# A Survey of Language Identification Techniques and Applications

Archana Garg
Department of Computer Science and Applications, Panjab University, Chandigarh, India
id.archana@yahoo.co.in

Vishal Gupta
UIET, Panjab University, Chandigarh, India
vishal@pu.ac.in

Manish Jindal
PURC, Muktsar, Punjab, India,
manishphd@rediffmail.com

*Abstract*—**Language Identification is the process of determining in which natural language the contents of the text is written. Language identification is always been an important research area which has been carried out from early 1970's. Still it is a fascinating field to be studied due to increased demand of natural language processing applications. In many applications, it works as a primary step of some larger process. In this paper, a number of applications are outlined where language identification is working successfully. Language Identification can be done using two types of techniques: computational techniques and non-computational techniques. Computational techniques are based on statistical methods and requires large set of training data for each of the language while non-computational techniques require that researcher must have extensive knowledge about the language to-be-identified. In this paper, a brief review of the few papers is presented which outlines the various statistical and non-statistical techniques that have been applied by the different researchers for language identification. Besides it, different researchers performed language identification for different type of documents such as monolingual, multilingual, long and short and for a particular set of languages.**

*Index Terms*—**Language Identification, Natural language processing, statistical methods, n-gram model, feature extraction.**

## I. Introduction

Language Identification is the process of finding in which natural language the contents of the text is written. Language identification is an extensive research area as it is most considered in natural language identification applications such as machine translation, information retrieval, summarization and question answering etc. which require prior language identification before processing. Language identification falls into two approaches: 1) non-computational; 2) computational. Non-computational approaches requires the researchers to have sufficient knowledge about the language to be identified such as diacritics and symbols, most frequent words used, character combinations etc. while computational approaches rely on statistical techniques rather than linguistic knowledge to solve related problems. It requires large training dataset for each to be identified language. Statistical approaches are categorized into two parts: 1) Training part and 2) classification part. In the training part, the feature extraction from the given training dataset, known as corpus, is done. In the classification part, the similarity measure between the training profile and the testing profile is found out and the most similar language is known as the language of the document. The general architecture of the language identifier is given by Padró and Padró (2004) in fig. 1 [54].

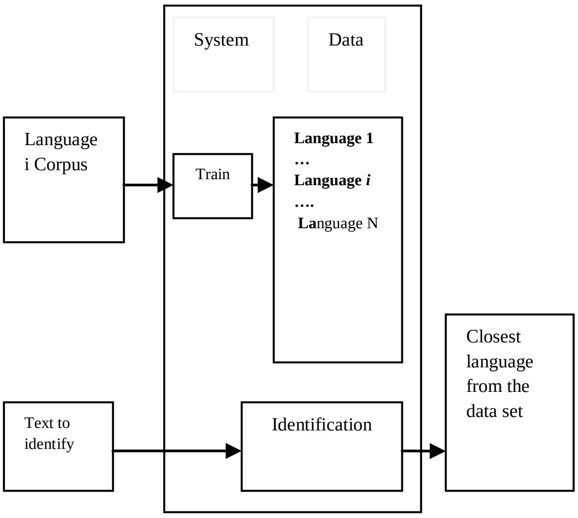

Fig.1: General architecture of a language identifier [54]

Different researchers have focused on different aspects of language identification such as identification of monolingual documents, multilingual documents, short documents, long documents, search engine queries, microblog posts, printed documents etc. In the next

© 2014 ACADEMY PUBLISHER
doi:10.4304/jetwi.6.4.388-400

section, a brief review of few of the papers has been presented which outlines the various techniques used by the different researchers for different language identification tasks.

## II. Language Identification Techniques

### *A. Language Identification Systems Using N-Gram Approach:*

Yang and Liang (2010) proposed a new approach for identification of natural language, i.e., joint approach based on N-Gram frequency statistics and Wikipedia. In this approach, the identification of languages in multilingual documents has been focused. Most of the methods of language identification require the large corpus as a training data for determining each language. But the joint approach elevated this problem by requiring a single language corpus i.e. local English corpus for all the languages to be identified. This approach is divided into two steps: (1) Implementation of segmentation algorithm on the text which uses N-Gram frequency statistics for segmentation into language-specific units; (2) Determining the language of each unit by utilizing the different language versions of Wikipedia. The data flow diagram of this approach is given in fig. 2. Wikipedia is well established database which supports 262 language versions of the documents and using it to natural language identification gives satisfactory accuracy rate. Multiple experiments done using this joint approach gives approximately 100% accurate results. [1]

Selamat (2011) proposed an improved N-Gram approach for the language identification of web documents. Improved N-Gram approach is based upon the combination of two conventional approaches, i.e., original N-Grams approach and a modified N-Grams approach. Original N-Gram approach uses the approach of Cavnar and Trenkle (1994) [5] which is based on rank order statistics of N-Gram profiles. The ONGA approach first finds the distance between training and testing data and the data with shortest distance is selected as the language of the document as given in fig. 3. One more enhanced approach i.e. modified N-Gram approached is proposed by Choong et al. [6] which removes the problem of dimensionality in ONG approach. The MNG approach uses the Boolean rate strategy to find the match between training and testing N-Gram profile. The Boolean rate is one if there is a match otherwise zero. After all the comparisons the match rate is found out and the language with highest match rate is known as the language of the document as given in fig. 4.

The improved N-Gram approach takes the advantage of ONG and MNG approaches by using N-Gram frequency of the testing profile and feature position of the training profile. The language of the document with maximum N-Gram frequency and minimum feature position is selected as the type of language. If it does not found then N-Gram frequency and feature position will be calculating again by reducing the size of testing data sets with 10%. Figure 5 shows the flow diagram of ING approach.

This approach is also experimented on multilingual sentence based identification of web pages. Recall, Precision, F-score are the measures for evaluation. The datasets of 1000 pages of each language is collected from BBC (British Broadcasting) news website for training and testing. The approach mainly works to validate the documents belonging to Arabic and Roman scripts. On evaluating the results, on web page identification is found 100% by using ING and MNG approach as compared to ONG with 66.60% (Precision), 61.41% (Recall) and 61.63% (F-Score) as well as on sentence based identification is 97.95% as compared to 39.57% (ONG) and 97.09% (MNG) approach.[7]

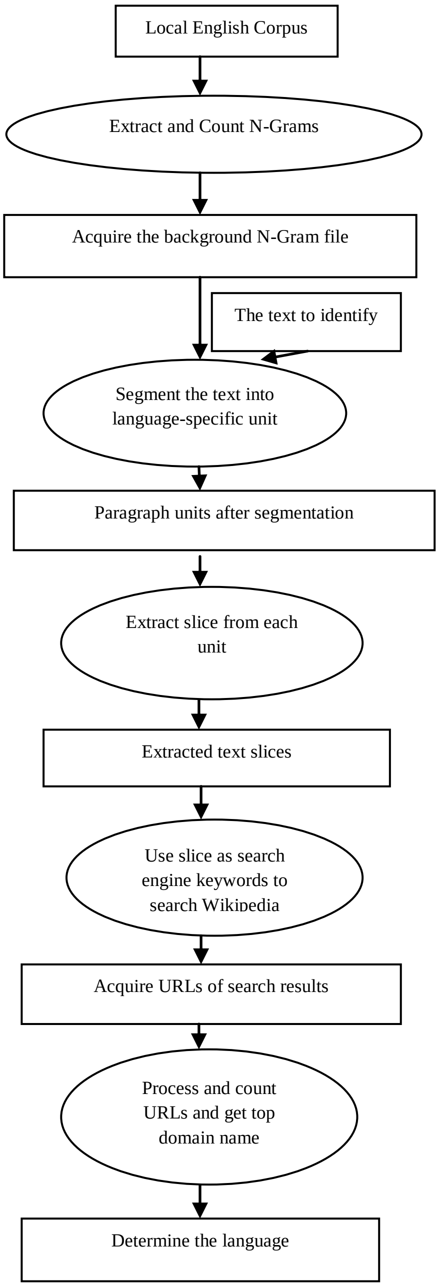
Local English Corpus
Extract and Count N-Grams
Acquire the background N-Gram file
The text to identify
Segment the text into language-specific unit
Paragraph units after segmentation
Extract slice from each unit
Extracted text slices
Use slice as search engine keywords to search Wikipedia
Acquire URLs of search results
Process and count URLs and get top domain name
Determine the language

Fig.2 Data flow of the proposed approach [1]

© 2014 ACADEMY PUBLISHER

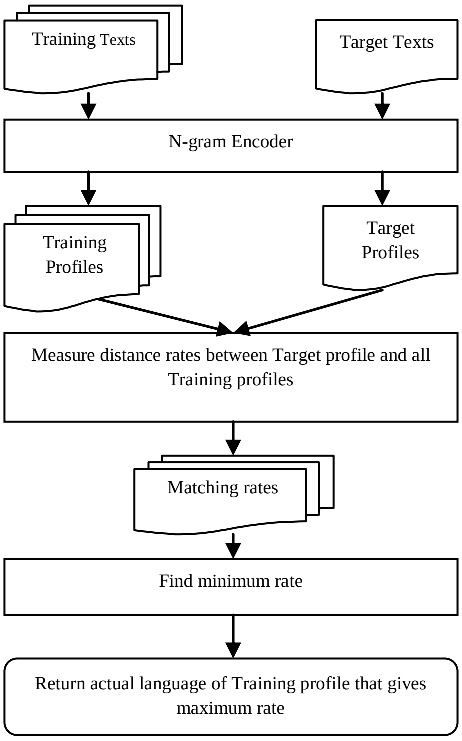
Training Texts
Target Texts
N-gram Encoder
Training Profiles
Target Profiles
Measure distance rates between Target profile and all Training profiles
Matching rates
Find minimum rate
Return actual language of Training profile that gives maximum rate

Fig.3 System flow of the original N-grams [5]

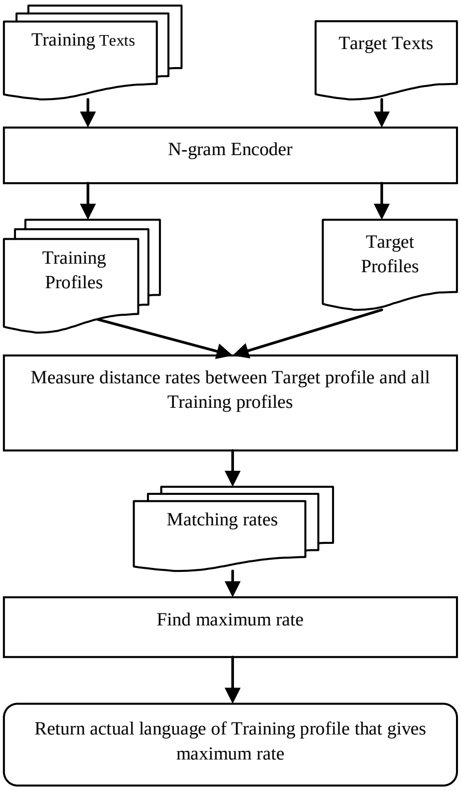
Training Texts
Target Texts
N-gram Encoder
Training Profiles
Target Profiles
Measure distance rates between Target profile and all Training profiles
Matching rates
Find maximum rate
Return actual language of Training profile that gives maximum rate

Fig.4 System flow of language identification process [6]

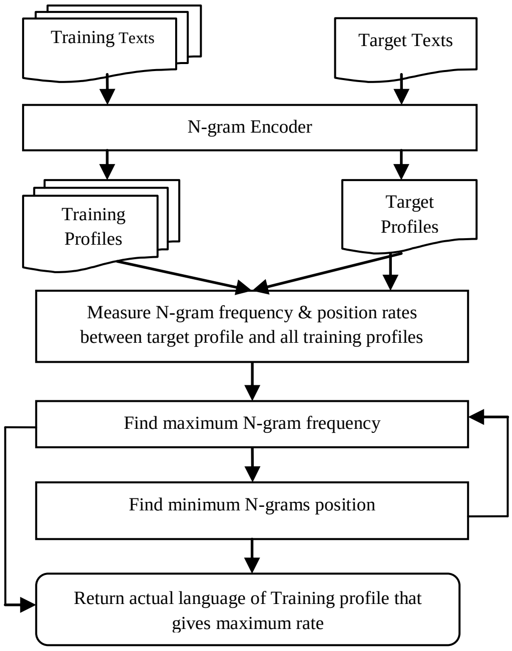
Training Texts
Target Texts
N-gram Encoder
Training Profiles
Target Profiles
Measure N-gram frequency & position rates between target profile and all training profiles
Find maximum N-gram frequency
Find minimum N-grams position
Return actual language of Training profile that gives maximum rate

Fig.5 The flow of the proposed improved N-grams (ING) approach [7]

Ng and Selamat (2011) have studied three identification methods i.e. distance measurement method, Boolean method and Optimum profile method. Distance measurement follows the approach given by Cavnar and Trenkle (1994) [5] . In this approach, training and testing profile are generated using N-Gram frequencies. These profiles are sorted on the basis of rank from the most frequent to least frequent. The distance is found out by comparing training and testing profiles and the minimum distance profile is known as the winner. Another method is the Boolean method [6] which is based on matching rate between training and testing profile for matching Boolean values are found either zero or one. If the match is found between testing and training profiles N-Grams then it returns one otherwise zero. The matching rate is found out and the highest matching rate is considered as the language of the document. Distance measurement suffers from the dimensionality problem and Boolean method fails when there is same N-Gram frequency of two or more languages. Optimum profile gives accurate results it uses both N-Gram frequency and N-Gram position features. Converge point is found out and the language with smaller converged point is selected as the language of documents. On evaluated these methods on European data sets of 23 languages, optimum profile gives maximum results of 91.52% as compared to distance measurement (72.98%) and Boolean method (85.01%).[8]

### *B. Language Identification Systems Using Profile Features:*

Padma et al. (2009) proposed an algorithm for identification of language in an Indian multilingual printed document containing the text in three languages – English (General Language), Hindi (National Language) and Kannada (Regional/State Language). The algorithm uses the local geometric approach to identify the type of language. The algorithm first fixed the bounding box for each text line and extracted the top and bottom profile features. The values for topmaxrow, topmaxrowno, bottommaxrow, bottommaxrowno are computed for each language and stored in a knowledge base. During testing, the feature values of the document images are calculated and compared with the values stored in the knowledge base and the correct type of the language is identified with the help of rule based classifier. The training data set of 800 text lines is used which comprises of 300, 200 and 300 text lines from Kannada, Hindi and English language. When evaluated on a set of 600 text lines, the system gives performance of average 95.4%.[2]

Vijaya and Padma (2009) proposed another algorithm based on feature value profiles to determine the language of multilingual document comprising of three languages- Kannada, English and Hindi. In this algorithm, top and bottom profile features of the text lines helped in calculating the feature value coeffprofile of each language which is stored in the knowledge base and worked as a test bed for testing data. The testing data feature values are compared with the values stored in the knowledge base and rule based classifier is used to determine the type of language. Multiple experiments shows the accurate results of average 96.6% [3].

### *C. Language Identification Using Different Classifiers:*

Baldwin and Lui [2010] have focused on language identification of monolingual documents. They have proved that the task of language identification is quite difficult over the datasets with larger number of languages with fewer amounts of training data per language. The experiments are made on three different sized datasets, i.e., Eurogov, TCL and Wikipedia having great variance in language of documents with different character encoding. It has been found that without [i] character encoding detection, the task of language identification is possible. In this paper, the authors [ii]focused on which tokenization strategy and classification model gives the best performance. Two tokenization strategies, i.e., Byte N-Gram based and Codepoint N-Gram based are used for document representation. A number of language identification models such as nearest neighbor (COSinn, SKEWinn, OOPinn), nearest prototype (COSam, SKEWam, Naive Bayes, Support Vector Machines are used. Experiments are made on 42 distinct classifiers with 7 models, 2 tokenization strategies( bigrams vs. codepoints) and 3 token n-gram orders( unigrams, bigrams and trigrams).s The models are evaluated using micro-averaged and macro-averaged Recall, Precision and F-score. The average performance per document is indicated by micro-averaged scores as well as average performance per language is indicated by macro-averaged scores. On evaluation, it is found that a simple 1-NN model with cosine similarity or an SVM with a linear kernel using a byte bigram or trigram document representation gives best performance.[4]

Gottron and Lipka (2010) have studied various approaches to identify the language of short, query-style text and compared which approach works better. In datasets they included single words and slightly longer texts. For determining the language of text, they use n-gram approach given by Cavnar and Trenkle[5], Na ïve Bayes classifier, Multinominal, Markov Processes given by Vojtesk and Bielikova[22], Frequency-Rank, Vector Space, LC4J approach which finds the cosine similarity between vector representation of n-grams of testing text and reference text in different language. For training, the 1,102,410 reference documents of news headlines are collected from Reuters Collections CV1[23] in English language and from CV2 in other languages such as German, Danish, Spanish, French, Dutch, Italian, Norwegian, Portuguese and Swedish. 20.048 single and unambiguous words of an average 8.1 characters long are extracted from bilingual dictionaries, i.e., from English to other language. All the above mentioned algorithms are used on unused Reuters headlines and single dictionary words for detecting the language and accuracies are found out for all settings of n. The authors found Na ïve Bayes classifier as the best approach for query style texts for large value of n. Markov and Frequency-Rank model drops its performance as the n-grams increases. LC4J does not detect any language in many cases.[24]

Botha and Bernard (2012) have proved that the classification accuracy of Language Identification is affected by various factors such as size of text fragments to be identified, the size of training data available, the classification features and the algorithm used and the language similarity to be identified. Larger fragments of text can improve classification accuracy as given by [32]. Different sizes of input text affect the classification accuracy as proved by [33]. The performance is improved when the size of the test segment is increased [34]. Training with different corpus sizes improves classification accuracy as found by [35]. The performance of different classifiers depends on the above factors mentioned. A comparison of different classifiers is given by [36]. This study is based on the 11 official languages of South Africa. The results of text based LID on South African languages have been published by [37] and a preliminary report is given by [38] on the current work. These 11 official South African languages consist of 9 languages from Bantu family of languages and 2 from Germanic family (English and Afrikaans). Bantu family of languages includes 4 in Nguni subfamily; 3 in Sotho subfamily and 2 of outside these subfamilies (Tsonga, Venda). Data is obtained from Professor D.J. Prinsloo of the University of Pretoria in all these languages using Web Crawler [39] and encoded in UTF-8 format to support special characters found in these languages. As the n-gram based model works well than word based models both for the short and larger text fragments [40] so the authors adopted the n-gram based approach as feature set. Besides it, they worked with n=3 and n=6. Three main classifiers as Na ïve Bayes classifier [41], difference-in-frequency classifier based on n-gram rank ordering method [5] and SVM classifier are used for text based LID. SVM classifier used the available software module with full SVM functionality [42]. Three window sizes are used for classification: 15 chars window,100 chars window and 300 chars window. 15 chars window is suitable for short text as it represents only two or three words. 100 chars window is more

© 2014 ACADEMY PUBLISHER

accurate in case of long sentences and 300 chars window (paragraph classification) is highly accurate. Text on datasets of each language is divided into 10 subsets out of which 9 are used for training and one is used for testing. During testing, classification is performed to find the error rate. The average results obtained from the 10 folds are determined to give an overall error estimate. On evaluating the results, it is found that the size of n-gram affects a lot in classification accuracy. In this study, 3-gram SVM model performs better as compared to other 3-gram models but shows similar results as compared to 6-gram NB classifier and outperformed 6-gram difference-in-frequency classifier on smaller character window. For larger character windows, both 6-gram based classifiers work well than SVM. Size of training data also affects as larger chars window gives more accurate results than short windows. In case of the similarity of languages to be identified, it is found that Sotho and Nguni language families show major error results. But when they are combined together in confusion matrix, it gives better results.[43]

### *D. Language Identification Using Baseline Approach:*

Milne et al. (2012) has proved that the language identified by short documents and long documents are different. The authors implemented two approaches - baseline approach and trigrams method on long as well as short documents from the data sets collected from Wikipedia and Europarl. Four languages are studied English, German, Spanish, and French. In baseline approach, a language profile of each language is built which contains top most frequent words. At testing, unique words in the test documents are compared with training profiles and language with more common words are recognized as the language of document. In trigrams method, the language profile is constructed using only trigram words. If the words are shorter than three grams then they are padded with space on the right. On testing, it is found that both approaches give good results on long documents (more than ten words) as compared to short documents (fewer than or equal to ten words). [9]

### *E. Language Identification Using Hybrid Approach:*

As language identification can be done by using the two approaches: 1) Non-statistical approach and statistical approach. Non statistical approaches are basically linguistic approaches which require complete knowledge about the rules of language used. Statistical approaches are basically machine learning approaches which requires less human efforts. In this study, Takçı and Ekinci (2012) uses hybrid approach i.e. combination of linguistic and statistical approach. The feature set is derived from the linguistic knowledge that contains letters and diacritics. These letter and diacritics are transformed into relative frequency by using statistical approach. The vector space model has been selected as presentation model For presenting the transformed data [47]. Each training and testing data is presented by the document letter frequency vector. This frequency vector is suitable for obtaining feature set of nine languages (English, French, German, Dutch, Italian, Portuguese, Turkish, Spanish, Swedish). Weighing factor is also used for increasing performance. The letters which passes in more languages will have small weighing factor. Besides it three classification algorithms are used- C-SVC (SVM for classification), MLP (Multilayer Perception), LDA (Linear Discriminant Analysis). SVM is a kernel based classification algorithm [10], MLP is the neural net based algorithm [11], LDA is a statistical based classifier [12]. The data set of 1800 documents collected from European Corpus Initiative (ECI) multilingual corpus [48] are used for experiments. Out of 1800 documents, 70% documents are used for training and 30% of testing of different text sizes such as 30, 60, 90, 100, 120, 180, 500 characters. For data mining tests, Tanagra 1.4.2 [13], machine learning software is used. On comparing the performance, it is found that C-SVC is best as per accuracy and LDA is best as per speed.[14]

### *F. Language Identification Using Character N-Gram Approach and Microblog Characteristics:*

Carter et al. (2011) have studied the language identification on microblog posts. The language identification of microblog posts is quite tedious task due to the short and idiomatic text as well as usage of mixed language in posts. The authors implemented TextCat algorithm based on N-gram approach described by Cavnar and Trenkle(1994) for training the language models on different datasets. Two language models: 1)Out-of-the-box which works as baseline and uses training data supplied by TextCat; 2) microblog which uses training sets of posts to retrain TextCat. Besides it, the authors identified five microblog characteristics which help in language identification: 1) blogger prior; 2) link prior; 3) mention prior; 4) tag prior; 5)conversation prior. These priors are combined in post dependent as well as independent way. In post dependent way, each post is observed individually and a post-dependent model, i.e., linear combination model is used to vary the weights of the priors that give optimal classification results for that specific post. This model uses two confidence metrics to find the language of post: the beam confidence and the lead confidence. In a post independent way, all posts are treated equally and post independent combination models are used for prior combination. These models include linear interpolation with post-independent weight optimization and majority voting. The dataset of 1000 microblog posts are collected from microblog platform, Twitter in five languages (Dutch, English, French, German. Spanish) based on their location. These 1000 tweets are split into training set of 400 tweets (for Textcat training), a development set of 100 tweets (for weight optimization) and test set of 500 tweets. These tweets are evaluated and extracted using different priors in five languages. On evaluating, it is found that the use of priors increase the accuracy 5% as that of baseline approach and post dependent combination of the priors achieves best performance.[15]

© 2014 ACADEMY PUBLISHER

### *G. Language Identification Using Artificial Neural Network:*

Dubaee et al. (2010) developed a method for classification of English and Arabic language identification by using the combination of wavelets and artificial neural networks (ANN). This method extracts required information from the web by identifying the language of users' query. The users' queries may be the words, short and long sentences, documents etc. The users' query is first converted into signals by using Unicode standard. These signals are classified into English and Arabic language according to the classifier used and feature extraction tool. The classifier is a feed-forward artificial neural network and wavelet transforms are used as feature vector. The methodology for classification of the processed signals is given in figure 6.

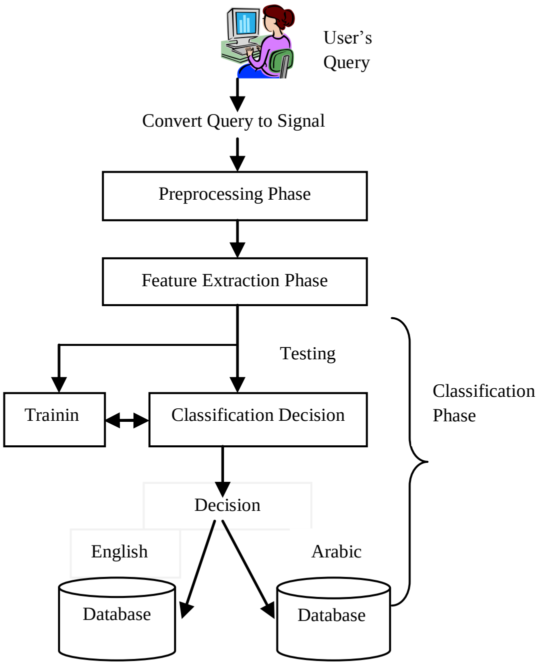

Fig.6 Methodology for the classification of the processed signal [16]

In the preprocessing phase, a good filter is selected from wavelet filters. 42 wavelet transform filters are evaluated out of which 3 filters, i.e., Haar, Bior 2.2, Bior 3.1 are found suitable for multilingual web information retrieval. In the feature extraction phase, the signals are mapped into feature vector by applying the tools, i.e., fast wavelet transform (FWT) decomposition with three filters and Power Deviation (PD) method[49] as given by the figure 7.

In the classification phase, the feature vector mapped by the feature extraction phase is recognized by ANN as given in the figure 8.

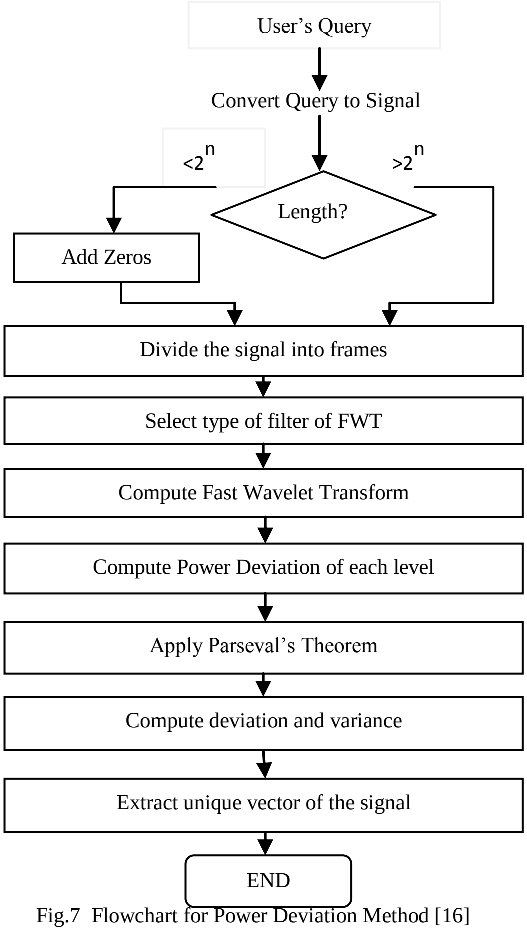
User's Query
Convert Query to Signal
<2^n
>2^n
Length?
Add Zeros
Divide the signal into frames
Select type of filter of FWT
Compute Fast Wavelet Transform
Compute Power Deviation of each level
Apply Parseval's Theorem
Compute deviation and variance
Extract unique vector of the signal
END

Fig.7 Flowchart for Power Deviation Method [16]

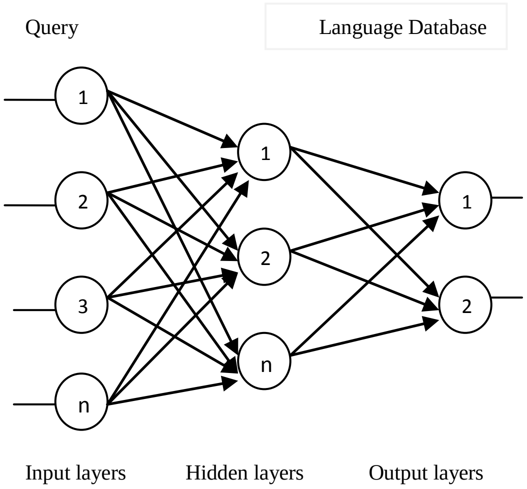

Fig. 8 The proposed architecture using ANN of PD method for English and Arabic languages [16]

The 104 WHO (World Health Organization) lines are used as dataset; out of which 26 lines from each language are used for training and rest are used for testing. On evaluation, it is found that best performance is given by Haar filter with 76.9% as classification ratio and time training 9.5 seconds.[16]

© 2014 ACADEMY PUBLISHER

### *H. Language Identification Using Langid.py Tool:*

Lui and Baldwin (2012) presented an Off-the-shelf language identification tool named as Langid.py. The authors uses Naïve Bayes classifier with a multinominal event model [17] and mixture of byte N-grams (1 ≤n≤4) for training it. Langid.py is full package consisting of two support modules: LDfeatureselect.py and Train.py. This tool is distributed with an embedded model which covers 97 languages and trained using the multi domain language identification corpus of Lui and Baldwin (2011) [50]. It performs tokenization and feature selection over the input document using Aho-Corasick String Matching algorithm [51] which uses Deterministic Finite Automation (DFA) states for processing the inputs. LDfeatureselect.py implements LD feature selection which is used to make the difference in Information Gain with respect to language and domain. This tool can be used in 3 ways : 1) command line tool which accepts the input through interactive queries; 2) Python library which provides a function that accepts text as input and finds the language of the text; 3) Web service which permits identification of language by means of HTTP PUT and HTTP POST request which return JSON-encoded responses Langid.py is evaluated on a number of language data sets such as Eurogov, TCL, Wikipedia [4], samples from a bio medical parallel corpus EMEA [18], Europarl [19]. The author compared the performance of langid tool with another language identification tools such as TextCat [5], LangDetect [52] and CDL [53]. On comparison it is found that langid.py is fastest, and more accurate than other tools. It also woks will on short input text such as microblog messages. Two data sets of twitter messages are used such as T-BE [20] and T-SC(Carter et al. , to appear) over both data sets, it shows better accuracy than other tools. [21]

### *I. Language Identification Using Different m=Models:*

Ceylan and Kim (2009) have studied the problem of identifying the language of search engine queries. This task is quite complicated due to the short length of queries. For this, the authors first proposed a method for automatically generating the dataset. The dataset has been constructed by retrieving the queries of 10 different languages along with the clicked urls from the Yahoo! Search Engine for the period of three months. These queries are then annotated and classified into three categories: 1) queries that does not contain any foreign terms; 2) queries that contain foreign terms but still be bringing web pages in same languages; 3) queries belonging to other languages. After that, three models, i.e., Statistical, Knowledge based and Morphological are implemented for language identification. These models are trained on Europarl Collection [19] and on the 10 languages. Statistical model uses a character based n-gram feature. Knowledge based model uses word based unigram feature. Morphological model uses affix information feature described in [25] gathered from the corpora for each language. Rank-order method is also implemented for comparison purpose. 3500 human annotated queries are tested on these models (consisting of 350 queries belonging to category-1 of each language)and shows that the statistical model gives the best results than all other models with the exception of some languages such as French, Spanish and Italian for which knowledge base model is best. After obtaining these results, the authors built a Decision Tree Classifier which improves the results by combining the output of previous models together with their confidence score. This classifier is trained with varying size of data (500, 1000, 5000, 10000 instances) of automatically annotated queries and evaluated on the same test set of 3500 human annotated queries. On evaluation, it is found as most satisfactory model giving outstanding results. This works better as the size of training data increases. A second decision tree classifier is also built by adding the non-linguistic feature to the existing system. This feature identifies the language that the user actually wanted for the results. So Classifier-2 boosts up the accuracy of results. [26]

### *J. Language Identification Using Centroid Based Approach:*

Takçı and Güngör (2012) have solved the language identification problem by applying centroid-based classification approach. Centroid –based classifier is a vector space-based classifier [27]. The authors considered the various problems come under vector space-based models such as high dimensionality of feature set, large time complexity and accuracy problem. In this study, they tried to overcome all these problems. They proposed a method named as ICF ( Inverse Class Frequency) which operates on small feature set and have linear time complexity and yields more successful results. During training, the centroid vector is formed which contains the central value of each class. The central value corresponding to the term denotes an average value for this term in all the documents in this class and acts as a representative of the whole class with respect to the term. Two main methods are there forming centroid vectors [44] : 1) Arithmetic Average Centroid (AAC); 2) Cumuli Geometric Centroid (CGC). In AAC method, mean value of the corresponding term weights is calculated for finding central value .In CGC method, the sum of the term weights is used as central value. The variation of AAC and CGC method [45] can be used where the weights in the centroid vector of a class are computed by taking the ratio of document and class for the terms. The one important factor in the text classification system is the term weighting factor [46]. For assigning the weights to each term, ICF scheme is used here. ICF makes use of term frequency distributions with the classes. In ICF, the centroid values are formed using AAC method first; then necessary modifications are made on the centroid values to obtain the new centroids. During testing, to find the correct class of new data, the similarity of the sample to each centroid is calculated using cosine similarity measure and it is assigned to the class yielding the maximum similarity value. In this study, the ECI/MCI multilingual corpus is used. This corpus consists of 27 languages. 90% of the data is used for training and 10% for testing. The experiments are performed on 9

© 2014 ACADEMY PUBLISHER

languages in the corpus. Feature set includes 54 features and is based on characters rather than words. For evaluation, precision, recall and f-measure metric are used and it is found that IFC gives best results as compared to other approaches such as Short-term, n-gram method, SVM, mutual information and chi-square methods in respect to time and space complexity.[28]

### *K. Language Identification Using Supervised Algorithm:*

Mayer (2012) proposed an algorithm for the case of multisite Internet domains. The algorithm works well for the language identification of short documents. As the statistical algorithms does not work well for the language identification of short documents [29] so a supervised algorithm is made by the author. The basic idea behind the study is to find the language of user created documents, i.e., eBay messages and tweets by using the site language or user profile language with some supervised algorithm. As language identification of email messages and tweets is quite difficult due to the short length as well as slang and abbreviations used by the tweets so the author implements the approach that finds the first two and last two words of the documents and builds the frequency tables of each word according to the site and user profile language. For training, data is collected from eBay Member to Member (M2M) emails and by using streaming API to capture 1% sample of public status updates made freely available by Twitter commonly known as “Sprinkler” data feed[30]. The site is extracted from the email template and use of site language gives 96% accuracy in identifying the language. 15K emails from June 2010 in seven languages is used for training. Similarly, in case of Twitter status updates, the site language is extracted from using the language field as well as location field from the user profiles as the combined precision gives more accuracy than using the user language alone. After collecting the data, it is preprocessed by removing the words having digits and accented letters in M2M mails but abbreviations containing digits and apostrophes are permitted in tweets. Then frequency tables are built containing the site or user language counts for each of the first two and combined table for last two words of eBay data as compared to four separate tables for each of the header and footer words of twitter data. At testing, first two and last two words are taken from each document and the frequency of each word $f_{kl}(w_k)$ across all the languages are extracted from the count tables k(k=1,…4) and a linear weighted sum is calculated. For the given weight, $W_k$(k=1,…,4)=1, the language score is computed for each language L as follows:

$$S_L=W_1f_{1L}(w_1)/f_1(w_1)+W_2f_{2L}(w_2)/f_2(w_2)+W_3f_{3L}(w_3)/f_3(w_3) + W_4f_{4L}(w_4)/f_4(w_4)$$

Here $f_k(w)$ denote the sum of frequencies for a word w in table k across all the languages. The language with the highest score is determined as the language of the document. These scores are combined with the site language by using the formula as given below:

$$SLS= WS\ \delta(L)+SL$$

For both eBay M2M and twitter status data, the prechosen weight $W_S$ =1.75. δ(L) is the indicator function which indicates 1 if L is the site language otherwise 0. This algorithm is then tested on eBay M2M data collected from different time periods in different volume labeled by site language and human. This algorithm is compared to other two industry standard algorithms such as zipping and character n-gram Naïve Bayesian algorithm with and without site language and results are found out on eBay data which shows this algorithm with the site language is proved to be best as having low runtime and high precision. Similarly, this algorithm is tested on twitter data and compared with only Naïve Bayesian algorithm with and without site language. Zipping algorithm is not taken into consideration in case of twitter data as it gives poor runtime and language identification accuracy on eBay data. Besides it, results are obtained corresponding to the two site language strategies: a) site language= user profile language; b) site language= language from location and user profile both. On evaluating, it is found that Bayesian algorithm with site language proves to best with high precision as compared to other models. [31]

## III. Applications Of Language Identification Systems

Language Identification is an interesting problem. In many applications, it works as a primary step of some larger process. It provides the facility of using background information about the language and using specialized approaches in many natural language processing tasks that deals with the collection of texts written in different language. With the increase of international communication and business, systems are required for correctly identifying the language of documents (emails, letters, web pages etc.). The task of language identification is working in various areas of natural language processing.

1. One active company in the field (Lextek) sells a professional Language identification program which provides many applications for it as email routing and filtering engines, text mining applications, encoding of WWW pages, information retrieval systems and content based and language specific web crawlers and search engines [55].
2. The bilingual corpus from the Internet is acquired by Resnik (1999) with the help of language identification techniques. The system called Strand is developed which can look for parallel translated pages and help to construct a bilingual corpus [56].
3. LID methods can also be used in distribution of languages on web pages. Langer (2001) developed a web crawler to get representative numbers on the distribution of the language on the web pages [57].

© 2014 ACADEMY PUBLISHER

TABLE I.
COMPARISON TABLE SHOWING LANGUAGE IDENTIFICATION METHODS AND THEIR ACCURACY

| Method Used | Author & Year | Approach Used | Accuracy Rate(%) | Limitations(if any) | Comments |
|---|---|---|---|---|---|
| Segmentation algorithm used on the text for segmenting the text into language specific units | Yang and Liang (2010) | Joint approach based on N-gram frequency statistics and Wikipedia | 100% approx. | -Lack of threshold length of the paragraph.<br>-Small paragraph does not extract sufficient amount of N-grams | Single language corpus i.e. Wikipedia for identifying multilingual documents |
| Top and bottom profile features and rule based classifiers | Padma et al. (2009) | Local geometric approach and profile features | 95.4% | Limited font size and type considered | Three regional languages are identified-Hindi, English, Kannada |
| Feature value coefficients and rule based classifier | Vijaya and Padma (2009) | Characteristic feature of top and bottom profile of text lines | 96.6% | Not suitable for handwritten text | Three regional languages are identified-Hindi, English, Kannada |
| Based on 42 different classifiers | Baldwin and Lui (2010) | Byte N-gram and code-point N-gram based approach | 1-NN model with cosine similarity as well as SVM with linear kernel is most accurate | Only monolingual documents can be identified | Three different sized datasets from (Eurogov, TCL and Wikipedia) are trained and tested |
| Improved N-gram algorithm | Selamat (2010) | N-gram approach | Web page identification:<br>Pre. -100%<br>Recall-100%<br>F1 – 100%<br>Sentence based identification:<br>Average-97.95% | Computational cost and scalability limitations are not considered | - ING approach (combination of ONG and MNG approach)<br>- used also for sentence based identification of multilingual web page |
| Optimum profile method | Ng and Selamat (2011) | N-gram approach | 91.52% | Only monolingual web pages are identified | Optimum profile method is compared with distance measurement and Boolean profile method |
| -Padded trigram method<br>-Cavnar and Trenle Method<br>-Langdetect<br>-Top 1000 words | Milne et al. (2012) | Baseline approach | Long documents give more accuracy than short documents. 99% approx. | These methods does not work well on short documents | Long and short documents of Europarl and Wikipedia are evaluated |
| Three classification algorithms used-C-SVC, MLP, LDA | Tak ç̇ and Ekinci (2012) | Hybrid approach | C-SVC is best with 100% accuracy on text size more than 100 byte | Not successful for short sized text | Weight factor increased language identification accuracy |
| Content based identification | Carter et al. (2013) | N-gram approach with microblog characteristics | 95.9% | -fluent multilingual post<br>-named entity error<br>-prior effects<br>-language ambiguous | Microblog text in five languages identified |
| Power deviation method | Dubaee et al. (2010) | Wavelet transforms and neural network | 76.9% with time training 9.5 seconds with Haar filters | Network needs more training and training sets for getting better results | Three filters i.e Haar, Bior2.2 and Bior 3.1 are evaluated on English and Arabiclanguages |
| Langid.py tool | Lui and Baldwin (2012) | Byte N-gram approach used | More than 90% on different test datasets | Langid.py tool is less speedy than CLD | Short as well as long documents are tested and compared |
| Three classifiers used-<br>-Statistical<br>-Knowledge based<br>-morphological | Ceylan and Kim (2009) | -character based n-gram feature<br>-word based unigram feature<br>-affix information feature | Stats.-82.7%<br>Know.-70.3%<br>Morp.-26.0%<br>Rank -65.2%<br>Decision-tree classifier-94.5% | Limited features are proposed for automated query taxonomy | -Language of search engine queries is identified<br>-rank order method is used for comparison purpose<br>-decision tree classifiers boosts up the accuracy |

| | | | | | |
|---|---|---|---|---|---|
| ICF method | Takçı and Güngör (2012) | Centroid based approach | Pre.-97.1%<br>Recall-97.5%<br>F – 97.3% | Successful only for language specific characters | ICF is compared with short-term, n-gram and SVM, mutual information and chi-square method |
| Supervised algorithm | Mayer (2012) | Used frequency tables created using first two and last two words of documents | Bayesian algorithm with site language proves to be best | Suitable for only short text | Language of short documents such as eBay mails and tweets is identified |
| - Naive Bayes<br>-Multinominal<br>-Markov<br>-Frequency-rank<br>-Vectorspace | Gottron and Lipka (2010) | N-gram based approach | Naive Bayes classifier performs best with 99.44% accuracy rate on headlines | Markov and frequency Rank method gives poor performance | Language of query-style short text identified |
| - Naive Bayesian<br>-Difference-in-frequency<br>-Support vector machine | Botha and Bernard (2012) | N-gram based approach with n=3 and n=6 | 3-gram SVM model and 6-gram NB classifier performs best | Nguni and Sotho language subfamilies gives poor performance | Three window sized, i.e, 15 chars, 100 chars, 300 chars, data are evaluated |

4. Grefenstette (1995) tells that language is essential to be identified for morphological processing of data. The language of the document needs to be known before applying stemming or parsing techniques. In case of spell checking, if a lexicon is used then the program must need before which language specific rules are applied [40].
5. LID is a preprocessing task for many applications of NLP. Tasks such as summarization, question-answering, machine translation, text categorization etc. need to know the language of the text before processing. Language identification is more crucial in bilingual and multilingual society where NLP applications have to deal with different customers communicating in different languages.
6. One of the most important areas of LID is spoken language identification. In the area of telephone based information services such as customer service, phone ordering, information hotline, phone banking and other call-centre/Interactive voice response based services, speech recognition system is helpful to transfer the incoming call to the corresponding agent.
7. Spoken Language identification is also helpful in defending against global terrorism. Government has deployed complicated systems to monitor communications among suspicious subjects. LID technology can detect the sensitive languages used by the terrorists during telephone conversation and for sending messages from one place to another place.
8. LID is a typical signal modelling and classification task. It can be used in many other areas as biometrics, human-computer interaction and pattern recognition tasks.
9. Thomson and Liron (2002) have defined the use of automatic language identification in Unified messaging system for text-to-speech conversion. A unified messaging system includes a voice gateway server coupled to an electronic mail system and a private branch exchange (PBX). The voice gateway server provides voice messaging services to a set of subscribers. Within the voice gateway server, a trigraph analyzer sequentially examines 3-character combinations within a text message; determines occurrence frequencies for the character combinations; compares the occurrence frequencies with reference occurrence statistics modeled from text samples written in particular languages; and generates a language identifier and a likelihood value for the text message. Based upon the language identifier, a message inquiry unit selects an appropriate text-to-speech engine for converting the text message into computer-generated speech that is played to a subscriber [58].
10. Llitjos and Black (2001) showed that language identification can improve the accuracy of letter-to-phoneme conversion. As the authors found human language identification very hard so they built Letter to Language Model (LLM) as well as a language identifier. LLM is trained on text corpora of 18 languages. This language identifier applies LLM on the input data (word) and for every trigram of the word. It calculates the probability of it belonging to all the languages by multiplying them by the relative frequencies for those trigrams in each one of the languages. This language probability information improves pronunciation accuracy of proper names [59].
11. Language identification is an essential tool used in machine translation. Machine Translation is the ability of the machine which translates the text or speech from one natural language to another. It basically substitutes the words in one natural language for words in another. Machine Translation process is classified into three modules: 1) language identification module; 2) transliteration module; 3) translation module. The language identification module identifies the language of the text or document by using some statistical or non-statistical techniques.

© 2014 ACADEMY PUBLISHER

The transliteration module performs mapping of pronunciation and articulation of words written in one script into another script preserving the phonetics. The machine translation module changes the language by preserving meaning. Li et al. (2007) use language identification in a transliteration system to account for different semantic transliteration rules between languages when the target language is Chinese [60]. Huang (2005) improves the accuracy of machine transliteration by clustering his training data according to the source language [61].

12. Language identification is a special case of text categorization. Text Categorization is the task of assigning predefined categories to free-text documents. When the documents are to be classified according to the language then it is necessary first to identify the languages of the documents to be classified. Traditionally, the identification of written language was done manually by identifying frequent words and letters known to be characteristic of particular languages. But now, computational approaches have been applied to the problem to make this task easy.
13. Tromp (2011) have shown the use of language identification in multilingual sentiment analysis on social media. The author studied that the multilingual sentiment classification works accurately when the process is divided into four steps; LID, part-of-speech tagging, subjectivity detection and polarity detection. Models utilize the language specific information at all the three steps after LID. If the language determined is incorrect then this error will affect the further steps of multilingual sentiment analysis. [62]

## IV. Conclusions

In this paper, various novel approaches proposed by different researchers and applications of language identification have been reviewed. Language identification is done using n-gram techniques, Centroid based techniques, different classifiers based techniques, supervised techniques, profile feature based techniques, artificial neural network based techniques, hybrid techniques etc. Different approaches worked for different set of languages and different type of documents and gives high level accuracy for identifying the text. A number of limitations are outlined by different researchers which can be the base of future research.

## V. Future Work

A number of novel approaches are proposed by different researchers for language identification. Most of the researchers worked for language identification of monolingual documents (web pages, search engine queries, microblog posts, tweets etc.). Multilingual documents are less considered. It is also found that different classifiers used by them for classification of documents also affect the accuracy of language identification. It is also studied that language of short documents are quite hard to determine as compared to long documents. So, in future works, the issue of multilingual documents can be considered. A hybrid model can also be proposed that can successfully identify the short and long documents both with high accuracy.

## References

[1] X. Yang, W. Liang, An N-Gram-and wikipedia Joint Approach to Natural language Identification, *IUCS*, 2010.

[2] M. C. Padma, P. A. Vijaya, P. Nagabhushan, Language Identification from an Indian Multilingual Document Using Profile Features, *International Conference on Computer and Automation Engineering ICCAE*, 2009, pp. 332-335.

[3] P. A. Vijaya, M. C. Padma, Text Line Identification from a Multilingual Document, *International Conference on Digital Image Processing ICDIP*, 2009, pp. 302-305.

[4] 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 ACL,* Los Angeles, California, June 2010, pages 229-237.

[5] W. B. Cavnar, J. M. Trenkle, N-gram –based text categorization, *in proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval* , Las Vegas, Nevada, USA, 1994, pp. 161-175.

[6] C. Choong, Y. Mikami, , C. Marasinghe, S. Nandasara, , Optimizing n-gram order of an n-gram based language identification algorithm for 68 written languages, *International Journal on Advances in ICT for Emerging Regions* 2(2), pp. 21-28, 2009.

[7] A. Selamat, Improved N-grams Approach for Web Page Language Identification, *N.T. Nguyen (Ed.): Transactions on CCI V, LNCS 6910*, pp. 1-26, 2011.

[8] C. Ng and A. Selamat, Improving Language Identification of Web Page Using Optimum Profile, *J.M. Zain et al.(Eds.): ICSECS 2011, Part II, CCIS 180,* pp. 157-166, 2011.

[9] R. M. Milne, R. A. O'Keefe, A. Trotman, A Study in Language Identification, *ADCS' 12, December 05-06, Dunedin*, New Zealand, 2012.

[10] T. Joachims, Learning to Classify Text using Support Andctor Machines, *Kluwer, Boston*, 2002.

[11] Haykin, Simon, *Neural Networks: A Comprehensiand Foundation (2 ed.)*, Prentice Hall, ISBN 0132733501, 1998.

[12] R. O. Dua, P. E. Hart, D. H. Stork, *Pattern Classification (2nd ed.),* Wiley Interscience, ISBN 0-471-05669-3, 2000.

[13] Tanagra Software Web Site, http://eric.univ-lyon2.fr/-ricco/tanagra/en/tanagra.html, [Access Date, 15 January 2009].

[14] H. Takçı, E. Ekinci, Minimal Feature set in language identification and finding suitable classification method with it, *in Procedia Technology 1*, pp. 444-448, 2012.

[15] S. Carter, W. Weerkamp, M. Tsagkias, Microblog language identification: overcoming the limitations of short, unedited and idiomatic text, *Lang Resources & Evaluation* 47, pp. 195-215, 2013.

[16] S. A. Al-Dubaee, N. Ahmad, J. Martinovic, V. Snasel, Language Identification using Wavelet Transform And artificial Neural Network, *In International Conference on Computational Aspects of Social Networks*, 2010.

[17] A. McCallum and K. Nigam, A comparison of event models for Naïve Bayes text classification, *In proceedings*

© 2014 ACADEMY PUBLISHER

*of the AAAI-98 Workshop on Learning for Text Catagorization, Madison, USA,* 1998.

[18] J. Tiedemann, News from OPUS-A Collection of Multilingual Parallel Corpora with Tools and Interfaces, *Recent advances in Natural language Processing*, V:237-248, 2009.

[19] P. Koehn, *Europarl: A parallel corpus for statistical machine translation*, MT summit,11, 2005.

[20] E. Tromp and M. Pechenizkiy, Graph-Based N-gram Language Identification on Short Texts, *In Proceedings of Benelearn The Hague, Netherlands*, pages 27-35, 2011.

[21] M. Lui and T. Baldwin, Langid.py: An Off-the-shelf Language Identification Tool, *In Proceeding of the 50th Annual Meeting of the Association for Computational Linguistics*, Pages 25-30, Jeju, Republic of Korea, 8-14 July 2012.

[22] P. Vojtek, M. Bielikova, Comparing natural language identification methods based on Markov processes. *In: Computer Treatment of Slavic and East European Languages, 4th Int. Seminar*, pp 271-282, 2007.

[23] D. D. Lewis, Y. Yang, T. Rose, F. Li, RCV1: A new benchmark collection for text categorization research, *Journal of Machine Learning Research 5*, pp. 361-397, 2004.

[24] T. Gottron and N. Lipka, A Comparison of Language Identification Approaches on Short, Query-Style Texts, *Advances in Information Retrieval, Lecture Notes in Computer Science,* Volume 5993, pp. 611-614, 2010.

[25] H. Hammarström, A naïve theory of affixation and an algorithm for extraction, *in proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL, New York City, USA,* Association for computational Linguistics, pages 79-88, June 2006.

[26] H. Ceylan and Y. Kim, Language Identification of Search Engine Queries, *in proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore,* pages 1066-1074, 2-7 August 2009.

[27] G. Salton, *Automatic Text Processing: The transformation, Analysis and Retrieval of Information by computer,* Addison Wesley, 1989.

[28] H. Takçi and T. Güngör, A high performance centroid-based classification approach for language identification, *in Pattern Recognition Letters 33*, pp. 2077-2084, 2012.

[29] Rehurek, Radim and Kolkus, Milan, Language Identificaion on the Web: Extending the Dictionary Method, *Computational linguistics and intelligent text processing, Lecture Notes in Computer Science,* volume 5449/2009, 357-368, DOI: 10.1007/978-3-642-00382-0_29, 2009.

[30] Twitter, Streaming *API*, Online documentation at https://dev.twitter.com/docs/streaming-api.

[31] U. F. Mayer, Bootstrapped Language Identification for Multi-Site Internet Domains, *KDD'12,* Beijing, China, pp. 579-585, August 12-16, 2012.

[32] K. N. Murthy, G. B. Kumar, Language identification from small text samples, *The Journal of Quantitative Linguistics 13 (1),* pp. 57-80, 2006.

[33] A. Poutsma, Applying Monte Carlo techniques to language identification, *In: proceedings of Computational Linguistics in the Netherlands*, 2001.

[34] T. Dunning, Statistical identification of language, *Technical Report CRL MCCS, Computing Research Lab, New Mexico State University*, pp. 94-273, 1994.

[35] M. Padro, L. Padro, Comparing methods for language identification, *In: Proceedings of the XX Congreso de la Sociedad Espanola para el procesamiento del Lenguage Natural*, 2004.

[36] G. R. Botha, E. Bernard, *Text-based Language Identification for the South African Languages*, University of Pretoria, 2007b.

[37] H. P. Combrinck, E. C. Botha, Text-based automatic language identification, *In: Proceedings of the 6th Annual Symposium of the Pattern Recognition Association of South Africa*, 1995.

[38] G. R. Botha, E. Bernard, Factors that affect the accuracy of text-based language identification , *In: Proceedings of the 18th Annual symposium of the Pattern Recognition Association of South Africa*, pp 7-12, 2007a.

[39] G. R. Botha, E. Bernard, Two approaches to gathering text corpora from the world wide web, *In: Proceedings of the 16th Annual symposium of the Pattern Recognition Association of South Africa*, p. 194, 2005.

[40] G. Grefenstette, Comparing two language identification schemes, *In: Third International Conference on Statistical Analysis of Textual Data*, 1995.

[41] A. Binas, Markovian Time Series Models for Language Identification, *Project Report*,Available:http://www.cs.toronto.edu/abinas/csc2515report.pdf(online), 2005.

[42] C. Chang, C. Lin, LIBSVM: A Library for Support Vector Machines, Available: http://www.csie.ntu.edu.tw/cjlin/libsvm (last accessed: 30/07/2007; online), 2001.

[43] G.R. Botha, E. Bernard, Factors that affect the accuracy of text-based language identification, *In: Computer Speech and Language*, 26, pp. 307-320, 2012.

[44] H. Guan, J. Zhou, M., Guo, A class-feature-centroid classifier for text categorization", *In: Proc*. www.Madrid, 2009.

[45] S. Tan, An improved centroid classifier for text categorization, *Expert Syst. Appl. 35 (1-2),* pp. 279-285, 2008.

[46] E. Leopold, J. Kindermann, Text Categorization with Support Vector Machines: how to represent texts in input space?, *Mach. Learn.*, 46 (1-3), pp. 423-444, 2002.

[47] G. Salton, M. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, 1983.

[48] European Corpus Initiatiand Multilingual Corpus I (ECI/MCI) (2005), http://www.elsnet.org/resources/ecicorpus.html, page last modified 29-03-2005.

[49] J. W. Resende, M. I. R. Chaves, C. Pnna, Identification of Power Quality Disturbances Using the MATLAB Wavelet Transform Toolbox, *IPST Conference,* 2001.

[50] M. Lui and T. Baldwin, Cross-domain feature selection for language identification, *In proceedings of 5th International Joint Conference on Natural Language Processing,* pages 553-561, Chiang Mai, Thailand, 2011.

[51] V. A. Alfred and J. C. Margaret , Efficient string matching: an aid to bibliographic search, *communications of the ACM,* 18(6):333-340, June, 1975.

[52] N. Shuyo, Language Detection library for java, 2010.

[53] M. McCandless, Accuracy and performance of google's compact language detector, October 2011. [blog entry; last visited in October 2012].

[54] M. PadróCirera, L. PadróCirera, Comparing methods for language identification, *In Procesamiento del lenguaje natural,* Barcelona: Sociedad Española para el Procesamiento del Lenguaje Natural, pp. 155-161, 2004.

[55] S. Kranig, Evaluation of language Identification Methods, *Bakalárska práca,* Universität Tübingen, Nemecko, 2005.

[56] P. Resnik, Mining the Web for Bilingual Text, *In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL99),* 1999.

[57] St. Langer, Sprachen auf dem WWW, *In: Proceedings der GLDV-Jahrestagung*, pp. 85-91, 2001.

[58] H. C. A. Hyde-Thomson, R. Liron, Unified messaging system with automatic language identification for text-to-speech conversion, *US Patent 6,487,533 B2*, Google Patents, 2002.

[59] F. Llitjos and A. W. Black, Knowledge of language origin improves pronunciation accuracy of proper names, *In Proceedings of Eurospeech,* pages 1919–1922, 2001.

[60] Li, K. C. Sim, J. S. Kuo, and M. Dong, Semantic transliteration of personal names. *In Proceedings of ACL*, pages 120–127, 2007.

[61] F. Huang, Cluster-specific named entity transliteration, *In Proceedings of HLT-EMNLP*, pages 435–442, 2005.

[62] E. Tromp, *Multilingual sentiment analysis on social media*, Master's thesis, Dept. Computer science, Eindhoven University of Technology, 2011.