ML1819 Research Assignment 2

(Language Detection of Web Pages Using N-Gram Approach for European Languages)

Team ID: 39
Task ID: 202

Task Title: Language Detection

Member 1: Mohd Tousif (18303317)

Member 2: Gauransh Bhutani (18303042)

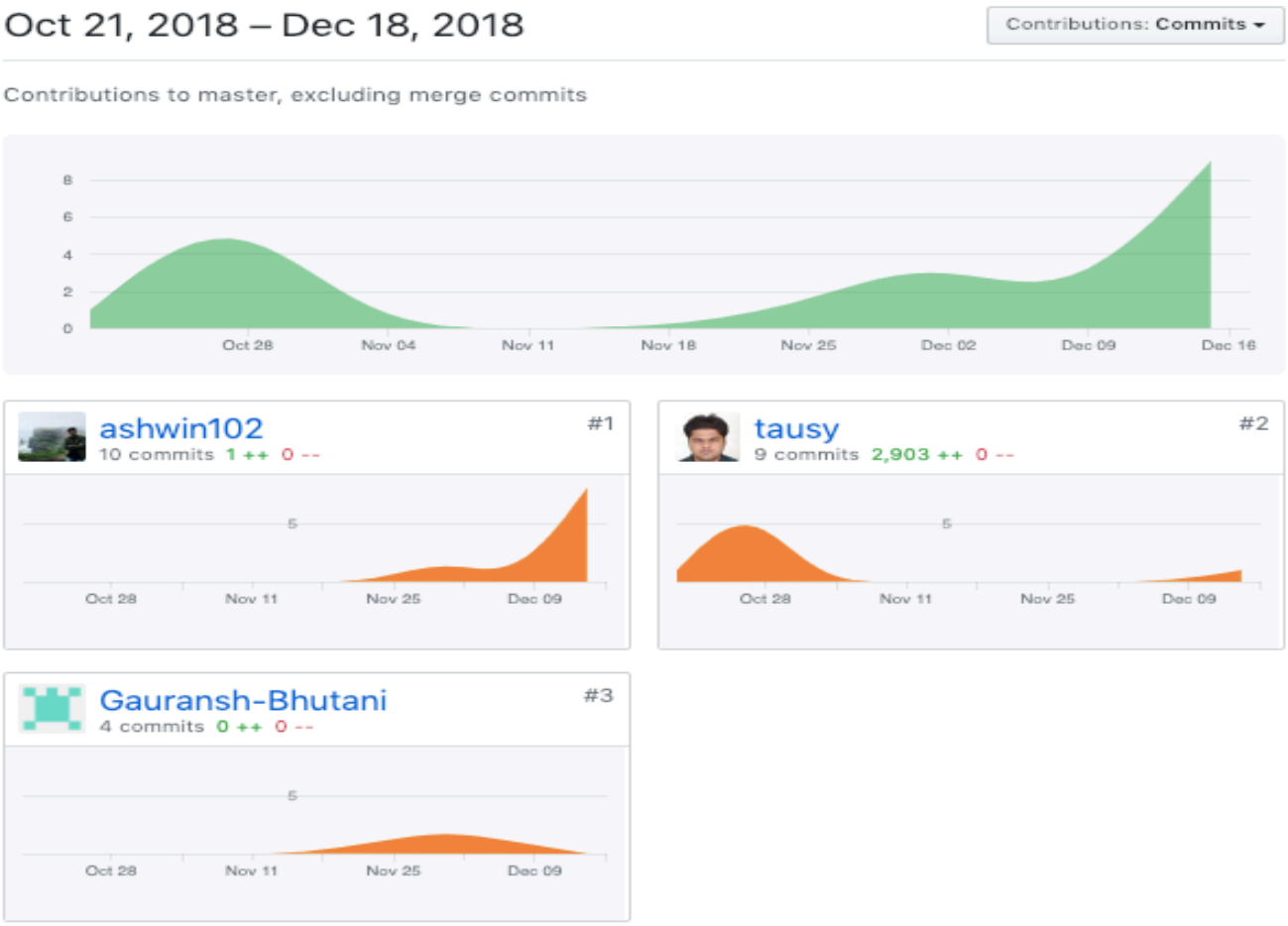
Member 3: Ashwin Kumar (18302700)

Word Count: 1576 (Excluding References, Diagram Captions)

URL (Source Code Repository): https://github.com/tausy/ML1819--task-202--Team-39

URL (Source Code Repository Activity): https://github.com/tausy/ML1819--task-202--Team-39/graphs/contributors?from=2018-10- 21&to=2018-12-18&type=c

Screenshot (Commit Activity):



Language Detection of Web Pages Using N-Gram Approach for European Languages

Mohd Tousif
MSc Data Science
Trinity College Dublin
mohdtout@tcd.ie

Ashwin Kumar
MSc Data Science
Trinity College Dublin
kumara1@tcd.ie

Gauransh Bhutani
MSc Data Science
Trinity College Dublin
bhutanig@tcd.ie

ABSTRACT

Language identification is predominantly used in numerous NLP systems as a seeding step to determine the language in which a web page is written. A huge amount of work has already been done in this area. This paper uses N-Gram approach to detect the language of a web page and uses WiLi-2018 (Wikipedia Language Identification dataset) [4] to detect the content of the text of European languages, in which they are inscribed.

Keywords

Language Detection, Language Identification, Natural Language Processing, N-Grams.

1. INTRODUCTION

The basis of this paper comes under the category of identification of European languages using WiLI 2018 dataset. Language detection can be achieved using Computational or Non-Computational technique. Statistical approaches can be divided into training and testing steps. In train step dataset is taken and corpus is created using feature extraction while in Classification step similarity between the training and the testing language profile is estimated and the most alike language is returned as the language of the given text. Non-Computational techniques require extensive knowledge about the particular language in advance, such as the character combinations, most frequent words used, etc. This paper uses n-gram approach, a statistical technique to detect the language of a given web page.

N-gram is a probabilistic language model which is used to detect the item in a sequence in the form of (n-1) Markov model. N-grams [2] typically refer to the sequence of words. A unigram is one word, a bi-gram is a progression of two words, a tri-gram is a sequence of three words. [3] One important point to consider is that the items within n-gram may not necessarily have any kind of relationship between them apart from the certainty that the word appears next to one another.

2. RELATED WORK

2.1 Cavnar and Trenkle Approach [2]

Cavnar and Trenkle (1994) used N-Gram based distance measure to accurately categorize the small documents and news articles. Distance measure was calculated between the language profile of training set data and the target document. This approach was primarily dedicated only towards the languages directly representable in ASCII.

2.2 Suzuki Algorithm [3]

Suzuki (2002) taken into account two predetermined values UB and LB to check the response of identification on the target text. Matching rate, a list of shift codons was calculated using Tri-Grams and compared to shift codons of the training set and the language for which the matching rate of training text is higher than the UB was assumed to be the target texts' language. This method was able to identify English, German, Portuguese and Romanian languages with high accuracy.

2.3 Language Identification of Web Pages [5]

In Bruno Martins (2005), automatic language identification of a given web page was achieved using the N-gram based algorithm and complemented it with a more efficient similarity measure and heuristics to handle the web pages in a better way. In this approach 23 different language models were constructed and tested upon 12 different languages, like, Danish, Dutch, English, French, German achieving an accuracy of 91.25%.

2.4 Language Identification of Web Pages Based on Improved N-gram Algorithm [6]

Choong et al. (2011) proposed a two-step process. Tri-Grams were used in combination with two heuristics namely byte-sequence HTML parsing and HTML character translation of web pages to obtain an accuracy of 94.04% for non-Latin (Asian and African) languages.

2.5 Graph-Based N-Gram Approach [8]

In Erik Tromp (2011), graph-based N-Gram approach was proposed for short and ill-written texts. A graph model was developed to predict the language of a text referencing to some previously unseen text. This approach was not applicable to multilingual documents.

2.6 The WiLI Benchmark Dataset for Written Language Identification [4]

In Martin Thoma(2018), open source dataset of short text extracts for 235 languages has been explained. It is a well-balanced classification dataset (with test–train split) that can be effectively used for language identification techniques.

3. METHODOLOGY

The approach presented by Choong et al [6] for N-gram generation only considered Tri-Grams for the detection of languages. Moreover, it typically focuses on Non-latin(Asian and African) languages. In our approach, we have considered Mono, Mono/Bi and Mono/Bi/Tri -Grams and tried to extend the N-gram approach to identify the European languages or Latin languages precisely.

$$R_i = \sum_{i=1}^{n} \frac{m_i}{n}$$

Where,

$$m_i = \begin{cases} 0 & if \ \ t_i \ \ did\ not\ match\ with\ T_j \\ 1 & if\ t_i\ matched\ with\ T_j \end{cases}$$

Equation 1: Comparison of Matching Rates

Ri is the rate at which the ith distinct N-Gram in the target profile (ti) matches with the jth distinct N-Gram in the training profile (Tj).

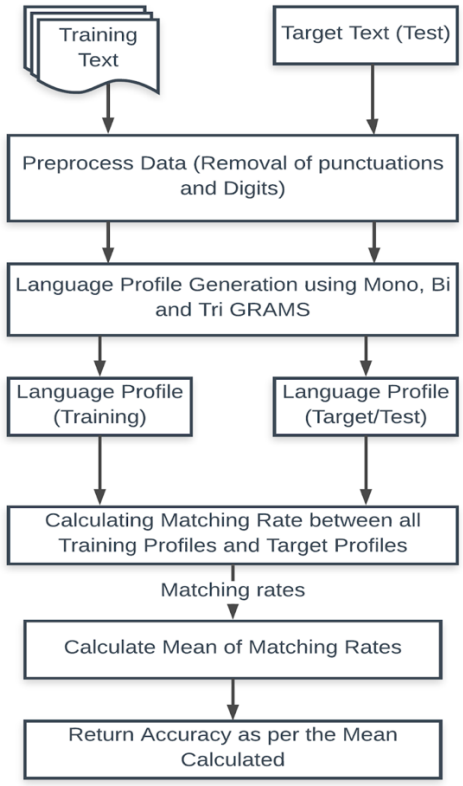


Figure 1. Flow of a Language Identification System [6]

A typical language identification system calculates language profiles from training texts of various language schemes using N-Grams(Figure-1). In a similar manner, the language profile for the target text is generated. The system then evaluates the matching rates of N-Grams between training and target profile. The matching rate is calculated as the average of match factors(mi) whenever an N-gram in target profile matched with a N-gram in the training corpus. Then the language with the highest match rate is returned as the language of a target web page given that the match rate is not lesser than the lower bound(LB). We have taken the value of LB as 0.5.

This paper uses the updated version of the famous WiLI-2018 dataset [4]. WiLI-2018 benchmark dataset is created from Wikipedia web pages written in multiple languages and publicly available. The dataset consists of a total of 235 languages and more than 0.1M web pages of different languages. We have targeted 34 European Languages amounting to 18000 web pages from this dataset.

4. RESULTS AND DISCUSSION

In this paper, the N-gram based language identification algorithm was reported and calculation was taken into account for 34 European languages. We have taken Mono, Mono and Bi, and Mono, Bi and Tri-grams into consideration for the language identification model. The model proposed provides an enhancement to the Choong et al [6] algorithm which was implemented for the Asian and African languages by using the tri-grams.

4.1 Using Mono-grams

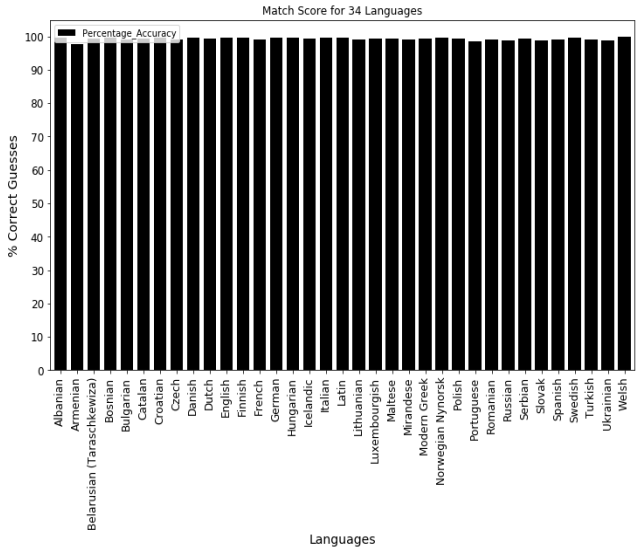


Figure 2. Results for Language Identification Algorithm considering Mono-grams

The above graph (Figure 2) represents the percentage accuracy of correct guesses in accordance with 34 European languages when only Mono-grams are considered. The model achieved an accuracy between 97.5% to 99.9% with the maximum accuracy for Welsh (99.85%) and minimum for Armenian (97.74).

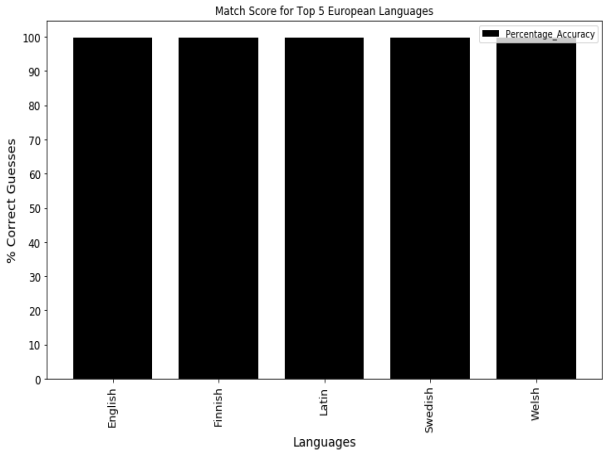


Figure 3. Percentage Results for Top 5 Languages considering Mono-grams

Results show the percentage accuracy of the top 5 languages with Welsh (99.85%) topping the charts followed by English (99.72%), Swedish (99.7%), Latin (99.67%) and Finnish (99.66%) obtaining the fifth spot (Figure 3).

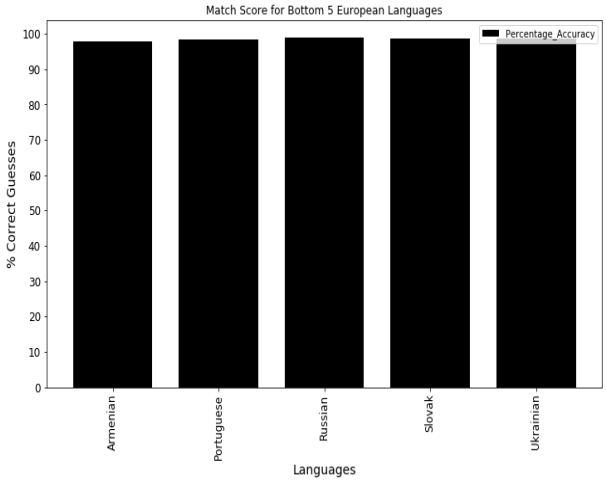


Figure 4. Percentage Results for Bottom 5 Languages considering Mono-grams

Results presented in the graph (Figure 4) show the accuracy of bottom 5 languages when Mono-gram is consiederd. Armenian (97.74%), Portugese (98.41%), Slovak (98.66%), Ukrainian (98.78%) and Russian (98.87%) are the bottom 5 languages identified by the system.

4.2 Using Mono and Bi-grams

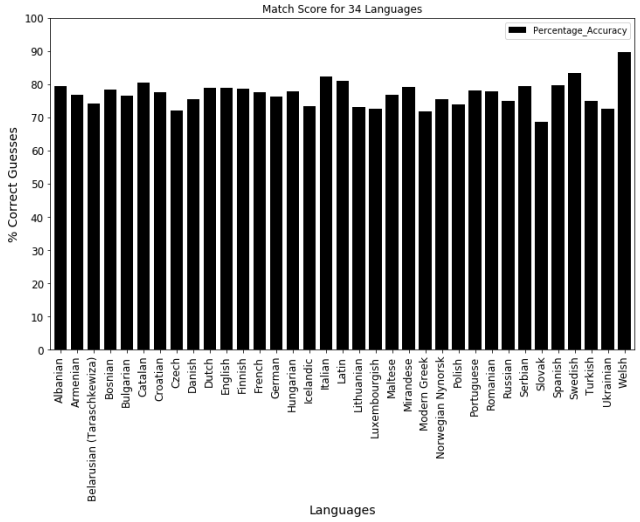


Figure 5. Results for Language Identification Algorithm considering Mono and Bi-grams

The above graph (Figure 5) shows the percentage accuracy of correct guesses in accordance with 34 European languages when Mono as well as Bi-grams are considered. As per the obtained results it can be seen that the achieved accuracy is lower as compared to considering only Mono-grams for the model. The model achieved an accuracy between 68% to 90% with the maximum accuracy for Welsh (89.8%) and minimum for Armenian (68.6%).

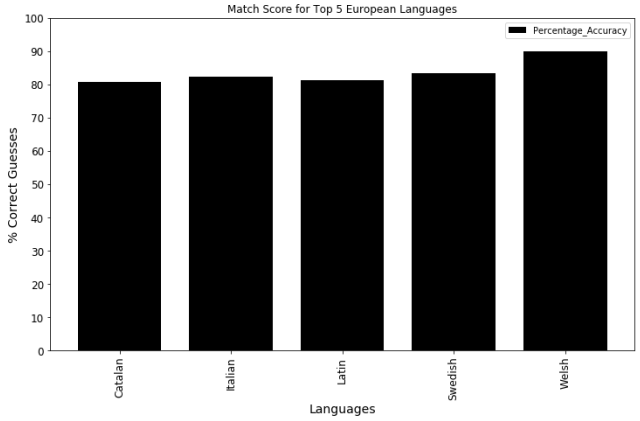


Figure 6. Percentage Results for Top 5 Languages considering Mono and Bi-grams

The above graph shows the percentage accuracy of the top 5 languages with Welsh (89.8%) retaining the top spot followed by Swedish (83.3%), Italian (82.3%), Latin (81.1%) and Catalan (80.6%) obtaining the fifth spot (Figure 6).

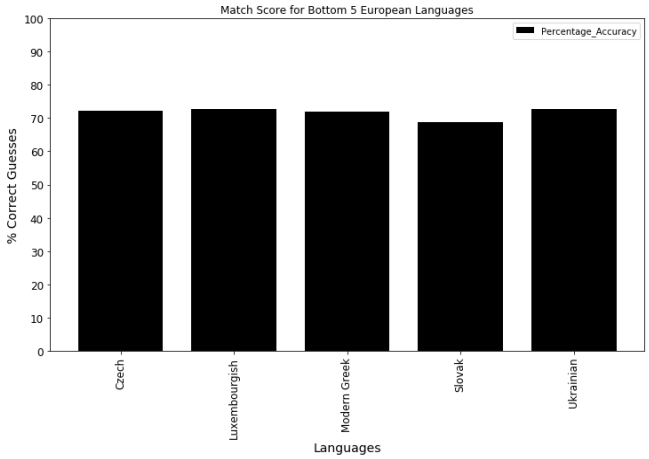


Figure 7. Percentage Results for Bottom 5 Languages considering Mono and Bi-grams

Results presented in the graph (Figure 7) show the accuracy of bottom 5 languages when Mono and Bi-grams are taken into account. Slovak (68.6%), Modern Greek (71.8%), Czech (72.1%), Ukrainian (72.63%) and Luxembourgish (72.68%) are the bottom 5 languages identified by the model

4.3 Using Mono, Bi and Tri-grams

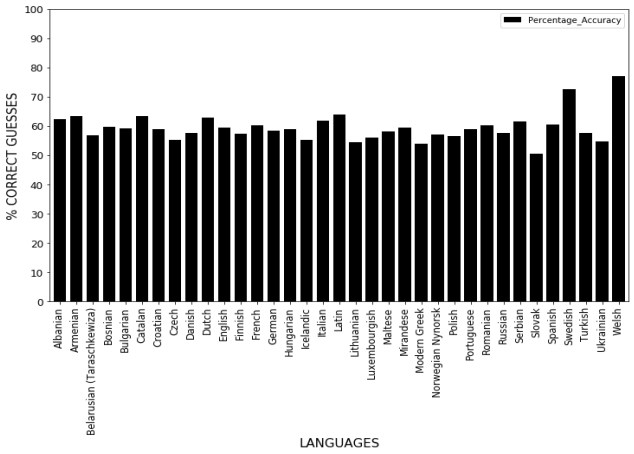


Figure 8. Results for Language Identification Algorithm considering Mono, Bi and Tri-grams

The above graph (Figure 8) presents the percentage accuracy of the correct guesses in accordance with Mono, Bi and Tri-grams. The achieved accuracy for different languages is between 50% to 77% with the maximum accuracy for Welsh (77%) and minimum for Slovak (50%).

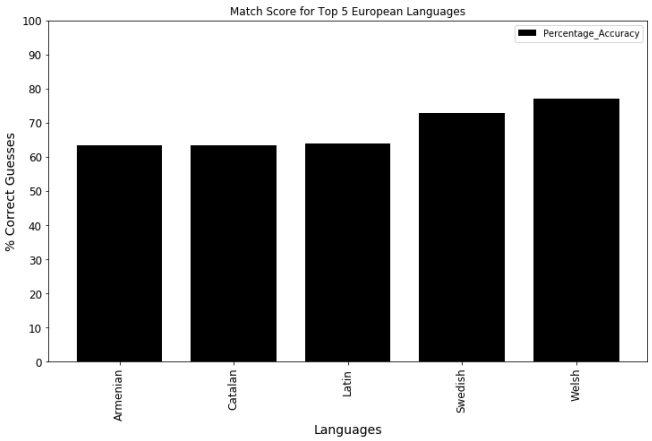


Figure 9. Percentage Results for Top 5 Languages considering Mono, Bi and Tri-grams

Results show the percentage accuracy of the top 5 languages with Welsh (77.05%) topping the charts as compared to other languages, followed by Swedish (72.74%), Latin (64.008%), Catalan (63.42%) and Dutch (62.811%), obtaining the fifth spot (Figure 9).

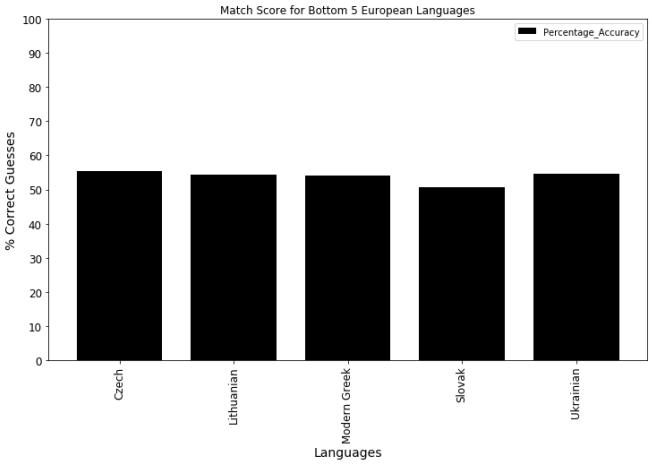


Figure 10. Percentage Results for Bottom 5 languages considering Mono, Bi and Tri-grams

Results presented in the graph (Figure 10) show the languages with relatively low accuracy. Slovak (50.55%), Modern Greek (53.95%), Lithuanian (54.42%) and Ukrainian (54.70%) are identified with a low accuracy by the system. Accuracy of the system can be enhanced further by taking heuristics into consideration which is a potential future work.

4.4 Comparison of different N-Grams Techniques

The table (Figure 11) presented below shows comparison of percentage accuracy for different N-gram approaches taken into account in this work.

LANGUAGES	LABEL	ACCURACY (%)		
		MONO	MONO_BI	MONO_BI_TRI
Welsh	cym	98.8507	89.8226	77.052816
Swedish	swe	99.7081	83.3129	72.74770576
Latin	lat	99.6781	82.3898	64.00870579
Armenian	hye	97.7405	76.9154	63.49032383
Catalan	cat	99.2318	80.6481	63.42693645
Dutch	nld	99.3524	79.018	62.81178246
Albanian	sqi	99.6134	79.3653	62.508136
Italian	ita	99.6512	82.3898	61.86833761
Serbian	srp	99.3713	79.6019	61.66896
Spanish	spa	99.9164	79.647	60.6317533
French	fra	99.0281	77.7318	60.403813
Romanian	ron	99.171	78.0089	60.32541487
Bosnian	bos	99.5668	78.3596	59.79788191
Mirandese	mwl	99..0649	79.1298	59.60185152
English	eng	99.7282	78.8535	59.512061
Bulgarian	bul	98.9811	76.4877	59.21277341
Portuguese	por	98.4196	78.2984	59.0268777
Hungarian	hun	99.629	77.8308	58.93988437
Croatian	hrv	99.501	77.6179	58.87997511
German	deu	99.4897	76.4432	58.58312387
Maltese	mlt	99.3536	76.7345	58.3004196
Danish	dan	99.5102	75.42	57.75643105
Turkish	tur	99.1382	75.0682	57.71716408
Russian	rus	99.8759	74.9274	57.67587165
Finnish	fin	99.666	78.7407	57.37806969
Norwegian Nynorsk	nno	99.5644	75.5234	57.24128266
Belarusian (Taraschkewiza)	be-tarask	99.2149	74.2289	56.86342429
Polish	pol	99.2439	74.0341	56.68512929
Luxembourgish	ltz	99.4245	72.6806	56.10486775
Icelandic	isl	99.2623	73.4829	55.3839586
Czech	ces	99.1211	72.1967	55.30869359
Ukrainian	ukr	99.789	72.6316	54.70052731
Lithuanian	lit	99.0375	73.0883	54.427767
Modern Greek	ell	99.197	71.8494	53.95981687
Slovak	slk	99.6601	68.6813	50.55039927

Figure 11. Accuracy table for different N-Gram approaches

The comparison of language detection system while using lowest degree of N-grams i.e. Mono-grams and various other higher order of N-grams shows that the accuracy is decreasing considerably with the addition of every higher degree N-grams into the approach (Figure 12).

While using only Mono-grams the accuracy is highest ranging from 99% to 97%. While the accuracy when using Mono, Bi and Tri-grams is considerably low ranging from 80% to 50%.

One common phenomenon while detecting languages is the match rate of N-grams in the train and test samples. It makes a big difference when higher order of N-grams is used. Consider the words 'bat' and 'cat', the monograms will only have one difference ('b' and 'c') while adding Bi-grams to it will make few more differences in the N-grams and hence the match rate will be affected.

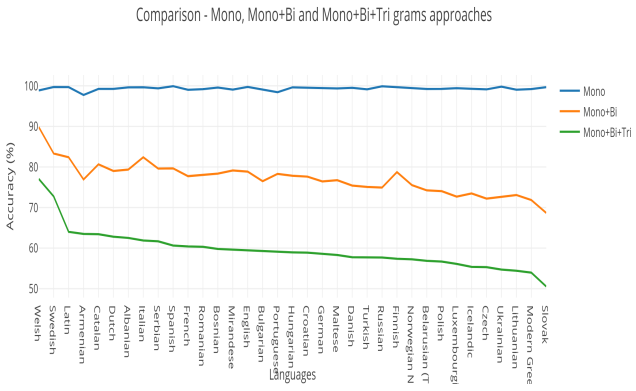


Figure 12. Accuracy comparison for different N-Gram approaches

5. LIMITATIONS

The proposed model uses Mono, Mono/Bi and Mono/Bi/Tri-Gram approach but higher accuracy can be achieved by implementing N-Grams of higher degree. The accuracy of the approach can be better estimated with other datasets. Accuracy of the approach can be improved with the high volume of training data.

6. REFERENCES

[1] Ali Selamat , Nicholas Akosu.*Word length algorithm for language identification of under-resourced languages.*

[2] W. B. Cavnar and J. M. Trenkle. *N-gram-based text categorization.* In Proceedings of SDAIR-94, the 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 161.175, Las Vegas, Nevada, U.S.A 1994.

[3] Izumi Suzuki, Yoshiki Mikami, Ario Ohsato, Yoshihide Chubachi. "A language and character set determination method based on N-gram statistics." ACM Transactions on Asian Language Information Processing (TALIP), 2002: 269-278.

[4] Martin Thoma, *The WiLI benchmark dataset for written language identification* arXiv:1801.07779v1 [cs.CV] 23 Jan 2018

[5] Bruno Martins, Mário J. Silva. *Language identification in web pages.* 2005 ACM symposium on Applied computing Santa Fe: ACM New York, NY, USA, 2005.76-768

[6] Yew Choong Chew, Yoshiki Mikami, Robin Lee Nagano, *Language Identification of Web Pages Based on Improved N-Gram Algorithm* IJCSI International Journal of Computer Science issues, Vol. 8, Issue 3, No. 1, May 2011 ISSN (Online): 1694-0814 www.IJCSI.org

[7] Chew Y. Choong,Yoshiki Mikami, C. A. Marasinghe and S.T. Nandasara. *Optimizing n-gram Orderof an n-gram Based Language Identification Algorithm for 68 Written Languages.* The International Journal on Advances in ICT for Emerging Regions 2009 02 (02) : 21 - 28

[8] Erik Tromp, Mykola Pechenizkiy. *Graph-Based N-gram Language Identification on Short Texts.* Department of Computer Science, Eindhoven University of Technology P.O. Box 513, 5600 MB, Eindhoven, The Netherlands.