International Journal of Research in Engineering and Innovation (IJREI)

journal home page: http://www.ijrei.com

ISSN (Online): 2456-6934

A Review on Authorship Profiling Approaches

K. Kotaiah Swamy[1], E.Ravi Kumar[2]

[1]*Assistant Professor, Dept of IT, Vardhaman College of Engineering, Hyderabad*
[2]*Associate Professor, Dept of IT, Vardhaman College of Engineering,Hyderabad*

Abstract

Authorship analysis is a text analysis technique that is visualized mainly in three different techniques namely Authorship Profiling, Authorship Identification and Plagiarism Detection. In this paper a brief survey on the recent developments in the area of author profiling approaches were presented. Authorship Profiling is to ascertain various authors' characteristics like age, gender, native country, native language, and degree of education and personality traits by analyzing their writing styles. In recent times, Author Profiling is popular in the fields of forensic analysis, security and marketing. Based on the popularity of the Authorship Profiling problem, multiple solutions were proposed by various researchers across the globe. Several researchers used different types of features to identify the writing style characteristics of authors. The main focus of this survey is to predict the demographic features of authors such as age, gender and personality traits based on the text corpus written by various authors.

Keywords: Author Profiling, gender prediction, age prediction, personality traits, stylometric features, machine learning algorithms, Accuracy.

1. Introduction

Author profiling is the task of determining demographic features of authors like native language, education, gender, age and personality traits of an author by understanding their writing styles. Author profiling is an important technique in the present information era which has applications in marketing, security and forensic analysis. Authorship Identification problem divides into Authorship Attribution and Authorship Verification. Authorship Attribution determines the author of a given anonymous text from known writings of many authors. Authorship Verification finds whether the given texts were written by the particular author or not by considering the writings of a same author.

Plagiarism Detection detects whether a given document is original or not. This approach is broadly categorized as text alignment and source retrieval. Text alignment is a process of matching the contents in terms of passages between two documents. Source retrieval is a process of searching for the similar sources of a suspicious document.

Author profiling helps in crime investigation to identify the perpetrator of a crime by considering the characteristics of writing styles. Social web sites are an integral part of our lives through which, crimes are cropping up like public embarrassment, fake profiles, defamation, blackmailing, stalking etc. To identify the perpetrator it is useful by understanding the writing style of perpetrator using Author Profiling. Forensics is a field to analyze the style of writing, signatures, documents, and anonymous letters to identify the terrorist organizations. In the marketing domain the consumers were provided with a space to review the product. Most of the reviews were not comfortable in revealing their personal identity. In this case these reviews were analyzed to classify the consumers based on their age, gender, occupation, nativity language, and country and personality traits. Based on the classification results, companies try to adopt new business strategies to serve the customers. Author profiling is also beneficial in educational domain by analyzing a large set of pupil. It helps in revealing the exceptional talent of the students. It also helps in estimating the suitable level of knowledge of each student or a student group in the educational forum.

In general every human being has his own style of writing and the writing style will not be changed and he continues to write the same style in Twitter tweets, blogs, reviews, and social media and also in documents. Exploiting the writing style in order to find the authors profile is the

Corresponding author: K. Kotaiah Swamy
Email Id: kthswamy@gmail.com

main focus of this paper such as age, gender, personality traits which reflects the basic personality of a person that is to be analyzed by using the writing styles.

The prime differentiating fact of human behavior is solely depends on personality. The psychology research literatures established a model called Big Five Personality Model which is modeled for describing and recognizing the personality. The five personality traits which were discussed in literature are openness, conscientiousness, extraversion, agreeableness and stability. These five personality traits were observed to be prime focusing points in order to understand the personality of an author. Openness is a kind of experience which is related to creativity, tolerance, imagination, curiosity, appreciation for culture and political liberalism. Conscientiousness is a kind of a measure which gives preference for an organized approach in life. Looking for the company of others and expressing positive emotions by seeking stimulation to the external world is the third personality trait called extraversion. The fourth trait agreeableness focuses on a cooperative compassionate behavior in maintaining friendly, positive social relations. The last personality trait stability measures the mood swings and emotions and the tendency towards guilt, anger, depression and anxiety. These five personality traits were predicted from writings of the authors in [59]. This survey considered various stylometric features and techniques discussed in [60].

In this paper an attempt is made to present the research advances in Author Profiling area from the last decade. Section 2 outlines the stylometric features and various feature selection and extraction methods to identify the most appropriate feature set. In Section 3 the profiling methods were discussed. In section 4 the evaluation of Authorship Profiling techniques were presented. In section 5 the results of various researches contribution were discussed.

2. Stylometric Features

In the literature the research on Authorship Profiling proposed a catalog of stylistic features to enumerate the writing style of authors. Every category of features has their own importance to predict demographic features of authors. The combinations of these features were also used to discriminate authors writing style. Table 1 shows the basic categories of features, the tools required and the resources for features measurement.

2.1 Lexical features

In general the research on text data considers the text as a sequence of tokens and each token corresponds to a word. Several functions in the grammar enumerate the variety of vocabulary of the text. Out of such functions type, token ratio is introduced by some researchers [8, 56]. One such function which finds the ratio between the total number of unique stems and the total number of words after applying stemming. The researchers excluded stop words from above said ratio and stemming was carried out using nltk implementation of the Snowball algorithm [56]. In another exploration the authors [2, 5] used hapax legomena (i.e., words occurring once) and Hapax dislegomena (The number of words that occur twice) to increase the vocabulary richness.

Table 1: Stylometric Features

Features		Tools and Resources
Character-based features	The total number of characters	Character dictionary
	The number of capitalized letters	Tokenizer
	Character N-Grams	Text::N-grams, Feature selector, N-gram Tokenizer
	The ratio of capital letters to total number of characters	Tokenizer
	The ratio of white-space characters to total number of characters	Tokenizer
	The ratio of tab spaces to total number of characters	Tokenizer
	The ratio of white spaces to nonwhite spaces	Tokenizer
	The ratio of capital letters to the lower case letters	Tokenizer
	The ratio of numeric data in the text	Character dictionary
	Frequency of special characters	Tokenizer
Lexical features	Total number of words	WordTokenizer, [Stemmer, Lemmatizer]
	Type/token ratio (verbosity)	WordTokenizer, [Stemmer, Lemmatizer]
	Word N-Grams	WordTokenizer, Ark-Tweet-NLP Tokenizer,
		GATE Twitter-specific Tokenizer, N-gram Tokenizer
	The number of positive emotional words	Tokenizer, RiTaWordNet
	The number of negative emotional words	Tokenizer, RiTaWordNet
	The number of patriotic words	Dictionary
	Number of acronyms	Tokenizer

	The number of words that occur only once (Hapax legomena)	Tokenizer
	The number of words that occur twice (Hapax dislegomena)	Tokenizer
	List of foreign words	Tokenizer, StanfordCoreNLP POS tagger
	Average word length	Tokenizer, [Sentence splitter]
	The number of capitalized words	Tokenizer
	The number of Words with repetitive letters	Tokenizer
	The maximum length of a word	Tokenizer
	The number of Words with numbers	Word dictionary
	The ratio whose length greater than five words to total words	Tokenizer
	The ratio of words shorter than three letters to total words	Tokenizer
	The ratio of distinct words to the total number of words in the text.	Tokenizer
Syntactic Features	N-Grams of POS-Tags	Tokenizer, Sentence splitter, POS tagger, Freeling tool, Penn Treebank tagset, PAROLE tagset, TwitIE, Core NLP Standford POS tagger (English and Spanish) , Tree Tagger (Italian and Dutch)
	Syntactic n-grams	Tokenizer, Stanford Core NLP for the English dataset, Freeling for the Spanish dataset, and Alpino1 for the Dutch
	Frequency of Function words	Tokenizer, Sentence splitter, POS tagger,
	The Number of contraction words	Tokenizer
	Frequency of punctuations	NLTK Tokenizer
	Stop words	Tokenizer , Snowball stop word list
	The proportion ratio of singular to plural nouns and proper nouns and pronouns	Tokenizer, Sentence splitter, POS tagger,
	Ratios of comparative and adverbs and superlative adjectives	Tokenizer, Sentence splitter, POS tagger,
	The ratio of punctuations to text	Tokenizer
	Spelling and Grammatical errors	Tokenizer, Orthographic spell checker, Language Tool, standard US/GB English or Spanish dictionary
	The ratio of future and past verb tenses	Tokenizer, Sentence splitter, POS tagger,
	The Number of words with hyphen	Tokenizer
Structural Features	Total number of paragraphs	Tokenizer
	Total number of sentences	Tokenizer
	Number of sentences per paragraph	Tokenizer
	Number of words per paragraph	Tokenizer
	Number of characters per paragraph	Tokenizer
	Average sentence length in terms of characters	Tokenizer, [Sentence splitter], Stanford CoreNLP tool
	Average sentence length in terms of words	Tokenizer, [Sentence splitter]
	HTML tags	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries
	The number of Hashtags	Tweet Tokenizer
	The number of Retweets	Tweet Tokenizer
	The number of mentions of users using the pattern @username	Tweet Tokenizer
	Number of URL’s used	Tokenizer, Specialized dictionaries
	Set of common slang vocabulary	Tokenizer, Specialized dictionaries
	Number of emoticons	Tokenizer, Specialized dictionaries
Content-specific Features	Frequency of content specific keywords	Tokenizer, Specialized dictionaries
	Topic specific features	Tokenizer, MALLET, LDA(Latent Dirichlet Allocation)
	LIWC words	Tokenizer, Specialized dictionaries
	MRC features	Tokenizer, Specialized dictionaries
	Sentiment words	Tokenizer, SENTIWORDNET 3.0, VADER sentiment analysis library

Readability Features	Flesch Reading Ease	Tokenizer
	Flesch Kinkaid Grade Level	Tokenizer
	Gunning Fog Index	Tokenizer
	Coleman Liau Index	Tokenizer
	LIX	Tokenizer
	RIX Readability Index	Tokenizer
	SMOG index	Tokenizer
	Automated Readability Index(ARI)	Tokenizer
Information Retrieval Features	Cosine	Query Analyzer , Zettair
	Okapi BM25	Query Analyzer , Zettair

The most straightforward approach for the researchers is to represent the document text by the vectors of word frequencies (8, 10). The studies which were witnessed in Author Profiling were based on the features of word combinations for representing the style. This phenomenon is similar to the conventional Bag of Words (BOW) representation and then the researchers were concentrated on the topic based classification (Michał Meina et al., 2013, K Santosh et al., 2013). In general the text is defined as a set of words and every word possess some frequency without focusing on the contextual information. Style based text classification is observed to possess a significant difference where in the best features were found to be the most common words that is to discriminate between the authors. In this text classification on topics the common words such as pronouns, articles, prepositions were generally removed from the feature set because of the fact that they do not possess any semantic relationship and are termed as function words. A highly affective and successful method to describe a word feature set is extracting the functional words from the available corpus (by including candidate authors) for Author Profiling. In the subsequent step it is necessary to make a decision to finalize the frequent words which are significant to be used as features. Bag of Words (BOW) is one of the common approach to represent the documents. This representation builds the feature vectors of documents by taking every term in the vocabulary as an attribute.

The size of the feature terms also place a predominant role in the document representation. Few researchers (4, 16) used 3000 frequent terms as features and incremented up to 50000 frequent terms. Seifeddine Mechti computed [17] ranked list of words that occur in corpus by using top 200 attributes. Upendra Sapkota used [23] the top 5000 words for the prediction of the gender but the results were poor with respect to the language English but achieved good results with respect to the language Spanish. The researchers also deduced that not only the features set size but the classification algorithms also has a significant role when the dimensionality of the problem increases which over fits the training data.

In order to take a benefit of the contextual information, n-grams of words (n adjacent words) were proposed as textual feature in [6, 9, 37, 44, 46, 49, 50]. Prior analysis says individual word features based classification accuracy dominated when word n-grams were considered. An acronym is an abbreviation, used as a word, which is formed from the initial components in a phrase or a word. The numbers of acronym words were used [29] by Juan Soler Company as a feature set.

Foreign words are those words which are mostly slangs used in internet like “Helloooo”, “Whaaaat”, “yipeee”, “ROFL” etc. Some researchers used [2, 5, 20] these foreign words for his research. Average word length as a feature is been used by [7, 20, 26] and the number of capitalized words as a feature is adopted in [3, 10, 15, 40, 42, 46, 50]. The number of Words with repetitive characters, the number of Words with numbers, The ratio of five letter words to the total words and the ratio of three letter words to the total words as features used by [25] and the ratio of distinct words to the total words in the text as features adopted by [8].

2.2 Character features

A text is a sequence of characters and various character based features were defined by researchers to differentiate the text. Edson R. D. Weren, et al used [8, 10] the total number of characters and James Marquardt et al used [3, 10, 15] the number of capital letters and Gilad Gressel used [5] the frequency of special characters. Christopher Ian Baker extracted [7] the ratio of capital letters to total number of characters, the ratio of white-space characters to total number of characters, the ratio of tab spaces to total number of characters, the ratio of white spaces to nonwhite spaces, the ratio of capital letters to lower case letters, the ratio of numeric data is also used as features in the text.

A more complicated and computationally naive approach is to mine the n-gram frequencies on the character level. For the stylistic purposes, character n-grams were considered to be most significant. The method of extracting the most frequent n-gram requires no special tools and is language-independent. The dimensionality of this representation is significantly increases when it is compared with the word-based approach. This phenomenon occurs when the character n-grams captures redundant information when character n-grams were used to represent a lengthy word. Magdalena used [14] frequencies of the most common character 4-grams of the considered documents. Erwan Moreau considered [19] character unigrams, trigrams and 5-grams for text characterization. Julio Villena is used [1] n-gram based character sequences based on distance among histograms for each attribute, this procedure achieves good accuracy results for gender prediction (over 70%) but lower results for age prediction. Erwan Moreau kept [19] 12,000 distinct n-grams by discarding the least frequent ones. Octavia-Maria found [56] that the best tf-idf features were observed at character-level where n-gram ranges from 2 to 6 and after this threshold, the system seemed to overfit. The categories of character n-grams that were prominent across different languages were not the same. For English and Spanish languages the Italian familial tokens feature did not improve the accuracy, whereas for Dutch language the familial tokens were one of the key features (Suraj Maharjan & Thamar Solorio, 2015). Magdalena Jankowska used [14] n-grams in which tokens were utf8-encoded characters.

2.3 Syntactic features

A function word is a word which is significantly less meaningful content. These are considered as structured grammatical words in English which has a structural relationship with other words in a sentence. These function words includes the grammatical aspects of English such as pronouns (she, they), determiners (the, that), prepositions (in, of), auxiliary verbs (be, have), modals (may, could), conjunctions (and, but) and quantifiers (some, both). Based on the prior work some researchers [34] used function words as features and proved that the male authors are tend to use more prepositions in their writings when compared to female authors. Gilad Gressel extracted [5] around seven features from the text which includes the grammatical aspects such as adjectives, nouns, determiners, pronouns, adverbs and foreign words.

Wee-Yong Lim used [15] specific pronouns in their work. Seifeddine Mechti identified [17] prepositions, pronouns, determiners, adverbs, verbs from the documents and found that for the gender dimension prepositions, pronouns and verbs were highly effective. Braja Gopa computed [20] the frequencies of the pronouns. Aditya Pavan generated [21] the frequencies of prepositions of authors in each document for age and gender prediction.

The morpho syntactic information tags were assigned to every word token based on the contextual information. This is a process carried out by a Part Of Speech (POS) tagger. This POS tagger identifying the styles of the authors quite accurately by using POS tag n-gram frequencies or POS tag frequencies [9, 11, 19, 22] from the unrestricted text. POS tag information provides the structural analysis of sentences and never reveals the fact about the combination of words to form phrases or high level structures.

Contraction is not a grammatical feature. It is a shortened form of two words where in the apostrophe acts a join in place of missing letter or letters. Some famous contractions are how's (how is), can't (cannot), I'm (I am) and Ma'am (Madam). Some researchers [4, 10, 15, 16] used contractions as a feature to identify the age and gender. Stopwords are the words which refer to the most commonly used words in any natural language. These stopwords are to be filtered either before or after the preprocessing and some researchers used these stopwords as a feature in their work [6, 10, 16].

While identifying the demographic features of authors, the frequency of punctuations were used by [4, 6, 10, 15, 16, 20, 22, 36, 46, 48, 51]. The proportion of plural and singular nouns, pronouns and proper nouns, the ratio of past and future verb tenses, ratios of comparative and superlative adjectives and adverbs were used by [25]. The ratio of punctuations to text were used by [7], spelling and grammatical errors were used by [3, 10, 15, 18] and the number of words with hyphen were used by [4, 16].

2.4 Structural Features

The style of a writer is observed by not only the features of style mentioned earlier but the statistics states that the structural information related to the paragraph length, number of special characters, sentence length, words per sentence and the style of writing lengthy complex sentences are the features which contribute for identification of style. In a process one has to have an idea about the conversation length, the usage of hyperlinks, images and the style used either at the beginning or at the end of the conversation. In an investigation Michał Meina discovered [13] that the conversation length as feature is useful in spam detection. They also observed that chatter bots performance contains similar conversations where in they used the similarity measure called Jaccard similarity coefficient on individual conversations and they grouped all the average edit distances in to the analytic data set.

It is observed from the literature that generally the higher age people use longer words with greater frequency and females wrote longer sentences than males. While predicting the demographic features of authors, total number of sentences were used by [8, 10, 46], the length of sentences and words and their proportions to each other, the length of documents and the ratio of words to five letters words and above and words shorter than three letters compared to all words were used by [25]. Average sentence length in terms of words were used by [2, 15, 20, 44, 52], the number of HTML tags were used by [3, 8], the number of URL's were used by [10, 15, 22], the set of common slang vocabulary were used by [13, 16, 51] and the number of emoticons were used by [3, 6, 10, 16, 25, 36, 48, 51, 53]. The structural group of features aimed to trace the characteristics of the text that were interdependent with the use of the twitter platform. While identifying gender, age and personality traits in the tweets features such as the number of @mentions, hashtags and URLs were used by [37, 40, 42, 44, 46, 47, 48, 50, 53, 55] and the number of retweets used by [40, 44, 51].

2.5 Content specific features

For age and gender profiling for a given corpus [15], it is observed that the content based features alone are more discriminative than the rest of the features. A slight decrease in accuracy is observed when the content based features were added to other features. James Marquardt extracted [3] fourteen terms as features from the MRC psycholinguistic database and 68 terms as features from Linguistic Inquiry and Word Count (LIWC) dictionary and the features concerned with negative, positive or neutral sentiment based expressed sentences. MRC data features captures the information about the word frequencies that predict the concepts of psycholinguistic features such as imagery, concreteness and familiarity. Lesly Miculicich Werlen [37, 44, 48] categorized motion, anger and religion based frequency of words that are helpful while classifying the age and gender in hotel reviews. Every individual is having his own style of writing. The style of writing never changes. The writing style, word choice and grammar rule is solely depend on the topic of interest and the differences were found with topic variations. It is observed that the gender specific topic will have an impact in their writing styles. It is observed that the female tend to write more about wedding styles and fashions and whereas male bloggers stress more on technology and politics. This phenomena when applied with reference to the age the people of 20’s write more about their college life and the people of 30’s write more about marriage, job and politics and more so the teenagers are tend to write about their friends and mood swings. With the above statistics it is evident that the content based features place a dominant role while distinguishing between the bloggers of different age groups. James Marquardt extracted some [20] features concerning about the positive, negative or neutral sentiments expression terminology using tools. These tools calculate a sentiment value for every word, with zero for neutral sentiment, negative value for negative sentiment and positive values corresponding to positive sentiment. Adam Poulston recognized [54] that the topic models are useful in developing Author Profiling systems across the number of languages and provide reasonable results without any additional features. Aditya Pavan et al. [21] considered the topic distribution model and used Latent Dirichlet Allocation (LDA) in order to get the topics in the documents using the probabilistic distribution function. LDA is a Bayesian hierarchical model which is modeled as an item of collection with a finite mixture of topics. In the process experimentation LDA is modeled on the set of topic probabilities while extracting the features.

2.6 Readability Features

Readability features are measured to specify and finding the complicacy in understanding a passage in English. There are many tests available for finding the complicacy in the written text and out of which some tests like Gunning Fog Index, SMOG index, Flesch Reading Ease, Coleman Liau Index, Flesch Kinkaid Grade Level, LIX, Automated Readability Index (ARI) and RIX Readability Index are familiar to categorize the texts. Several researchers ([3], [5], [8], [18], [24]) were used readability features along with other features to predict gender and age dimensions of authors but it is observed that the impact of these features is more on the [18] accuracies.

2.7 Information Retrieval Features

Information Retrieval is a process of finding relevant documents based on the user input such as key words or example documents. The Information Retrieval (IR) System indexes the complete set of words in a document. Edson R. D. Weren employed [8] around 30 IR-based features. They used the text to classify a query and retrieved similar ‘k’ texts. The ranking is a process which is evaluated by the cosine or okapi metrics. Cosine features computation is based on the aggregation function on top-k results for which gender/age group results in response to a query composed by the keywords in the text that were to be classified. For this feature set, queries and documents were compared using the cosine similarity and Okapi BM25 score. Edson R. D. Weren applied [24] same set of features to different corpus and observed that 51% documents were correctly classified for gender dimension and 55% documents were correctly classified for age dimension.

2.8 Feature selection and Extraction

The selection of features was carried out by Fermín L. Cruz using [11] the chi-square correlation measure between the feature and the output classes. The Jaccard similarity coefficient was applied [13] over the texts to focus on the bag of words that are informative words rather than the frequent ones. Wee-Yong Lim applied [15] Principal Component Analysis (PCA) as a method which transform the high dimensional data into a lower dimensional space linearly for simple representation of the data. A. Pastor Lopez-Monroy followed [16] some ideas from Concise Semantic Analysis (CSA) to use a low dimensional representation with high level of representativeness. Erwan Moreau used [19] various classical distance measures like Euclidean, Cosine to identify the selected n-grams in frequent n-grams. Lucie Flekovayz used [25] an approach called Information Gain which ranks and prune the feature space by using the top 1500 features.

To represent documents Miguel A. Álvarez-Carmona brought [36] ideas from the information retrieval field by exploiting the Latent Semantic Analysis (LSA). LSA represents terms and documents into a new semantic space. This is carried out by performing a Singular Value Decomposition (SVD) using a Term Frequency Inverse Document Frequency (TFIDF) matrix. Edson Roberto

Duarte Weren selected [39] a subset of 6 to 22 features from 288 features by using BestFirst subset evaluator. It is proved that usage of subsets had advantages in nearly all predictions except age prediction. Lesly Miculicich Werlen used [48] four feature selection methods that were evaluated to determine the suitability of the data such as Manual selection, Information Gain, Odd ratio and Support Vector Machine Recursive Feature Elimination (SVM RFE). Scott Nowson compressed [50] the index of features by using truncated singular value decomposition. Ifrah Pervaz used [52] WEKA's attribute selection approach for selecting best features from the complete feature set by exploring four evaluators and two search methods. They covered the role of stylistic features in identification of author personality traits. For this purpose they figured out 29 features and performed different experiments on these features by comparing accuracies using all features, then checking the accuracy for single feature and finally using the subsets of the features and found that best results achieved by using feature subsets.

By topic specific features [18] it is understood that the coefficients corresponds to the document representation is represented with 150 (for each language) linear statistical topics that are estimated using Latent Semantic Analysis (LSA) technique. The methodology proposed [10] for the Spanish corpus focuses on the use of graphs as a strategy for feature extraction. They used a graph-based representation for extracting n-grams of words with the SUBDUE tool. Aditya Pavan used [21, 43] a generative model called Latent Dirichlet Allocation (LDA) while extracting a set of topics. Suraj Maharjan used [46, 54] the gensim Python library for LDA topic extraction. Caitlin McCollister used [47] the topic modeling software package MALLET to construct models of 100 topics each for the four languages. Juan-Pablo used [53] a syntactic parser to extracting syntactic n-grams from dependency trees.

3. Authorship Profiling Methods

In every Authorship Profiling problem, the training corpus consists of a set of text samples and a set of candidate authors of known authorship. The test corpus contains the set of text samples of unknown authors which has to be profiled individually to each candidate author. Many Author Profiling approaches were observed in most of the cases the training text was treated cumulatively (per author) or individually. In some research instances it is observed that the available training text per author was concentrated in a big file in order to extract a cumulative representation of author's style. In such cases the researchers disregarded the differences between the texts written by the same author. This type of approach is called profile based approach.

In other approaches multiple training text samples were required per author to develop an accurate profiling model. That is, every training text is represented individually as a separate instance of authorial style, this type of the approach is called instance-based approach. In this paper the difference between instance based and profile-based approaches were considered based on the property of the profiling methods since the review determines the philosophy of every method. It emphasizes the writing style of each method that attempts to handle whether the existing style is general for each author or it is a separate style for each document.

3.1 Profile-based Approaches

In this profile based approach the available training texts per author was concatenated to get a single text file. This single big file is used to extract the different properties of the author's style. The text from the testing corpus is compared with every author file and measure was applied and the most suitable author was estimated. In this case only a big concatenated file per author is being considered and individual text samples representation is ignored. This results in eliminating the differences between the text files of the same author. And also it is observed that the extracted stylometric features of a concatenated file are different from each of the training text features. The profile-based approach uses the training process and in this training phase consists of the retrieval process of profiles for all the candidate authors. Alberto Bartoli found [38] that the tweets of the problem instances in each of the subsets were authored by the same person. For this reason it is decided to build a new training set by replacing each of those subsets with a single problem instance in which the corpus is the union of all the tweet sets of the subset. Maite Gimenez considered [40] an approach that joins all tweets for each user, thereby making sample for each user. In some approaches authors were used [42, 56] the tweets as a text source for classification and tweets were joined in order to create larger documents of text.

3.2 Instance-based Approaches

In the modern Authorship Profiling approaches, the known authorship text sample is considered as an instance and every text sample as a unit. Here in this process every text sample from the training corpus is represented as a vector of attributes and a chosen classification algorithm is trained based on the instance set of known authorship (training set) in order to develop a profiling model. This model is able to predict the demographic features of an unknown text. It is understood that these classification algorithms requires more training instances per class for extracting a reliable model. The text samples were long enough which represents their style adequately. The lengths of the text samples were varied in size and presented in literature. It is observed that when the text block length decreases then the accuracy reduced. The selection process of the training text sample instance is not a trivial process and observed that it affects the performance of the profiling model. In the process it is described that the vector space models comprises with the most of the instance-based approaches which are followed by the similarity-based models.

3.2.1 Vector Space Models

Each document is a vector in vector space model and an entry is the frequency of a given feature in the document. Different researchers were used a variety of powerful machine learning and statistical algorithms to build a classification model by using these feature vectors. While identifying the demographic features of authors, Support Vector Machines (SVM) Classifiers were used by [3, 9, 15, 23, 32, 38, 40, 41, 42, 44, 53, 54, 56], decision trees through the J48 algorithm is used by [17, 20, 24, 26, 43], Expectation Maximization Clustering (EMC) algorithm is used by [4], random forest classifier is used by [5, 10, 18, 30, 38, 51, 55], logistic regression classifier is used by [6, 12, 25], the algorithm was developed using the Perl programming language with functions was used by [7], Liblinear classifier was used by [16, 36], rule based algorithm-JRip was used by [11], Common N-Gram (CNG) classifier was used by [14], Maxent Classifier was used by [21], custom stochastic gradient descent algorithm was used by [27, 37], Sequential Minimal Optimization (SMO) algorithm was used by [28, 31], REPTree (a fast decision tree learning algorithm) was used by [29], IBK(IB1) classifier was used by [30], Exponential Gradient (EG) algorithm was used by [32], Multi-Class Real Winnow learning (MCRW) algorithm was used by [34].

It is evident that some of these algorithms were in a position to handle sparse, high dimensional and noisy data, by giving an expressive representations of text. Support Vector Machines (SVM) model generally avoids overfitting problems and suggested as a best solution. The vector space models which are having class imbalance problem because of which the effectiveness is observed to be diminished. In a detail approach in these models the training set is initially rebalanced by segmenting a particular author’s text samples according to the size of each class. In this way longer text samples formed for majority authors and shorter text samples were produced for minority authors.

3.2.2 Similarity-based Models

Similarity based models emphasizes on calculating the pair wise similarity measures between the known and unseen text samples and there by estimate the most likely author by using nearest neighbor algorithm. To predict traits (gender, age and personality) of Twitter users Piotr Przybyła applied [55] this procedure. The idea is to start with exploring close similarities between writings, and then tries to discover more complex dependencies. More specifically, in order to predict traits for a new user, one must find the most similar user in the training data. When the similarity is sufficiently close, one can assign traits of the found user to the new user. Otherwise an advanced classification model is used to predict the traits. In this case it is to found the nearest neighbour of the new user in the training data in order to determine the nearest neighbour, and then Euclidean distance is used for all available features. Whenever the distance is less than a certain threshold then the traits were assigned to the nearest neighbour to the new user. Mirco Kocher used [45] three nearest neighbors according to a simple distance metric called SPATIUM-L1 based on the L1 norm.

4. Evaluation

The Authorship Profiling techniques were significantly applied to several models and presented in the literature to predict demographic features of the authors including English literature [1-26, 28, 29, 32, 33, 34, 35, 36-56], Greek literature [31], Vietnamese literature [30], Spanish literature [1-27, 36-56], Dutch literature [36-43, 45-56], Italian literature [36-43, 45-52, 54-56] etc.

Other than literature, different evaluation corpuses were used by several researchers for Authorship Profiling studies to cover various text domains such as essays written by psychology undergraduates [33], e-mail messages [28], Vietnamese blogs [30], the Greek blogosphere [31], blogs [1-27], social media posts [1-9], Reviews [1-9], Twitter tweets [1-9, 36-56] etc. In this analysis some of the corpuses primarily collected for text analysis tasks but these corpuses were also been used for Author Profiling studies. Here are the few corpuses that are used for this process including parts of British National Corpus (BNC) (Moshe Koppel et al., 2002), International Corpus of Learner English (Moshe Koppel et al., 2005, Shlomo Argamon et al., 2009), NY Times Opinion Blog corpus (Juan Soler Company & Leo Wanner, 2014).

The Authorship Profiling method requires assessment of demographic features of an author and its performance under different conditions. The vital assessment parameters which are to be considered for Authorship Profiling are training and testing corpuses in terms of the number of candidate authors, the document length and size and the division of training corpus over the authors based on the fact that whether the corpus is balanced or imbalanced. Identification of the feature vectors used to build the classification model and finding of suitable classification algorithms that also affect the effectiveness of Author Profiling techniques. Identification of features common to multiple languages, recognition of appropriate features for predicting various demographic features, run time of a system, the performance evaluation measures were used by the researchers which shows impact on evaluation of Author Profiling techniques.

Several researchers experiment their Author Profiling techniques for multiple natural languages. Some researchers [1-26, 44] assessed their technique for two languages namely Spanish and English. And another group of researchers [36-43, 45-52, 54-56] evaluated their method for four languages namely English, Spanish, Dutch and Italian. And others [53] evaluated their technique for three languages, namely English, Spanish and Dutch corpora.

Different authors concentrated on prediction of different types of demographic features. Some group of researchers [1-28, 30, 33, 34, 36-56] predicted gender and age. Another group of researchers [36-56] predicted big five personality traits namely openness, extraversion, agreeableness, conscientiousness, stableness in tweets. Other researchers [28, 33, 35] predicted nativity language of authors and [28, 30] predicted location and [30] predicted occupation and [29, 31, 32] predicted gender only.

Most of the researchers used the same set of features for predicting demographic features of the authors in various languages. It is necessary to take care while considering the similar set of features. Such care is necessary while considering the set of capital letters. In this case whether the capital letter is a first letter of a sentence or a first letter of a word is the concern. But such features were not suitable for indic languages. Upendra Sapkota used [23] the top five thousand frequent words for the prediction of the gender but the results were poor with respect to the language English but achieved good results with respect to the language Spanish. The familial tokens were one of the key feature for language Dutch but these features were not effective for the languages such as English, Spanish and Italian [46].

In Authorship Profiling studies, Run time plays a major role to evaluate the efficiency of the technique used to predict the demographic features of the authors. In 2013 PAN competition the researcher [26] used less run time of 10.26 minutes with an approach of only readability features and obtained the 8th position in English and 13th in Spanish. Seifeddine Mechti took [17] 11.78 days as runtime and approached the task with content features, obtained the 3rd. position in English and 21th in Spanish. The vast majority of approaches took a few hours of run time. Suraj Maharjan [6] has used the technique MapReduce on a Hadoop cluster when the data is amounting to huge number. Using this technique they completed the training phase in a short time where in the filename with the class information is considered as a key and the file content is as a value so as to generate the necessary n-gram. The total classification process was completed in 26 minutes and it is observed as 69 hours for other researchers.

Different evaluation metrics namely accuracy, precision, recall, F-measure were used by the researchers to check the performance of their models. Accuracy is considered as a scoring metric to evaluate the effectiveness of the system. Accuracy in this context is the ratio of number of test documents that were correctly predicted to the total number of test documents. Recall of a classifier is the proportion of positive demographic features that are correctly predicted. Precision of a classifier is the ratio of the number of correct predictions to the total number of correctly and incorrectly predictions. The F-measure is the harmonic mean of recall and precision.

5. Discussion

A. Pastor Lopez-Monroy achieved [4] best accuracies of 0.6795 and 0.3974 for the gender and age predictions in English blogs and 0.5893 and 0.4821 for the gender and age predictions Spanish Blogs of PAN 2014 [58]. Shrestha achieved [6] a best joint accuracy of 0.2062 for English and 0.2845 for Spanish Social Media corpus of PAN 2014 competition [58]. A. Pastor Lopez-Monroy achieved [1] best accuracies of 0.7208 and 0.4935 for the gender and age predictions respectively in English and Shrestha achieved [6] best accuracies of 0.6556 and 0.6111 for the gender and age predictions respectively in Spanish Twitter corpus of PAN 2014 competition [58-75]. A. Pastor Lopez-Monroy achieved [4] best accuracies of 0.6809 and 0.3337 for the gender and age predictions respectively in English Reviews of PAN 2014 competition [58].

Michał Meina achieved [18] a best accuracy of 0.5921 for the gender prediction in PAN 2013competition [57], along with they achieved best total accuracy of 0.3894 in English blogs data. A. Pastor Lopez-Monroy achieved [16] a best accuracy of 0.6572 for the age prediction in English blogs data PAN 2013 competition [57]. K Santosh achieved [22] a best accuracy of 0.6473 for the gender prediction in PAN 2013 competition [57], in addition to it they achieved best total accuracy of 0.4208 in Spanish blogs data. A. Pastor Lopez-Monroy achieved [16] a best accuracy of 0.6558 for the age prediction in Spanish blogs data PAN 2013 competition [57].

Miguel A. Álvarez-Carmona achieved [36] a global accuracy of 0.7906 for English, 0.8215 for Spanish and 0.9406 for Dutch languages and Carlos E. González-Gallardo achieved [41] a global accuracy of 0.8658 for Italian in PAN 2015 competition [59].

Maria De-Arteaga found [12] good results while distinguishing gender by using all features in English and Spanish and observed that for age range and found the supervised attributes are the better predictors. They were identified the best statistical features, that were found based on bayes theorem while predicting the age and the gender. They were also observed that the stylistic and lexical features were more discriminative features to distinguish the age than the gender and also found that the pre-established lists of words are useful while estimating the age but not useful for distinguishing the gender.

Dominique Estival used [28] a corpus of 9836 e-mails of 1033 authors which contains English, Spanish, Arabic language e-mails. While predicting native language of authors they recognized 689 features of character-level, lexical, and structural features. They used many machine learning algorithms for this analysis but Random forest algorithm gave a overall accuracy of 0.8422. While predicting the education dimension the machine learning algorithm named bagging gave an accuracy of 0.7998 by using the function words as features. The SMO machine learning algorithm gave an accuracy of 0.8113 for country dimension by using all features. While predicting the gender Juan Soler Company achieved [29] a good accuracy of 0.8283 by using the sentence based, character based, syntactic and word based features.

Dang Duc Pham used [30] the corpus of 3524 Vietnamese Weblog pages of 73 bloggers. They exploited 298 features including Lexicon, Character-based, Content-Specific, Document-based, Paragraph-based, Word-based, Structural, Line-based, POS-based, Function words features and applied them on following ten machine learning algorithms namely Neuron Network (Multilayer Perceptron), IBk (IB1), ZeroR, Bagging, Decssion Tree J4.8, SMO, NaiveBayes, BayesNetwork, Random Forest and RandomTree. Out of these algorithms IBK gave the best accuracy of 0.8212 for occupation and 0.7800 for location dimension.

Shlomo Argamon used [33] the corpus of International Corpus of Learner English (ICLE) which is a culmination of non-native English speakers from various countries who were learning English as a second language and the corpus was tested to predict the age, gender and native language. He also used essays of 251 psychology undergraduates at the University of Texas at Austin for neurotism prediction. They considered five sub-corpora namely Russian, Czech Republic, Bulgaria, French and Spanish from ICLE. They used 258 authors writings from each sub corpus to avoid class imbalance problems. While predicting the age, gender, nativity language and neurotism they observed that style based features gave an accuracy of 65.1%, content based features gave an accuracy of 0.823 and both style based and content based features together gave an accuracy of 0.793. They concluded that the content based features gave best results than the combination.

Moshe Koppel achieved [35] an accuracy of 0.802 while predicting the nativity language by using 1035 features including 250 rare POS bigrams, 400 standard function words, 185 error types and 200 letter n-grams. Juan-Pablo Posadas-Duran showed [53] that the use of syntactic n-grams along with other specific tweet features (such as number of retweets, frequency of hashtags, frequency of emoticons, and usage of referencing urls) gave good results while predicting the personal traits but their usage is not successful while predicting the age and gender.

References

[1] Julio Villena-Román, José Carlos González-Cristóbal, "DAEDALUS at PAN 2014: Guessing Tweet Author's Gender and Age ", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[2] Seifeddine Mechti, Maher Jaoua, Lamia Hadrich Belguith, "Machine learning for classifying authors of anonymous tweets, blogs and reviews", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[3] James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, Martine De Cock, "Age and Gender Identication in Social Media", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[4] A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, and Luis Villaseñor-Pineda "Using Intra-Profile Information for Author Profiling", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[5] Gilad Gressel, Hrudya P, Surendran K, Thara S, Aravind A, Prabaharan Poornachandran, "Ensemble Learning Approach for Author Profiling", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[6] Suraj Maharjan, Prasha Shrestha, and Thamar Solorio, "A Simple Approach to Author Profiling in MapReduce", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[7] Christopher Ian Baker, "Proof of Concept Framework for Prediction", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[8] Edson R. D. Weren, Viviane P. Moreira, and José P. M. de Oliveira, "Exploring Information Retrieval features for Author Profiling", Proceedings of CLEF 2014 Evaluation Labs, 2014.

[9] Satya Sri Yatam, T. Raghunadha Reddy, "Author Profiling: Predicting Gender and Age from Blogs, Reviews & Social media", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 3 Issue 12, December-2014.

[10] Yuridiana Aleman, Nahun Loya, Darnes Vilarino, David Pinto, "Two methodologies applied to the Author Profiling task", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[11] Fermín L. Cruz, Rafa Haro R, and F. Javier Ortega, "ITALICA at PAN 2013: An Ensemble Learning Approach to Author Profiling", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[12] Maria De-Arteaga, Sergio Jimenez, George Duenas, Sergio Mancera and Julia Baquero, "Author Pro_ling Using Corpus Statistics, Lexicons and Stylistic Features", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[13] Delia-Irazú Hernández, Rafael Guzmán-Cabrera, Antonio Reyes, and Martha-Alicia Rocha, "Semantic-based Features for Author Profiling Identification: First insights", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[14] Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios, "CNG text classification for Authorship Profiling task", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[15] Wee-Yong Lim, Jonathan Goh and Vrizlynn L. L. Thing, "Content-centric age and gender profiling", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[16] A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Esaú Villatoro-Tello, "INAOE's participation at PAN'13: Author Profiling task", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[17] SeifeddineMechti, Maher Jaoua,Lamia Hadrich Belguith, "Author Profiling Using Style-based Features", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[18] Michał Meina, Karolina Brodzi´nska, Bartosz Celmer, Maja Czoków, Martyna Patera, Jakub Pezacki, and Mateusz Wilk, "Ensemble-based classification for Author Profiling using various features", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[19] Erwan Moreau and Carl Vogel, "Style-based distance features for Author Profiling", Proceedings of CLEF 2013 Evaluation Labs, 2013.

[20] Braja Gopal Patra, Somnath Banerjee, Dipankar Das, Tanik Saikh, Sivaji Bandyopadhyay, "Automatic Author Profiling

Based on Linguistic and Stylistic Features”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[21] Aditya Pavan, Aditya Mogadala, Vasudeva Varma, “Author Profiling using LDA and Maximum Entropy”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[22] K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma, “Author Profiling: Predicting Age and Gender from Blogs”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[23] Upendra Sapkota, Thamar Solorio, Manuel Montes-y-Gómez, and Gabriela Ramírez-de-la-Rosa, “Author Profiling for English and Spanish Text”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[24] Edson R. D. Weren, Viviane P. Moreira, and José P. M. de Oliveira, “Using Simple Content Features for the Author Profiling Task”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[25] Lucie Flekova and Iryna Gurevych, “CanWe Hide in theWeb? Large Scale Simultaneous Age and Gender Author Profiling in Social Media”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[26] Lee Gillam , “Readability for Author Profiling?”, Proceedings of CLEF 2013 Evaluation Labs, 2013.

[27] Jonathan Schler, Moshe Koppel, Shlomo Argamon , James Pennebaker, “Effects of Age and Gender on Blogging”.

[28] Estival D., Gaustad T., Pham S. B., Radford W., and Hutchinson B. “Author Profiling for English Emails”. 10th Conference of the Pacific Association for Computational Linguistics (PACLING, 2007), 2007.

[29] Juan Soler Company1, Leo Wanner. “How to Use Less Features and Reach Better Performance in Author Gender Identification”. The 9th edition of the Language Resources and Evaluation Conference(LREC), 26-31 May, (2007).

[30] Dang Duc, P., Giang Binh, T., Son Bao, P.: Author Profiling for vietnamese blogs. Asian Language Processing, 2009 (IALP '09), pp. 190-194. (2009).

[31] Dang Duc, P., Giang Binh, T., Son Bao, P.: Authorship Attribution and Gender Identification in Greek Blogs. 8th International Conference on Quantitative Linguistics (QUALICO), April 26-29, 2012 , (2012).

[32] Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), pp. 401-412 (2002).

[33] Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2):119.(2009).

[34] Schler, J., Koppel, M., Argamon, S., and Penebaker, J. “Effects of Age and Gender on Blogging”. AAAI Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), AAAI Technical report SS-06-03, (2006).

[35] Koppel. M., Schler. J., Kfir Zigdon, “Determining an Author's Native Language by Mining a Text for Errors”. eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. August 21-24, (2005).

[36] Miguel A. Álvarez-Carmona, A. Pastor López-Monroy,Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, and Hugo Jair Escalante, “INAOE’s participation at PAN’15: Author Profiling task”, Proceedings of CLEF 2014 Evaluation Labs, 2015.

[37] Mounica Arroju, Aftab Hassan, Golnoosh Farnadi, “Age, Gender and Personality Recognition using Tweets in a Multilingual Setting”, Proceedings of CLEF 2014 Evaluation Labs, 2015.

[38] Alberto Bartoli, Andrea De Lorenzo, Alessandra Laderchi, Eric Medvet, and Fabiano Tarlao, “An Author Profiling Approach Based on Language-dependent Content and Stylometric Features.”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[39] Edson Roberto Duarte Weren, “Information Retrieval Featuresfor Personality Traits”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[40] Maite Giménez, Delia Irazú Hernández, and Ferran Pla, “Segmenting Target Audiences: Automatic Author Profiling Using Tweets.”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[41] Carlos E. González-Gallardo, Azucena Montes, Gerardo Sierra, J. Antonio Núñez-Juárez, Adolfo Jonathan Salinas-López, Juan Ek, “Tweets Classification Using Corpus Dependent Tags, Character and POS N-grams”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[42] Andreas Grivas, Anastasia Krithara, and George Giannakopoulos, “Author Profiling using stylometric and structural feature groupings”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[43] Hafiz Rizwan Iqbal, Muhammad Adnan Ashraf, Rao Muhammad Adeel Nawab, “Predicting an author’s demographics from text using Topic Modeling approach”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[44] Yasen Kiprov1, Momchil Hardalov2, Preslav Nakov3, and Ivan Koychev4, “SU@PAN'2015: Experiments in Author Profiling”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[45] Mirco Kocher, “UniNE at CLEF 2015: Author Profiling”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[46] Suraj Maharjan and Thamar Solorio, “UsingWide Range of Features for Author Profiling”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[47] Caitlin McCollister1, Shu Huang2, and Bo Luo1, “Building Topic Models to Predict Author Attributes from Twitter Messages”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[48] Lesly Miculicich Werlen, “Statistical Learning Methods for Profiling Analysis”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[49] Fahad Najib, Waqas Arshad Cheema, Rao Muhammad Adeel Nawab, “Author’s Traits Prediction on Twitter Data using Content Based Approach”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[50] Scott Nowson, Julien Perez, Caroline Brun, Shachar Mirkin, and Claude Roux, “XRCE Personal Language Analytics Engine for Multilingual Author Profiling”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[51] Alonso Palomino-Garibay1, Adolfo T. Camacho-Gonzalez1, Ricardo A. Fierro-Villaneda2, Irazu Hernandez-Farias3, Davide Buscaldi4, and Ivan V. Meza-Ruiz2, “A Random Forest Approach for Authorship Profiling”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[52] Ifrah Pervaz, Iqra Ameer, Abdul Sittar, Rao Muhammad Adeel Nawab, “ Identification of Author Personality Traits using Stylistic Features”, Proceedings of CLEF 2015 Evaluation Labs, 2015.

[53] Juan-Pablo Posadas-Durán, Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, Alexander Gelbukh, and Obdulia Pichardo-Lagunas, “Syntactic N-

grams as Features for the Author Profiling Task", Proceedings of CLEF 2015 Evaluation Labs, 2015.

[54] Adam Poulston, Mark Stevenson, and Kalina Bontcheva, "Topic Models and n–gram Language Models for Author Profiling", Proceedings of CLEF 2015 Evaluation Labs, 2015.

[55] B. Naresh Kumar Reddy, N. Venktram and Sireesha, "An Efficient Data Transmission by using Modern USB Flash Drive," International Journal of Electrical and Computer Engineering, Vol. 4, Number 5, pp. 730-740, 2014.

[56] J. V. N. Ramesh, B. Naresh Kumar Reddy, V. V. Murali Krishna, B. M. Kumar Gandhi, V. Shiva and M. Dronika Devi, "An Effective Self-Test Scheduling For Real-time Processor Based System," International Journal of Smart Home, Vol. 9, Number 3, pp. 101-112, 2014.

[57] K.Lavanya, B. Naresh Kumar Reddy, C.Naga Raju and K.Sridhar, "Framework for Enhancing Level of Security to the ATM Customers with DCT based Palm Print Recognition," International Journal of Applied Engineering Research, Vol. 9,Number 19, pp. 5345-5351, 2014.

[58] G.L.N.Murthy, B.Anuradha, CH. Siva Rama Krishna, B. Naresh Kumar Reddy and J.V.N.Ramesh, "Effective utilization of labeling algorithms for Hippocampus segmentation," European Journal of Scientific Research, Vol. 134, Number 2, pp. 206-211, 2014. (SCOPUS indexed Journal).

[59] B.Venkateswara Reddy, P.Satish Kumar, P.Bhaskar Reddy and B. Naresh Kumar Reddy "Identifying Brain Tumour From MRI Image Usingmodified FCM and Support Vector Machine," International Journal of Electrical and Computer Engineering, Vol 4, Issue 1, pp. 244-262. 2013.

[60] B. Naresh Kumar Reddy, Vasantha.M.H. and Nithin Kumar Y.B., "A Gracefully Degrading and Energy-Efficient Fault Tolerant NoC Using Spare core," 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI 2016), Pennsylvania, U.S.A., pp. 146-151, 2016.

[61] Vijaya Sree Boddu, B. Naresh Kumar Reddy and M. Kranthi Kumar, "Low-Power and Area Efficient N-bit Parallel Processors on a Chip," 13th International IEEE India Conference INDICON 2016, pp. 1-4, 2016.

[62] G. L. N. Murthy, B. Anuradha, Siva Rama Krishna, B. Naresh Kumar Reddy and R. Sithara, "Slice specific atlas independent hippocampus segmentation using simple labeling,"10th International Conference on Intelligent Systems and Control (ISCO), pp. no. 1-5, 2016.

[63] B. Naresh Kumar Reddy, Vasantha.M.H., Nithin Kumar Y.B. and Dheeraj Sharma, "Communication Energy Constrained Spare Core on NoC," 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Dallas, U.S.A., pp. 1-4, 2015.

[64] B. Naresh Kumar Reddy, Vasantha.M.H., Nithin Kumar Y.B. and Dheeraj Sharma, "A Fine Grained Position for Modular Core on NoC," IEEE International Conference on Computer, Communication and Control, Sep 2015.

[65] B. Naresh Kumar Reddy, N. Suresh, J. V. N. Ramesh, T. Pavithra, Y. Krupa Bahulya, Pranose J Edavoor and S. Janaki Ram, "An Efficient Approach for Design and Testing of FPGA Programming using LabVIEW," 4th International Conference on Advances in Computing, Communication and Informatics, Aug 2015.

[66] B. Naresh Kumar Reddy, et al.,, "An Efficient Low Power High Performance in MPSOC," Proceedings of the Third International Symposium on Women in Computing and Informatics, pp. 708-711, India — August 10 - 13, 2015 .

[67] G S Ajay Kumar Reddy, S. V. Jagadesh Chandra and B. Naresh Kumar Reddy, "Developing the fabricated system of automatic vehicle identification using RFID based poultry traceability system," International Conference on Information Communication and Embedded Systems, pp. 1-6, 2014.

[68] B. Naresh Kumar Reddy, M. Naraimhulu, S. V. Sai Prasad, K. Khaja Babu and S. V. Jagadeesh Chandra, "An Efficient Online Mileage Indicator by Using Sensors for New Generation Automobiles," IEEE Bangalore Section technically co-sponsor on 2nd International Conference on Advanced Computing, Networking and Security, pp. no. 198-203, Dec, 2013.

[69] Narasimhulu. M, B. Naresh Kumar Reddy, Subrahmanya Sharma .G," Designing of a Smart Car Using ARM7", International Journal of Advances in Engineering & Technology, Vol. 3, Issue 2, pp. 185-191, 2012

[70] Piotr Przybyła and Paweł Teisseyre, "What do your look-alikes say about you? Exploiting strong and weak similarities for Author Profiling.", Proceedings of CLEF 2015 Evaluation Labs, 2015.

[71] Octavia-Maria S, ulea1;2 and Daniel Dichiu, Bitdefender Romania, "Automatic Profiling of Twitter Users Based on Their Tweets.", Proceedings of CLEF 2015 Evaluation Labs, 2015.

[72] Pan 2013 Author Profiling competition,http://www.uni-weimar.de/medien/webis/events/pan-13/pan13-web/author-profiling.html.

[73] Pan 2014 Author Profiling competition,http://www.uni-weimar.de/medien/webis/events/pan-14/pan14-web/author-profiling.html.

[74] Pan 2015 Author Profiling competition,http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-profiling.html^.

[75] Stamatatos. E. "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for Information Science and Technology, 60(3), pp. 538-556, 2009, Wiley.