Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Authorship Attribution of Punjabi Poetry using SVM Classifier

Navinder Kaur, Amandeep Verma
Department of Computer Science, Punjabi university
Punjab, India

Abstract*— Authorship attribution (AA) refers to the problem of identifying the author of an unseen text. From the machine learning point of view, AA can be viewed as a multiclass, single-label text-categorization task. This task is based on the assumption that the author of an unseen text can be discriminated by comparing some laxical features extracted from that unseen text with those of texts with known authors. In this paper Authorship Attribution process is applied on punjabi poetry corpus consisting of punjabi poetry written by 10 different poets. Several features such as word ngram, character ngram are used as input to a linear SVM (Support Vector Machine) classifier and the validation of the proposed system is evaluated on the basis of Precision, Recall, F-score and Accuracy.

Keywords*— Authorship Attribution, SVM Classifier, Punjabi poetry corpus, Tf-idf Weighting, Word n-gram, Char n-gram

I. INTRODUCTION

Natural language processing (NLP) [1] is the technology for dealing with human language, which appears in writings, publications, stories, emails, web pages, tweets, product descriptions, newspaper stories, social media, in thousands of languages and diversity etc. It is used to build language models which can be used to distinguish between languages. The goal of NLP is to build a representation of text that adds structure to unstructured natural language. Word structure refers to syntactic or semantics of text. Syntactic deals with capturing grammatical relationship among text and semantics means capturing the meaning devolved by the text. Authorship Attribution (AA) has been area of interest from field of text mining, NLP. Authorship Attribution (AA) is an act of discovering the author of anonymous document. It is required to Figure out the author of the historical writings whose authors are not known. There were disputes regarding the authorship of old and historical important text so, there was a need to develop a system which can be proficient to recognise the true author of the text. So Authorship Recognition (AR) comes in existence [2]. AR is divided into various fields: - Authorship attribution, Authorship verification, Authorship Discrimination, Plagiarism detection and Data indexing & segmentation. Authorship attribution is a stylometry research field which is used to identify the author of unattributed text using various techniques of text mining. Authorship Verification includes the testing of the authorship by verifying whether the giving piece of text is written by claimed author or not. It can also be checked by giving negative examples. For example, two well-known cases of disputed texts in English includes the disputed Federalist papers [3] and the 15th book of Oz [4]. AR also plays a vital role in identifying the plagiarism i.e. whether the given piece of text belongs to other author or not [5]. Discrimination of authorship checks whether two different texts written by same author [6]. AA is used in computer forensic, criminal law, military intelligence and humanities research. Several features such as characters, Character-n-grams, Words, Words-n-grams, Rare words, Textual feature, function words, Rhythm feature [7] can be extracted for AA. It can be implemented using various classifiers namely SMO-SVM (Sequential minimal Optimization based Support Vector Machine), Stamatatos distance, Manhattan Distance , MLP (Multilayer Preceptron), Linear Discriminant Analysis, Maximum entropy learning [2] [8] . It will help in allocating the author to the unseen text based on feature words, which includes writing style of author, punctuation marks, words which may be independent of the writing.

Literature shows various techniques available which determine the author of document. According to the literature most of work was done on English and Arabic language. Indian regional languages such as Hindi, Punjabi etc. lack behind in this area. Punjabi is an Indo-Aryan language which is 10th most spoken language in the world. It is the native language of about 150 million people [9] moreover, it is very rich language consisting of vast literature but due to lack of awareness of researchers not much work is done on Punjabi language. Although some topic summarisation, topic modelling, name entity recognition systems has been developed for Punjabi language but AA system is not yet developed. In this paper, Authorship Attribution System is implemented using Python 3.4 for AA of Punjabi poetry. The whole process consists of training the network for 856 poems whose authors are known then the anonymous poem is given as input to the network which gives the most appropriate name of the author as output.

II. RELATED WORK

Authorship attribution research starts in 1887, when Mendenhall [8] reported using word-length distributions to study certain of work of John Stuart Mill , comparing them to work by others on the same topic. **K. luyckx** *et al.* uses the feature extraction techniques to test and verify the results [10]. Distribution of n-grams of lexical features which are

identified by Chi-square metric is represented numerically in feature vectors. Both eager and lazy supervised learning methods are used for classification. **K. Shaker** *et al.* reviewed the hybrid of evolutionary search and LDA approach [8]. In this approach usage of function words that are specific words which are used by the author in distinct way which may or may not relate to the subject matter, is main criteria for specifying the author of the text. Above approach is tested on Arabic text as well as on English text which consist of about 70 function words. **S. Wang** *et al.* hinge on Rhythm feature for Chinese text AA [11]. Rhythm matrix represents timing pattern among the words in the text consisting of few bytes in memory. The rhythm matrix of each text writing is created and compared to find out the similarity between them. The similarity is measured based on 2 parameters: Euclidean distance and Kullback-Leibler Divergence which gives an accuracy of 80%. **S. Raghavan** *et al.* signifies uses of Probabilistic Context-Free Grammars (PCFG) for AA [12] which involves constructing grammar of text given and using same as language model for classification. The test document given is parsed using authors grammar and assigns it to author whose PCFG produced highest likelihood for the document. PCFG-E model is group of bag-of-words, Max Entropy classifier and n-gram language model. **T. Solorio** *et al.* takes the idea of specifying similarity among text and different authors to generate modality [13]using Meta features which are helpful in short text i.e. online forum domain. The goal of author is to build a model which learns the distinguishable characteristics in the written work of the author. **Escalante** *et al.* purposes the use of Local Histograms (LH) of character n-grams [14] Which retained the sequential information in the document. Histograms are widely used in text categorisation as well as document visualization. Locally weighted bag of words (LOWBOW) framework is used to create the local histogram which are then smoothed by a kernel. LH calculated at the similar locations is compared to each other using 3 approaches: Diffusion, Euclidean distance and chi-square. **S. Ouamour** *et al.* put forth the character N-gram approach [2]. Authors examine the authorship of Arabic books written by ten Arabic travellers. Different features such as characters, character-bigram, character-trigram and character-tetra gram are used for authorship attribution. Stamatatos distance, Manhattan distance, Multi-Layer Perceptron (MLP) and Support Vector Machines (SVM) are used for classification. **R. Ramezani** *et al*. uses 29 distinct textual features for identifying the author [15]. AA attribution can be considered as multiclass and single-label text-categorization task from point of view of machine learning. 3 distinct classifiers are used SVM, K-NN, and C5. **Y. Seroussi** *et al.* use authorship attribution of informal text such as e-mails, blogs with topic modelling [16]. Disjoint Author-Document Topic (DADT) model was proposed that projects authors and documents to two disjoint topic spaces. Latent Dirichlet Allocation (LDA), Author-Topic (AT) and DADT models are implemented on formal as well as informal.

III. DATASET

A. Punjabi poetry Corpus

The poetry corpus is a store having collection of poems related to a particular poet. The poetry of 10 different poets is collected from various websites like http://www.punjabi-kavita.com/ and http://www.shivbatalvi.com/. The poetry corpus includes 865 poems with 175,231 words for training dataset and 456 poems with 80,512 words for testing dataset shown in Table 1. This whole process is carried out to train the network using appropriate set of poetry so that if anonymous poem is given as input it can distinguish it correctly.

TABLE I: STATISTICS OF POETRY CORPUS

	Number of poems	**Number of words**
Training dataset	865	175,231
Testing dataset	456	80,512
Total	1321	255,743

A detailed description of dataset of the proposed study is illustrated in Table 2. The famous poems which are used in the study are given here corresponding to their poets. Further, the count of poems used for training and testing purpose is shown in the corresponding columns. Following it, total number of poems and words are summed up.

TABLE II: DETAILED DESCRIPTION OF DATASET

S. no.	**Name of poet**	**Famous poem**	**Number of poems for training purpose**	**Number of poems for testing purpose**	**Total number of poems**	**Number of words**
1	Baba Bulle Shah	ਬੁੱਲ੍ਹਾ ਕੀ ਜਾਣਾ ਮੈਂ ਕੌਣ	51	51	102	7662
2	Bawa Balwant	ਤੇਰਾ ਇਕ ਦਿਲ ਹੈ ਜਾਂ ਦੋ ?	16	15	31	4293

3	Bhai Gurdass	ਸਤਿਗੁਰ ਨਾਨਕ ਪ੍ਰਗਟਿਆ ਮਿਟੀ ਧੁੰਧ ਜਗ ਚਾਨਣ ਹੋਆ ॥	101	51	152	6561
4	Bhai Veer Singh	ਮਹਿੰਦੀਏ ਨੀ ਰੰਗ ਰੱਤੀਏ ਨੀ !	101	50	151	14519
5	Dhani Ram Chatrik	ਇਕ ਦਿਨ ਕੁੜੀ ਮੁਟਿਆਰ ਇਕ ਨਜ਼ਰ ਆਈ	101	51	152	48417
6	Guru Nanak Dev Ji	ਵਿਦਿਆ ਵੀਚਾਰੀ ਤਾਂ ਪਰਉਪਕਾਰੀ ॥	101	51	152	8619
7	Mohan Singh	ਅਸੀਂ ਨਿਮਾਣੇ ਸਾਵੇ ਪੱਤਰ	101	51	152	18988
8	Professor Pooran Singh	ਲੂਣਾਂ	101	46	147	29920
9	Shiv Kumar Batalavi	ਗ਼ਮਾਂ ਦੀ ਰਾਤ ਲੰਮੀ ਏ ਜਾਂ ਮੇਰੇ ਗੀਤ ਲੰਮੇ ਨੇ ।	101	39	140	29480
10	Waris Shah	ਹੀਰ	101	51	152	7042

IV. METHODS AND PERFORMANCE MEASURES

A. Feature Extraction

Lexical features have been used in the proposed system which can help in precisely identifying the author of the given poem. Word n-gram and character n-gram feature are extracted and are used to train SVM classifier.

Word N-Gram: - The collocation order of N words could be taken into account which is called word N-Gram and is based on this assumption that the authors usually tend to use special words together. This feature is extracted only from Text documents.

N-Gram Character: - This feature is similar to the word N-Gram except instead of words, the collocation and the order of N characters is considered [15]. This feature is extracted only from Text documents.

Features are extracted using the tf-idf (term frequency- inverse document frequency) vectorisation method. Before implementing tf-idf vectorisation it is required to understand and calculate Term Frequency and inverse Document Frequency.

Term Frequency: The weight value is assigned to each term in a document, that depends on the number of occurrences of the term in the document [17]. The simplest approach is to assign the weight to be equal to the number of occurrences of term t in document d. This weighting scheme is referred to as term frequency and is denoted $tf_{t,d}$, with the subscripts denoting the term and the document in order.

Suppose there is a set of Punjabi text documents and wish to determine which document is most relevant to the query "mW dw ipAwr" . Term frequency is calculated by counting the number of times each term "mW", "dw", "ipAwr" occurs in each document.

Inverse Document Frequency: Term frequency suffers from a crucial problem as all terms are considered equally important when it comes to assessing relevancy on a query. In fact certain terms have little or no discriminating power in determining relevance.

For instance, a collection of documents on the "ਓਲੰਪਕ ਖੇਡਾਂ" is likely to have the term ਖੇਡਾਂ in every document. To overcome this problem, mechanism of inverse document frequency is introduced. The proposed mechanism attenuates the effect of terms that occur too often in the collection to be meaningful for relevance determination [17]. It can be achieved by lowering down the weights of terms with high collection frequency, defined to be the total number of occurrences of a term in the collection.

The idea would be to reduce the TF weight of a term by a factor that grows with its collection frequency. Instead, the document frequency DF, defined to be the number of documents in the collection that contain a term t will be used for this purpose.

TABLE III: COLLECTION FREQUENCY (CF) AND DOCUMENT FREQUENCY (DF)

Word	**CF**	**DF**
ipAwr	10422	8760
mW	10440	3997

The reason to prefer DF to CF is illustrated in Figure 5.1, where a simple example shows that collection frequency (CF) and document frequency (DF) can behave differently.

Inverse document frequency can be calculated from DF as below in the equation where, N is the total number of documents in a collection

$$idf_t = \log \frac{N}{df_t} \quad (1)$$

Thus the idf of a rare term is high, whereas the idf of a frequent term is likely to be low. The behaviour of idf has been shown in Table 4 that is calculated corresponding to given document frequency of different words with 807,791 number of document.

TABLE IV: DESCRIPTION OF INVERSE DOCUMENT FREQUENCY

Term	**df_t**	**idf_t**
ਓਲੰuiਪਕ	18,165	1.65
ਖੇਡਾਂ	6723	2.08
ipAwr	19,241	1.62
mW	25,235	1.5

Tf-idf weighting:- The definitions of term frequency and inverse document frequency is combined, to produce a composite weight for each term in each document [17]. The tf-idf weighting scheme assigns to term t a weight in document d given by

$$tf - idf_{t,d} = tf_{t,d} \times idf_t \quad (2)$$

In other words, tf-idf, td assigns to term t a weight in document d that is:

- Highest when t occurs many times within a small number of documents
- Lower when the term occurs fewer times in a document, occurs in many documents
- Lowest when the term occurs in virtually all documents.

At this point, each document can be viewed as a vector with one component corresponding to each term in the poetry corpus, together with a weight for each component that is given by equation (2).

B. SVM classifier

A Support Vector Machine (SVM) is a discriminative classifier prescribed by a separating hyper plane [18] shown in Figure 1. The hyper planes maximizes the distance between itself and nearest training point called as margin which can be used for classification, regression, clustering or other tasks. The nearest data point to the margin is known as support vector. Linear SVM is used in proposed methodology which uses "LinearSVC" class

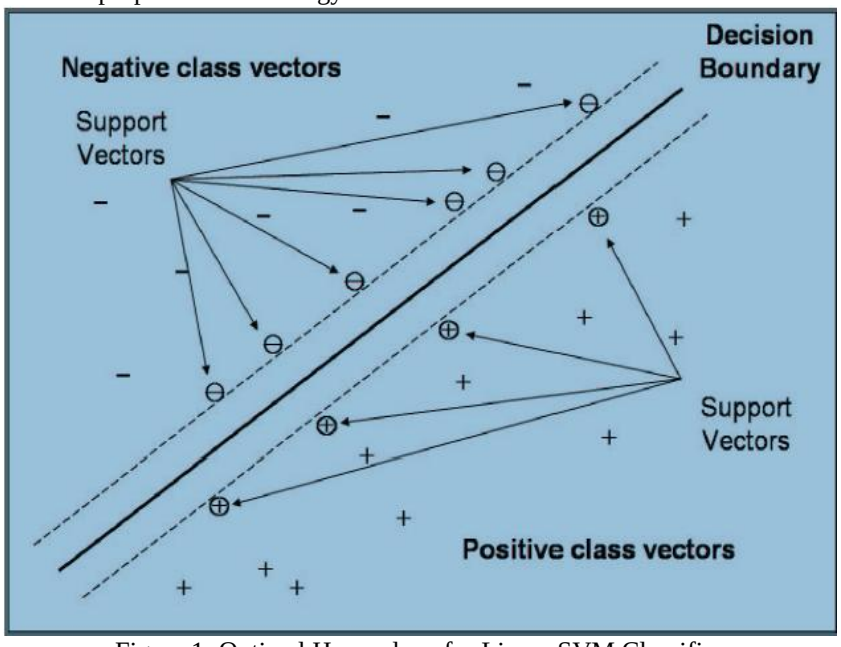

Figure 1: Optimal Hyperplane for Linear SVM Classifier

C. Performance parameters

The selection of the acceptable metrics in evaluating the performance of the Authorship Attribution system is significant for the result and the validation of the system. The parameters preferred to evaluate the performance of the proposed system are Precision, Recall, F-score and Accuracy.

1) **Precision (P):** Precision measures the ratio of relevant output instances to the total instances obtained from output [19] . It is given as:-

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Where TP is the number of true positives and FP is the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. In the proposed methodology precision can be calculated as.

$$Precision = \frac{Correctly\ identified\ author\ of\ the\ poem}{All\ poems\ retrieved\ as\ belonging\ to\ a\ class}$$

2) **Recall (R):** Recall is ratio of relevant output instances to the total instances [19] moreover; it can also be represented as in equation 4.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Where *TP* is the number of true positives and *FN* the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. This parameter is included to observe nature of output. In this paper Recall can be defined as

$$Recall = \frac{Number\ of\ correctly\ identified\ poems}{total\ number\ of\ poems\ used\ for\ testing}$$

3) **F-Score (F):** It is also known as balanced F-score or F-measure. It gives the overall performance of the proposed system [20]. High value of precision leads to more correctly identified authors than incorrect ones, while high recall means AA model has returned most of the relevant results [20]. F-measure is used to represent the correctness of the AA model. Equation 5 gives the formula for F-measure

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

4) **Accuracy:** The accuracy is the proportion of the test results that is true positive and true negatives among total number of cases.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

V. EXPERIMENTS AND RESULTS

In the proposed system experiment of the Authorship Attribution on Punjabi poetry of 10 different poets is represented. Several features are tested: Char unigram, char bigram, char trigram, word unigram, word bigram and word trigram. SVM classifier is employed by using nltk tool kit [21] for Authorship attribution. For the accomplishment of this task number of poems for the training varies for each author according the availability of their poems, same is the case with number of poems used for testing purpose. The statistics of the database are given above in Table1. Figure 2 and Table 5 shows the Precision, Recall, f-score and accuracy % obtained using different features. Precision value of 81% is obtained from word unigram feature. This precision score is best among all the features that have been employed in this experiment. The same feature obtained the highest recall and f-score value. It is important to mention that an accuracy of 79% is obtained by char tri- gram feature.

TABLE V: PRECISION, RECALL, F-SCORE AND ACCURACY PERCENTAGE

	Precision	**Recall**	**F-score**	**Accuracy**
word unigram	81	78	77	78
word bigram	59	58	56	57
word trigram	49	19	15	19
char unigram	51	50	47	49
Char bigram	72	69	68	69
Char trigram	72	68	39	79

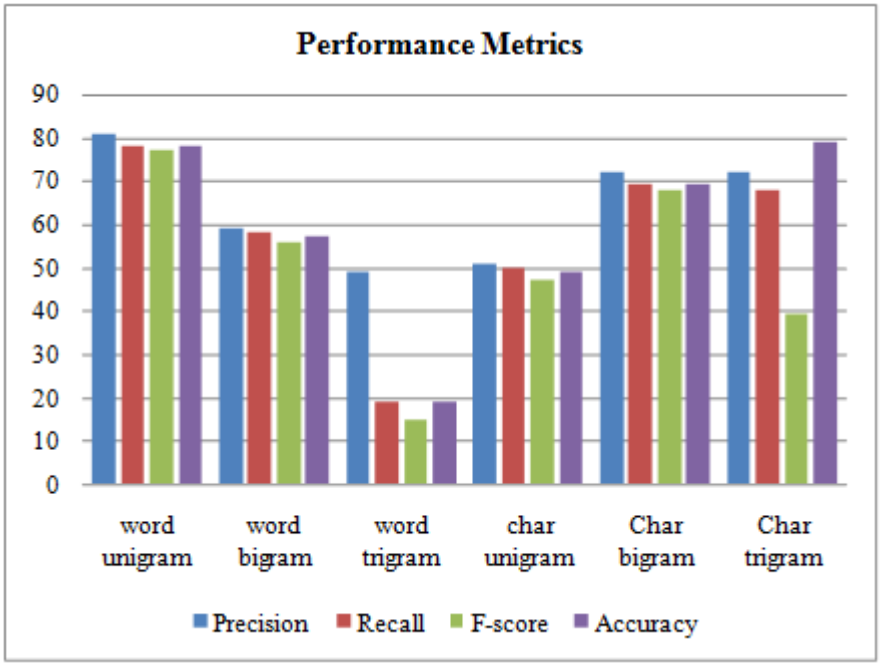

Figure 2: Performance metrics percentage for different used features

Figure 3 summarizes the overall results into some classes of features: the average Precision, Recall, F-score and Accuracy for word- gram and average Precision, Recall, F-score and Accuracy obtained by character- gram feature. It can be clearly observed that the character based features are better than the word based features with 68% Precision, 62% Recall, 51% F-score and 65% Accuracy.

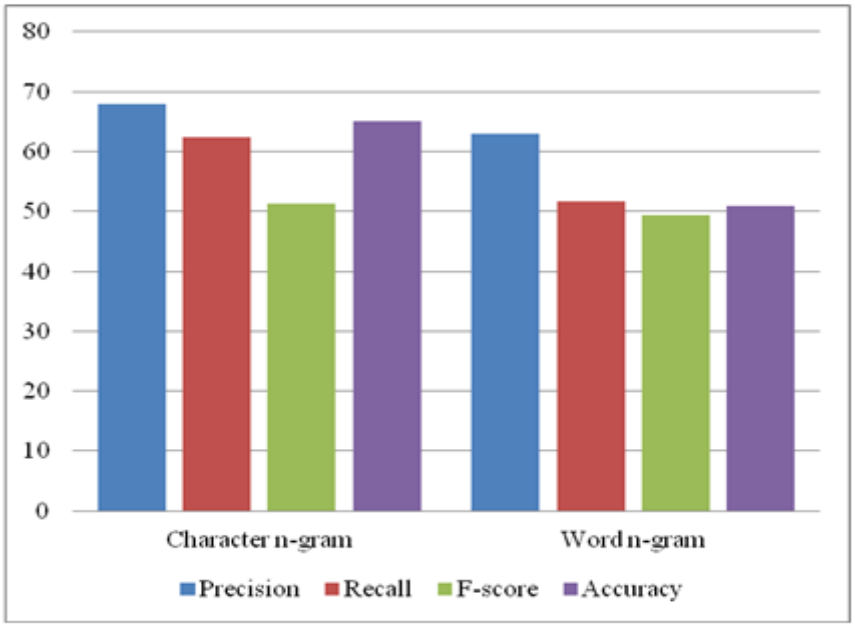

Figure 3: Average Precision, Recall, F-score and Accuracy using SVM classifier

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion

In this research work a new Authorship Attribution task has been experimented on Punjabi poetry written by 10 different poets. Several state-of-the-art features have been tested for Punjabi poetry. For the classification purpose SVM classifier is used which uses 856 poems for training and 465 poems for the testing. Experiments of authorship attribution on poems of 10 different poets, which have been done separately for each feature, show the following remarkable points:

- The character-based features are better than the word -based features;
- We notice a failure for the classifiers using the word-based features (attribution score about 10% and 70%);
- The best features are character-trigrams (accuracy score of 79%).
- Although the size of the texts was too small (between 209 words and 800 words in the text), the performance of the classifier are really interesting (80% of good attribution)

- This new work of authorship attribution, which is one of the rare works done on Punjabi text, shows a real motivation and interest for this language. It also shows that, the rules are almost the same for both English and Punjabi.
- Consequently, one can exploit most of the
- Knowledge that have been acquired over the time with the English language.

B. Future Scope

Future work concerns the implementation of Authorship Attribution on Punjabi literature. Various other lexical, syntactic, semantic or application specific features can be used to extract the features. The poetry corpus used in the study is limited only to 10 Poets. The corpus can be designed again with more poems and authors to enhance the current poetry corpus. Future work concerns the use of other classifiers, in order to see if it could be possible to outperform the present attribution score. It can also be used in combination with topic modelling so that the performance of the system can be enhanced.

REFERENCES

[1] "NLP." [Online]. Available: https://www.coursera.org/course/nlp. [Accessed: 25-Nov-2014].

[2] S. Ouamour and H. Sayoud, "Authorship attribution of ancient texts written by ten Arabic travelers using character N-Grams," in *Proceedings of International Conference on Computer, Information and Telecommunication Systems (CITS)*, 2013, pp. 1–5.

[3] D. Mosteller, F., Wallace, *Inference and Disputed Authorship: The Federalist*. Readings: Addison-Wesley, 1964.

[4] J. Binongo, *Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution*. 2003, pp. 9–17.

[5] K. Robin and S. Conrad, "A Set-Based Approach to Plagiarism Detection Notebook for PAN at CLEF," in *Proceedings of PAN 2012 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF*, 2012.

[6] H. Sayoud, "Author Discrimination between the Holy Quran and Prophet's Statements," *Lit. Linguist. Comput.*, vol. 27, no. 4, pp. 427–444, 2012.

[7] S. Ouamour and H. Sayoud, "Authorship attribution of ancient texts written by ten arabic travelers using a SMO-SVM classifier," in *Proceedings of International Conference on Communications and Information Technology (ICCIT)*, 2012, pp. 44–47.

[8] K. Shaker and D. Corne, "Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis," in *Proceedings of IEEE UK Workshop on Computational Intelligence (UKCI)*, 2010, pp. 1–6.

[9] "Punjabi language." [Online]. Available: http://punjabiworld.com/Punjabi-Culture/punjabi-history.html. [Accessed: 26-Nov-2014].

[10] K. Luyckx and W. Daelemans, "Authorship attribution and verification with many authors and limited data," in *Proceedings of 22nd international conference on computational Linguistics*, 2008, pp. 513–520.

[11] S. Wang and B. Yan, "Authorship attribution for chinese text based on sentence rythm features," in *Proceedings of IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT)*, 2010, pp. 61– 64.

[12] S. Raghavan, A. Kovashka, and R. Mooney, "Authorship Attribution Using Probabilistic Context-Free Grammars," in *Proceedings of the Association for Computational Linguistics Conference Short Papers*, 2010, vol. 6, pp. 38–42.

[13] T. Solorio, S. Pillay, S. Raghavan, and M. Montes-y-Gómez, "Modality Specific Meta Features for Authorship Attribution in Web Forum Posts," in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 2011, pp. 156–164.

[14] H. J. Escalante, S. Nicol, D. L. Garza, and M. Montes-y-g, "Local Histograms of Character N -grams for Authorship Attribution," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 288–298.

[15] M. Ramezani, R. ; Sheydaei, N. ; Kahani, "Evaluating the Effects of Textual Features on Authorship Attribution Accuracy," in *Proceedings of 3th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2013, pp. 108–113.

[16] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship Attribution with Topic Models," *Assoc. Comput. Linguist.*, vol. 40, no. 2, pp. 269–310, 2014.

[17] C. D. Manning, P. Raghavan, and H. Schütze, "Scoring, term weighting and the vector space model," in *Introduction to Information Retrieval*, Cambridge University Press, 2009, pp. 109–133.

[18] C. Colin and N. Cristianini, "Simple learning algorithms for training support vector machines, Technical Report," University of Bristol, 1998.

[19] R. Jizba, "Measuring Search Effectiveness," 2008. [Online]. Available: https://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching_-_Recall_Precision.pdf. [Accessed: 18-Apr-2015].

[20] D. M. W. Powers, "Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation," 2007.

[21] B. Stevern, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly Media lnc, 2009.

