N-gram approach for gender prediction

T Raghunadha Reddy
Research scholar, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India
E-mail: trnreddy543@gmail.com

B Vishnu Vardhan
Professor, Dept of CSE, JNTUH College of Engineering, Jagtiyal, Karimnagar, Telangana, India
E-mail: mailvishnu@yahoo.com

P Vijayapal Reddy
Professor, Dept of CSE
Matrusri Engineering college, Hyderabad, Telangana, India
E-mail: drpvijayapalreddy@gmail.com

***Abstract*—The Internet was growing with huge amount of information, through Blogs, Twitter tweets, Reviews, social media network and with other information content. Most of the text in the internet was unstructured and anonymous. Author Profiling is a text classification technique that is used to predict the profiling characteristics of the authors like gender, age, country, native language and educational background by analyzing their texts. Researchers proposed different types of features such as lexical, content based, structural and syntactic features to identify the writing styles of the authors. Most of the existing approaches in Author Profiling used the combination of features to represent a document vector for classification. In this paper, a new model was proposed in which document weights were calculated with combination of POS N-grams and most frequent terms. These document weights were used to represent the document vectors for classification. This experiment was carried out on the reviews domain to predict the gender of the authors and the achieved results were promising when compared with the existing approaches in Author Profiling.**

***Index Terms*— Author Profiling, Gender prediction, POS N-gram, Text classification, Term Frequency**

I. INTRODUCTION

The internet has become unmanageably big and day-by-day it is increasing exponentially through social media, blogs and reviews. Most of this information is written by various authors in different contexts. The availability of such anonymous information challenges the researchers and information analysts to develop automated tools for analyzing such information. In this regard, Author Profiling is a popular technique to extract such information from the texts by analyzing author's writing styles [1].

Authorship analysis is performed in three different ways such as Authorship Identification, Plagiarism Detection and Author Profiling [2]. Firstly, Authorship Identification is the process of finding the author of a given document. It is performed in two ways that are Authorship Attribution and Authorship Verification. Authorship Attribution predicts the author of a given anonymous document by analyzing various documents of multiple authors [6]. Authorship verification finds whether the given document is written by a particular author or not by analyzing the documents of a single author [3].

Secondly, Plagiarism Detection detects the author's contribution in the given document. It is performed in two steps. The first is source retrieval and the second is text alignment [13]. Source retrieval process retrieves the possible sources of suspicious documents that contain the content of a given document from document collections. Text alignment finds the matching percentage of a given document content from suspicious documents [13].

Author Profiling is an important technique in the present information era which has applications in marketing, security and forensic analysis [1]. 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. Forensics is a field to analyze the style of writing, signatures, documents, and anonymous letters. Author Profiling helps in crime investigation and forensic analysis to identify the perpetrator of a crime with the characteristics of writing styles. In the marketing domain the consumers were provided with a space to review the product. Most of the reviewers were not comfortable in revealing their personal identity. These reviews were analyzed to classify the consumers based on their age, gender, occupation, nativity language, 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 and also helps in estimating the suitable level of knowledge of each student or a student group in the educational forum.

The authorship attribution task has traditionally been carried on data from small set of documents written by various authors [5]. This task is more difficult to identify an author for larger data sets, involving more number of authors. In such cases, Author Profiling is a good alternative solution and provides clues to find the identity of authors. Author Profiling is also used to find the characteristics of the author even when the documents of the given author are not in the training data.

In general every human being has his own style of writing and it will not be changed while writing in Twitter tweets, blogs, reviews, social media and also in documents. Men use more number of determiners and quantifiers and woman use more number of pronouns than men in their writings [1]. Similarly the male authors stress more on topics related to sports, politics and technology whereas the female authors write about topics like beauty, kitty parties and shopping [2]. Prior works [2, 4] found that the male authors use more

978-1-5090-1560-3/17 $31.00 © 2017 IEEE
DOI 10.1109/IACC.2017.167

prepositions in their articles and blog posts when compared to female authors. Generally the writing styles of the authors vary based on the selection of topics and the writing styles like choice of words and grammar rules. In an observation [11], it is informed that females write more about wedding styles and males write more about technology and politics. Further females use more adjectives and adverbs than male authors.

This paper is organized in five sections. The contributions of various researchers for Author Profiling were discussed in section II. The proposed approach is explained in section III. Section IV addresses the corpus characteristics and various approaches to gender prediction. Section V concludes the work and future scope.

II. Related work

In Author Profiling, the main concentration of the researchers is to extract the features that are suitable for differentiating the writing styles of the authors. Various researchers proposed different types of features such as lexical features, structural features, content based features, syntactic features and semantic features [21]. Ensemble classifier is a combination of several weak classifiers, which works more effectively than the individual ones. As in [8], ensemble based learning approach implemented with two classifiers. One classifier used the document vectors with frequency of Bag Of Words and N-grams of POS tags. Another classifier used a set of 17 normalized stylistic features to represent the document. They observed that the use of complex set of POS tags increase the accuracy of gender and age prediction for Spanish language.

The features including part of speech N-grams, structural features, exploration of sequences of part of speech, test difficulty, dictionary based features, errors, topical features, topical centroids and structural centroids for English and Spanish language were used in [9]. It is observed that the exploration of sequences of part of speech N-gram features alone obtained a good result for gender and age prediction.

In general word N-grams comes under content based features and POS N-grams are considered as style based features. Several researchers used the combination content based features and style based features [12, 15, 18]. Another researcher [10] used 12000 most frequent distinct N-grams, which include character N-grams, word N-grams and POS N-grams were considered in their work. The structural features, readability measures and function words like POS unigrams play a vital role in prediction of gender and age of the authors [14]. The stylistic features like character N-grams and POS N-grams achieve good results for dense texts were [16].

Some researchers used combinations of syntactic, stylistic and semantic features [7]. It is observed that their system results a poor performance for gender prediction in English language with N-grams of POS tags. Five classes of features, including content based features, surface features, syntactic features like N-grams of POS tags, readability measures and semantic features were extracted for representing a document [8]. They observed that the POS unigrams for English language and POS trigrams and quadrigrams for Spanish language plays an important role to increase the prediction accuracy of gender and age. It also observed that the content based features perform better than the stylistic features.

As in [17], four types of features such as lexicon features, twitter specific features, orthographic features and term level features like word N-grams and POS N-grams were identified. In their observation word 3-grams and word 4-grams not improved the performance, but the POS tags, word unigrams and bigrams perform most sustainable performance on test datasets. It is also observed that the orthographic features were useful to improve the accuracy of gender and age, but these are not useful to improve the accuracy of personality traits.

The syntactic N-grams like POS tags, words, lemmas and relations were extracted to construct a dependency tree [19]. In their observation the syntactic N-grams along with specific tweet features achieved good results for predicting personality traits, but this combination was not successful for predicting the gender and age.

The N-gram language model, topic model and POS tags are used in their experiment [20]. It is observed that the POS N-grams were the best feature for gender and age prediction in English language, but these features are not included in their final experiment because they were not finding compatibility POS tagger for Spanish, Dutch and Italian languages and also observed that the topic based features alone were not performed well, but the combination of topic based features and N-grams shows good performance.

III. Proposed approach

In this approach, a new document representation technique is introduced to predict the gender of anonymous text in the reviews domain. In this model, $\{T_1, T_2, \ldots\ldots, T_n\}$ denotes the collection of vocabulary terms, $\{D_1, D_2, \ldots\ldots, D_m\}$is a collection of documents in the total corpus, TWM is the weight of the term in the total documents of male corpus, TWF is the weight of the term in the total documents of female corpus, DWM is the document weight specific to male profile, DWF is the document weight specific to female profile.

The step wise procedure of proposed approach is as follows:

Step 1: Collect the reviews corpus.
Step 2: Preprocessing techniques were applied on the corpus
Step 3: Extraction of most frequent terms.
Step 4: The term weight measure is used to calculate the term weights specific to each profile group.
Step 5: The Document weights are calculated to each profile group by aggregating term weights in the document.
Step 6: Generating document vectors by using weights of the documents.
Step 7: The classification model is created using document vectors.
Step 8: The test documents are passed through Step 2 to Step 6 to construct the document vectors. These document vectors are inputted to the classification model to identify the profiles of the each test document.

Fig 1 describes the model of the proposed approach.

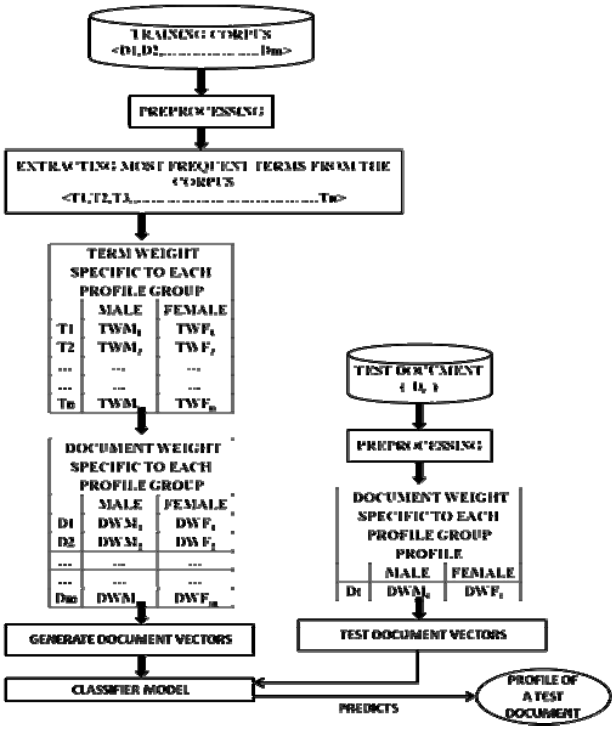

Fig 1. Proposed Model

In this approach, the accuracy for gender prediction depends on the efficiency of term weight measures and document weight measures used. Two profile groups were considered for gender profile such as male and female. The subsection A describes the proposed term weight measure. The proposed document weight measure is explained in subsection B.

A. Term weighting based on non uniform term distribution

In general, different authors were concentrated on various issues in the writing of reviews about products. Most of the terms in the reviews text were not uniformly distributed. In information theory, the information content value of a document is inversely proportional to the probability of occurrence of the feature in a document. An instance of a feature that occur in all the review documents has less information value when compared to those that occur in few documents. The proposed weight measure gives more weight to the terms that are not uniformly distributed across the documents. Term frequency in a specific document and the term frequency in all the documents in a profile group corpus are used to calculate the weight of the term in a specific profile group.

Let $P = \{p_1, p_2, \ldots\ldots, p_q\}$ is the set of profiles, $V = \{t_1, t_2, \ldots\ldots\ldots, t_n\}$ is a collection of vocabulary terms for analysis. Equation 1 is used to calculate the term weight specific to a profile. In equation 1, $tf(t_i, p_j)$ is the total frequency of term t_i in all the documents labeled with profile group p_j. $tf(t_i, d_k)$ is the term frequency of t_i in the document d_k .

$$W_{tij} = w(t_i, p_j) = \sum_{k=1}^{m} \left(\frac{tf(t_i, d_k)}{tf(t_i, p_j)} \log \left[\frac{1 + tf(t_i, d_k)}{1 + tf(t_i, p_j)} \right] \right) \quad (1)$$

B. Profile specific document weight measure

In this measure, each document weight calculated by aggregating the weighted representations of terms specific to that document. The proposed document weight used the weight of the term in the individual document and the weight of the term specific to the profile group. Equation 2 is used to calculate the weight of document specific to each profile group. In equation 2, Term Frequency Inverse Document Frequency (TFIDF) measure is used to calculate the weight of the term in a particular document. TFIDF measure gives the weight to a term based on the frequency of the term in a document and the number of documents contains the term in a total corpus of documents.

$$W_{dkj} = \sum_{t \in d_k, d_k \in p_j} TFIDF(t_i, d_k) * W_{tij} \quad (2)$$

$$TFIDF(t_i, d_k) = tf(t_i, d_k) * \log \left(\frac{|D|}{|1 + DF_{ti}|} \right) \quad (3)$$

where, W_{dkj} is the weight of document d_k in the profile p_j , $|D|$ is the total number of documents in the profile group, DF_{ti} is the number of documents contains the term t_i in the corpus of profile p_j .

The document vectors used for classification are represented using equation 4.

$$Z = \bigcup_{d_k \in p_j} (z_k, c_j) \quad (4)$$

here, $z_k = \{W_{dk1}, W_{dk2}, \ldots\ldots, W_{dkq}, Cj\}$ is a document vector for document d_k. W_{dkq} is the weight of a document d_k specific to profile group p_q, and c_j is a class label of profile p_j .

IV. EMPIRICAL EVALUATIONS

A. Corpus characteristics

The corpus was collected from TripAdvisor.com, and it contains 4000 reviews about different hotels. The corpus was constructed carefully to ensure its quality with regard to text cleanliness and annotation accuracy. In order to make this dataset applicable to Author Profiling and to ensure its quality, the following steps are adopted. First, reviews containing less than 5 lines of text were excluded from our dataset. Second, the reviews which are written in English language were only

considered. Third, it was considered the reviews written by the authors whose gender was given in their user profile. The corpus is balanced in terms of gender dimension for male and female profile groups contain 2000 reviews of each.

B. Evaluation measures

Most of the researchers used various measures such as Precision, Recall, F1-score and Accuracy proposed for finding their system performance. In this work, Accuracy is used to measure the performance of the gender prediction model. Accuracy is the ratio of number of documents profiles correctly predicted to the total number of documents in the corpus.

$$Accuracy = \frac{Number\ of\ \text{documents correctly predicted their}\ gender}{Total\ number\ of\ documents}$$

C. Bag Of Words (BOW) representation using Part Of Speech (POS) N-grams

In the English language, words are the basic units to write the sentences. Based on their functionality and usage the words are categorized into different types of part of speech that are nouns, pronouns, verbs, adverbs, adjectives, conjunctions, prepositions and interjections. In this approach, part of speech information is used in order to represent a document.

In the experiment, 10-fold cross validation is used to evaluate the document vectors. In 10-fold cross validation, the original corpus is randomly partitioned into 10 subsamples. Of the 10 subsamples, 9 subsamples are used as training data for generating the classification model, and the remaining one subsample is used to validate the model. The cross validation process is repeated until every subsample is used exactly once as the validation data. In this work, the performance of the gender prediction was tested on various classifiers such as Naive Bayes Multinomial (Probabilistic), Simple Logistic and Logistic (functional), IBK (lazy), Bagging (Ensemble/meta) and RandomForest (Decision Tree).

TABLE I. The accuracy percentages of BOW approach with most frequent POS N-grams

Number of Features / Classifier	200	400	600	800	1000
Naive Bayes Multinomial	64.11	64.32	64.85	64.48	64.48
Simple Logistic	61.05	59.67	59.83	59.14	58.61
Logistic	61.79	60.10	56.87	52.90	55.60
IBK	54.33	53.49	52.49	52.12	51.96
Bagging	59.94	60.09	61.13	60.29	60.01
Random Forest	62.58	62.79	61.63	60.62	60.89

In the Bag Of Words representation the documents are represented as vectors, each value in a vector is a frequency of the feature in a document. In this experiment, the frequencies of POS N-grams (n=1,2,3) are used as features to represent the document. The Stanford part of speech tagger is used for part of speech tags to the terms in the document. The features were considered from 200 to 1000 most frequent POS N-grams with an increment of 200 in each iteration. Table 1 shows the accuracies of BOW approach for different classifiers. The Naive Bayes Multinomial classifier attained a good accuracy for 600 most frequent POS N-grams compared to other classifiers. The Naive Bayes Multinomial classifier obtained the same result for most frequent 800 and 1000 POS N-grams. In all the classifiers, the accuracy is decreased when the number of features was increased. Only frequencies of POS N-grams are not sufficient to increase the accuracy of gender prediction.

D. Weighted representation of Part Of Speech (POS) N-grams

Part of speech information is used to compute a term weight. The term weight measure is used for different purpose, which is computed from part of speech statistics not from term frequency statistics. The POS information is used in the form of N-grams (POS N-grams), which are contiguous POS sequences. The proposed model as described above is used to generating the document vectors. Most frequent POS unigrams, POS bigrams, POS trigrams are used as features.

The proposed term weight measure is used to calculate the POS N-gram weight based on the frequency of the POS N-gram in a specific document and the frequency of POS N-gram in all the documents of a specific profile group. The proposed document weight measure compute the weight of document by aggregating the weights of POS N-grams that are occurred in a document.

TABLE II. The accuracy percentages of proposed approach when POS N-grams were used as features

Number of Features / Classifier	200	400	600	800	1000
Naive Bayes Multinomial	62.2	66.95	67.6	66.9	68.15
Simple Logistic	48.4	48.35	48.3	48.4	48.3
Logistic	62.15	66.7	68.6	68.2	70.9
IBK	50.45	50.4	53.65	53.85	55.8
Bagging	52.4	50.45	52.1	53.05	53.5
Random Forest	50.8	49.8	52.55	53.45	54.4

These document vectors were given to different classifiers for generating classification model. Table II shows the accuracies of gender prediction for various classifiers. The logistic regression classifier achieved a good accuracy for gender prediction among other classifiers. The results were not satisfied when only POS N-grams were used as features to calculate the weight of the document. In the next section, the POS N-grams frequencies were added to document weights that are calculated with the weights of the most frequent terms.

E. *document weights with most frequent terms and POS N-grams*

The frequencies of POS N-grams are combined with the document weights. The document weights are calculated by extracting 1000 most frequent terms from the corpus. The proposed term weight measure is used to calculate the weights of the terms specific to profiles. The document weight specific to profile group is calculated using document weight measure. These document weights were used to prepare document vectors.

Two preprocessing techniques such as stop words removal [22] and stemming [23] were applied on the corpus before extracting most frequent terms from the corpus. For extracting POS N-grams, one preprocessing technique such as removal of punctuation marks from the corpus was applied. The Stanford POS tagger is used to give tags to the terms in the document.

Table III shows the results of gender prediction when combination of features was used to generate the document vectors. The Naive Bayes Multinomial classifier obtained a good accuracy for gender prediction when 1000 most frequent terms are used for document weights and 800 POS N-grams combination was used compared to other classifiers. Except bagging classifier, in other classifiers the accuracy of gender prediction was decreased when the number of features was increased. The combination of features achieved good results for gender prediction compared to previous approaches.

TABLE III. The accuracy percentages of gender prediction using combination of POS N-grams and document weights

Number of Features / Classifier	1000 terms + 200 POS	1000 terms + 400 POS	1000 terms + 600 POS	1000 terms + 800 POS	1000 terms + 1000 POS
Naive Bayes Multinomial	86.68	86.47	86.63	86.91	86.73
Simple Logistic	61.05	59.67	59.83	59.14	58.62
Logistic	85.78	83.10	82.27	81.37	80.39
IBK	54.28	53.49	52.48	52.17	52.01
Bagging	59.99	59.78	60.36	60.38	60.41
Random Forest	63.85	62.58	61.99	61.31	61.15

V. CONCLUSIONS AND FUTURE SCOPE

In this approach, Naive Bayes Multinomial and Logistic classifiers achieved good results for gender prediction in reviews domain. It is observed that the proposed approach attained good results compared to BOW approach and also observed that only POS N-grams were not sufficient to improve the performance of the gender prediction. In the proposed approach, three measures were used such as term frequency in individual document, term frequency in all documents of specific profile group and the TFIDF scores of terms. The prediction accuracies were affected based on the term weight measure and document weight measure.

As a future work, experiments were performed with various profile combinations to increase the accuracy of the prediction of the profiles. Future efforts put into introducing the sentiment analysis to identify the differences in style of text written by various authors. Further it is planned to evaluate this approach on different datasets as well as to explore the profile specific features to improve the performance.

REFERENCES

[1] Koppel M. S. Argamon and A. Shimoni, Automatically categorizing written texts by author gender, Literary and Linguistic Computing, pages 401-412, 2003.

[2] J. Schler, Moshe Koppel, S. Argamon and J. Pennebaker (2006), Effects of Age and Gender on Blogging, in Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, March 2006.

[3] Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. J. Mach. Learn. Res. 8, 1261–1276 (Dec 2007).

[4] Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker: Gender Differences in Language Use: An Analysis of 14,000 Text Samples. Discourse Processes pp. 211–236 (2008).

[5] E. Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. JASIST.

[6] M. Sudheep Elayidom , Chinchu Jose , Anitta Puthussery ,Neenu K Sasi "Text classification for authorship attribution analysis", Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, September 2013.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, Benno Stein: Overview of the 6th International Competition on Plagiarism Detection. CLEF (Working Notes) 2014: 845-876.

[14] 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.
[15] 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.
[16] 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.
[17] Yasen Kiprov1, Momchil Hardalov2, Preslav Nakov3, and Ivan Koychev4, "SU@PAN'2015: Experiments in Author Profiling", Proceedings of CLEF 2015 Evaluation Labs, 2015.
[18] 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.
[19] 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.
[20] Adam Poulston, Mark Stevenson, and Kalina Bontcheva, "Topic Models and n–gram Language Models for Author Profiling", Proceedings of CLEF 2015 Evaluation Labs, 2015.
[21] T. Raghunadha Reddy, B.VishnuVardhan, and p.Vijaypal Reddy, "A Survey on Authorship Profiling Techniques", International Journal of Applied Engineering Research, Volume 11, Number 5 (2016), pp 3092-3102.
[22] http://members.unine.ch/jacques.savoy/clef/index.html
[23] Porter, M.F., (2002) "Developing the English Stemmer", http://snowball.tartarus.org/