (n.d.). *Nuer (Naath)*. Retrieved December 6, 2019, from www.gurtong.net
19. Hakkinen, J., & Tian, J. (2001). n-gram and decision tree based language identification for written words. *Automatic Speech Recognition and Understanding*, 335–338.
20. Hassan, R. J. (1981). The Oromo Orthography of Shaykh Bakri Saḍalō. *Bulletin of the School of Oriental and African Studies*, 550-566.
21. Haykin, S. (1994). *Neural networks: a comprehensive foundation.* Prentice Hall PTR.

22. Hayward, & Hassan. (1981). The Oromo Orthography of Shaykh Bakri Saṗalō. *Bulletin of the School of Oriental and African Studies*, 551.
23. Heine, B. (1981). The Waata Dialect of Oromo: Grammatical Sketch and Vocabulary.
24. House, S. A., & Neuburg, P. E. (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustical Society of America*, 708–713.
25. Hudson, G. (2009). "Amharic". The World's Major Languages. 594–617.
26. Ingle, N. (1980). A Language Identification Table. Technical Translation International.
27. Ishwaran, H., & Rao, J. (2009). Decision tree: introduction. In *Encyclopedia of Medical Decision Making* (pp. 323–328).
28. J.Hakkinen, J. (2001). n-Gram and Decision Tree Based Language Identification For Written Words. *Workshop on Automatic Speech Recognition and Understanding, Trento*, 335-339.
29. Jimma_Times. (n.d.). *Online Afaan Oromoo–English Dictionary*. Retrieved December 6, 2019, from http://www.jimmatimes.com/article/ARTS/ARTS/Ethiopia_Online_Afaan_Oromoo_English_Dictionary/32153
30. Judith, R. (2013). Globally Speaking: Motives for Adopting English Vocabulary in Other Languages (Multilingual Matters). 165.
31. Kawachi, K. (2007). A grammar of Sidaama (Sidamo), a Cushitic language of Ethiopia: Doctoral dissertation. State University of New York.
32. Kerlinger, F. N. (1973). *Foundations of behavioral research.* New York.
33. Kikui, G.-I. (1994). Identifying the coding system and language of on-line documents on the internet. *Proceedings of the 16th International Conference on Computational Linguis- tics*, Denmark.
34. King, B., & Abney, S. (2013). Labeling the languages of words in mixed- language documents using weakly supervised methods. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
35. Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? Nature biotechnology.
36. Kizitus Mpoche, T. M. (2006). Language, literature, and identity. 163–164.
37. Kralisch, Anett, & Mandl, T. (2006). Barriers to information access across languages on the internet: Network and language effects. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, volume 3*.
38. Kruengkrai, Srichaivattana, Sorlertlamvanich, & Isahara. (2005). Language Identification Based on String Kernels. *IEEE International Symposium on Communications and Information Technology*.
39. Kullback, S., & Leibler, R. (1951). On information and sufficiency. In *Annals of Mathematical Statistics* (pp. 79–86).
40. *Language Access Act Fact Sheet*. (2011). Retrieved April 13, 2019, from http://ohr.dc.gov/sites/default/files/dc/sites/ohr/publication/attachments/LAAFactSheet-English.pdf
41. Lewis, I. (1998). Peoples of the Horn of Africa: Somali, Afar and Saho. *Red Sea Press*, 11.
42. L-F.Zhai, M. X. (2006). Discriminativelytrainedlanguagemodelsusing support vector machines for language identification. *The Speaker and Language Recognition Workshop*, 1-6.
43. Lipschutz, S., & Lipson, M. (2009). Linear Algebra (Schaum's Outlines). McGraw Hill.

44. Ljubesic, Fiser, & Erjavec. (2014). Tweet-CaT: a Tool for Building Twitter Corpora of Smaller Languages. . *Proceedings of the Internacional Conference on Language Resources and Evaluation (LREC)*, 2279–2283.
45. Lodhi, Shawe-Taylor, Cristianini, & Watkins. (2002). Text classification using string kernels. *Journal of Machine Learning Research*.
46. Lui, M. (2014). Generalized Language Identification. *PhD thesis* , 2014.
47. Lui, M., & Baldwin, T. (2011). Cross-domain feature selection for language identification. *Proceedings of the 5th International Joint Conference on Natural Language Processing*.
48. Manning, C., & Schu ̈tze, H. (1999). Foundations of Statistical Natural Language Processing. *MIT Press*.
49. Martins, B., & Silva, M. (2005). Language Identification in Web Pages. *Proceedings of the 20th ACM Symposium on Applied Computing (SAC)*, 763-768.
50. Meyer, R. (2006). Amharic as lingua franca in Ethiopia. *Journal of African Languages and Linguistics*, 117–131.
51. Nguyen, D., & Dogruoz, A. S. (2013). Word Level Language Identification in Online Multilingual Communication. *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 857–862.
52. Omniglot. (2017). *Afar alphabets, pronunciation and language*. Retrieved December 6, 2019, from http://www.omniglot.com/writing/afar.htm
53. Omniglot. (2017). *Nuer alphabets, pronunciation and language*. Retrieved December 6, 2019, from http://www.omniglot.com/writing/nuer.htm
54. Omniglot. (2017). *Somali alphabets, pronunciation and language*. Retrieved December 6, 2019, from http://www.omniglot.com/writing/somali.htm
55. Padro, M. P. (2004). Comparing methods for language identication. *Proceedings of the XX Congreso de la Sociedad Espanola para el Procesamiento del Lenguage Natural*.
56. Palmer, D. (2010). Text Processing. (N. Indurkhya, & F. Damerau, Eds.) *Handbook of Natural Language Processing*, 9–30.
57. Pankurst, A. (1991). Indigenising Islam in Wällo: ajäm, Amharic verse written in Arabic script. *Proceedings of the Xlth International Conference of Ethiopian Studies, Addis Ababa*.
58. Peng, F., & Schuurmans, D. (2003). Combining naive Bayes and n-gram language models for text classification. *Springer*.
59. Peng, F., Schuurmans, D., & Wang, S. (2003). Augmenting Naive Bayes Classifiers with Statistical Language. *University of Waterloo Canada*.
60. Powers, & David, M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 37–63.
61. Powers, & David, M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies, II*, 37–63.
62. Prager, J. (1999). Linguini: language identification for multilingual documents. *Journal of Management Information Systems*, 71-101.
63. Python. (2018). *Python*. Retrieved February 2, 2019, from Python: https://www.python.org/
64. Raymond, G. J. (2005). Ethnologue: Languages of the World. *Dallas: Summer Institute of Linguistics*.

65. Rehurek, Radim, & Kolkus, M. (2009). Language identification on the web: Extending the dictionary method. *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009)*, 357-368.
66. Reynar, P. S. (1996). Language identification: Examining the issues. *Proceedings of the 5th Symposium on Document Analysis and Information Retrieval*, 125-135.
67. Rodolfo, F. (2003). Akkälä Guzay. *Encyclopaedia Aethiopica: A-C. Wiesbaden: Otto Harrassowitz KG*, 169.
68. S. MacNamara, P. C. (1998). Neural networks for language identification: A comparative study. *Information Processing and Management*, 395-403.
69. Shannon, C. (1951). Prediction and Entropy of Printed English. *Bell system technical journal*.
70. Simões, A., Almeida, J. J., & Byers, S. D. (1998). Language Identification: a Neural Network Approach. *ACM*.
71. Stephanie. (2018). *Experimental Design*. Retrieved February 4, 2019, from Statistics How To: http://www.statisticshowto.com/
72. Teahan, W. J. (2000). Text classification and segmentation using minimum cross- entropy. *Proceedings of the 6th International Conference Recherche d'Information Assistee par Ordinateur (RIAO'00)*, 943–961.
73. Truica, C.-O., Velcin, J., & Boicea, A. (2015). Automatic Language Identification for Romance Languages using Stopwords and Diacritics. *University Politehnica of Bucharest*.
74. University of Pennsylvania, S. o. (1977). Afaan Oromo.
75. WebCorp. (2018). *WebCorp*. (B. C. University, Producer, & Birmingham City University) Retrieved February 2, 2019, from WebCorp: http://www.webcorp.org.uk/live/
76. Yamaguchi, Hiroshi, & Tanaka-Ishii, K. (2012). Text segmentation by language using minimum description length. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, 969-978.
77. Zampieri, M., Gebre, B. G., & Diwersy, S. (2012). Classifying pluricentric languages: Extending the monolingual model. *Proceedings of the Fourth Swedish Language Technlogy Conference (SLTC)*.