Authorship Attribution in Bengali Language

Shanta Phani
Information Technology
IIEST, Shibpur
Howrah 711103, West Bengal, India
shantaphani@gmail.com

Shibamouli Lahiri
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109
lahiri@umich.edu

Arindam Biswas
Information Technology
IIEST, Shibpur
Howrah 711103, West Bengal, India
abiswas@it.iiests.ac.in

Abstract

We describe Authorship Attribution of Bengali literary text. Our contributions include a new corpus of 3,000 passages written by three Bengali authors, an end-to-end system for authorship classification based on character n-grams, feature selection for authorship attribution, feature ranking and analysis, and learning curve to assess the relationship between amount of training data and test accuracy. We achieve state-of-the-art results on held-out dataset, thus indicating that lexical n-gram features are unarguably the best discriminators for authorship attribution of Bengali literary text.

1 Introduction

Authorship Attribution is a long-standing and well-studied problem in Natural Language Processing where the goal is to classify documents (often short passages) according to their authorship. Different flavors of the problem treat it as either "closed-class" (train and test authors come from the same set), or "open-class" (test authors may be different from train authors). A related variant is Authorship Verification, where the goal is to verify if a given document/passage has been written by a particular author via, e.g., binary classification.

Although Authoship Attribution in English has received a lot of attention since the pioneering study of Mosteller and Wallace (1963) on the disputed *Federalist Papers*, equivalent work in Bengali – one of the most widely spoken South Asian languages – has spawned only three strands of research till date (Das and Mitra, 2011; Chakraborty, 2012; Jana, 2015). Part of the reason behind this lack of research progress in Bengali Authorship Attribution is a shortage of adequate corpora and tools, which has only very recently started to change (Mukhopadhyay et al., 2012).

In this paper, our contributions are as follows:

- **Corpus**: a new corpus of 3,000 literary passages in Bengali written by three eminent Bengali authors (Section 3).
- **Authorship Attribution System**: a classification system based on character bigrams that achieves 98% accuracy on held-out data (Section 4).
- **Feature Selection**: six types of lexical n-gram features, and selection of the best-performing combination on an independent development set (Section 5).
- **Learning Curve**: how the performance on held-out data changes as the number of training instances varies (Section 6).
- **Feature Ranking**: most discriminative lexical features by Information Gain (Section 7).
- **Feature Analysis**: frequency analysis of discriminative features, grouped by authors (Section 7).

We would like to mention that there are many different ways in which our Authorship Attribution system could potentially improve or be extended (more powerful learning algorithms; syntactic, semantic and discourse features; etc). However, given that we already achieved impressive accuracy values on held-out data (Section 6), such improvements would necessarily be incremental, unless new corpora are introduced that warrant different feature sets and/or classifiers.

2 Related Work

A general overview of the topic of Authorship Attribution has been given in the surveys by Juola (2006), Stamatatos (2009), and Koppel et al. (2009). Unsurprisingly, there are many recent studies in English Authorship Attribution. Seroussi et al. (2012) showed that author-topic model outperforms LDA for Authorship Attribution tasks with many authors. They came up with a combination of LDA and author-topic model (DADT – disjoint author-document-topic model) that outperforms the vanilla author-topic model

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 100–105,
Trivandrum, India. December 2015.

Author		Overall	Train	Test	Development
Rabindranath	Mean #words	221.18 (26.08)	215.28 (27.21)	228.60 (23.73)	225.56 (23.09)
	Mean #characters	3232.61 (385.44)	3139.87 (405.12)	3352.16 (339.98)	3298.53 (338.50)
Sarat Chandra	Mean #words	188.99 (29.57)	186.18 (30.29)	192.49 (30.05)	191.12 (26.97)
	Mean #characters	2661.23 (421.84)	2628.64 (432.10)	2695.18 (423.85)	2692.47 (393.02)
Bankim Chandra	Mean #words	841.05 (256.86)	832.38 (250.32)	846.44 (261.31)	853.01 (264.55)
	Mean #characters	13259.56 (4188.04)	13153.46 (4083.46)	13325.55 (4284.03)	13405.75 (4290.50)

Table 1: Corpus statistics. Values in parentheses are standard deviations. Mean and standard deviation are taken across passages.

and an SVM baseline. Seroussi et al. (2014) further showed state-of-the-art performance on PAN 11, blog, IMDB, and court judgment datasets.

As we discussed in Section 1, Authorship Attribution in Bengali is a relatively new problem. Among the three studies we found, Chakraborty (2012) performed a ten-fold cross-validation on three classes (Rabindranath, Sarat Chandra, others) with 150 documents in each, and showed that SVM outperforms decision tree and neural network classifiers. The best accuracy was 84%. An earlier study by Das and Mitra (2011) also worked with three authors – Rabindranath, Bankim Chandra, and Sukanta Bhattacharya. They had 36 documents in total. Unigram and bigram features were rich enough to yield high classification accuracy (90% for unigrams, 100% for bigrams). However, their dataset was not very large to draw reliable conclusions. Further, the authors they experimented with had very different styles, unlike our (more difficult) case where two of the authors often had similar styles in their prose (Rabindranath and Sarat Chandra).

Jana (2015) looked into Sister Nivedita's influence on Jagadish Chandra Bose's writings. He notes that "The results reveal a distinct change in Bose's writing style after his meeting with Nivedita. This is reflected in his changing pattern of usage of these three stylistic features. Bose slowly moved back towards his original style of writing after Nivedita's death, but his later works still carried Nivedita's influence." This study, while interesting, is not directly comparable to ours, because it did not perform any classification experiments. Among other recent studies in Authorship Attribution in Indian languages, Nagaprasad et al. (2015) worked on 300 Telugu news articles written by 12 authors. SVM was used on word and character n-grams. It was observed that F-score and accuracy decrease as size of training data decreases, and/or the number of authors increases.

Bobicev et al. (2013) looked into Authorship Attribution in health forums. In their 30-class classification problem, orthographic features performed well, and Naive Bayes was shown to perform better than KNN. The best accuracy was close to 90%.

Bogdanova and Lazaridou (2014) experimented with cross-language Authorship Attribution. They designed cross-language features (sentiment, emotional, POS frequency, perceptual, average sentence length), and posited that Machine Translation could be used as a starting point to cross-language Authorship Attribution. Using six authors' English books and their Spanish translations, they obtained 79% accuracy with KNN. The best pairwise accuracy was 95%. Nasir et al. (2014) framed Authorship Attribution as semi-supervised anomaly detection via multiple kernel learning. They learned *author regions* from the feature space by representing the optimal solution as a linear mixture of multiple kernel functions.

Luyckx and Daelemans (2008) introduced the important problem of *Authorship Verification*. To model realistic situations, they experimented with 145 authors and limited training data (student essays on Artificial Life). They showed that Authorship Verification is much harder than Authorship Attribution, and that more authors and less training data led to decreased performance. Memory-based learning (e.g., KNN) was shown to be robust in this scenario. An interesting study was presented by van Cranenburgh (2012), where he focused on *content words* rather than *function words*, and showed that tree kernels on fragments of constituency parse trees provide information complementary to a baseline trigram model for Authorship Attribution. Literary texts from five authors were used, and the best (combined) accuracy reached almost 98%.

Sarawgi et al. (2011) attempted to remove *topic bias* for identifying gender-specific stylistic markers. They used deep syntactic patterns with PCFG, shallow patterns with token-level language models, morphological patterns with character-level language models, and bag of words (BoW) with MaxEnt classifier. Per-gender accuracy reached 100% using morphological features on blog data. On paper data, BoW features also reached 100% per-author accuracy for both male and female authors.

3 Corpus

In this work, we focused on Authorship Attribution of Bengali *literary text*, in keeping with prior studies(Das and Mitra, 2011; Chakraborty, 2012; Jana, 2015). Note that with the emergence of social media, it would be completely valid to pursue this problem on news articles, tweets, Facebook posts, online forum threads, blog entries, or other social media outlets. However, the problems with such data are: (a) they are less clean than literary text, leading to a lot of surface variation, and (b) the number of authors is

essentially unbounded, thereby rendering the problem more difficult and lowering accuracy values (Layton et al., 2010; Schwartz et al., 2013).

We chose three eminent Bengali authors for our study, and extracted 1000 passages from the works of each author. The authors are:

1. **Rabindranath Tagore** (1861-1941)
2. **Sarat Chandra Chattopadhyay** (1876-1938)
3. **Bankim Chandra Chattopadhyay** (1838-1894)

Note that all three are male authors, and lived during the golden age of Bengali Renaissance, thus their writing styles could often be very similar – echoing the premises of the original Mosteller and Wallace study (1963). Besides, these authors have an extensive repertoire of works (novels, essays, poetry, songs, dramas, short stories, reviews, letters, critiques, etc) that have been completely digitized for researchers to leverage (Mukhopadhyay et al., 2012).[1]

We sampled 1,000 random passages for each author as follows. We first removed poetry and songs because they are not uniformly distributed across all three authors. Thereafter, we merged the remaining prose in a single large file, and sampled 25 random fragments for each passage (25K fragments in total). We have taken necessary care to ensure that passage contents were *disjoint*.

The above procedure yielded a balanced corpus of passages for the three authors. The corpus is realistic, because it embodies the *fragmentary* nature of realistic authorship attribution scenarios where all too often texts are not recovered in their entirety. Furthermore, it sidesteps the problem of unequal sample lengths (e.g., by having whole documents or books as samples). Our corpus statistics are shown in Table 1. Note that the corpus has been divided into (balanced) train, test, and development sets, with 1,500 samples in the train set, and 750 samples in the test and development sets. Table 1 shows that passages from Bankim Chandra are the longest (on average), followed by Rabindranath and Sarat Chandra. The reason is the former's usage of complex and formal language constructs in his writings which typically led to longer and more intricate fragments.

4 Authorship Attribution System

We pose the problem as one of supervised classification. With three classes, our accuracy on held-out data reaches 98%.[2] In accordance with previous research, we found that the best results are obtained from *most frequent lexical n-grams*. Among the features we experimented with are:

- **Stop words**: 355 Bengali stop words.[3]
- **Uni, bi, and trigrams**: Word n-grams (n = 1, 2, 3) that are most frequent on the complete dataset.
- **Character bi and trigrams**: Character n-grams (n = 2, 3) that are most frequent on the complete dataset. Whitespace characters were ignored.

We have tried three feature representations on the above categories – binary (presence/absence), tf, and tfidf. While it may seem that such shallow features are not enough to capture the variability and complexity of individual authors, as we shall see in Section 6, the impressive performance values we obtained dispel such doubts.

We used three classifiers from Weka (Hall et al., 2009) – Naive Bayes (NB), SVM SMO, and J48 decision tree – to test their performance on the development set. As shown in Table 2, J48 performs significantly worse than NB and SMO across the board, whereas NB and SMO perform close to each other. We chose NB as our final classifier. This decision is guided by the fact that NB is simpler to conceptualize and implement, and faster to train than SMO.

Note further from Table 2 that word unigrams give the best performance. However, as we shall see in Section 5, best values are obtained from character bigrams (tf), so our final system consists of 300 most frequent character bigrams (tf) as features on Naive Bayes classifier.

5 Feature Selection

As we see from Table 2, best numbers are in the region of stop words, word unigrams, character bigrams, and character trigrams. It is therefore instructive to look into what performance benefit we can achieve by varying the number of features in these categories, along with their representation (binary/tf/tfidf). The results are shown in Figure 1. We observed that the best development accuracy of 97.87% was obtained for 300 character bigrams (tf) feature combination, so we used that combination for our final system.

Note from Figure 1 that in almost all cases, increasing the number of features led to improved performance on the development set. However, overfitting is clearly visible for character bigrams and trigrams beyond a certain number of features (around 300). This observation offers a completely organic feature selection strategy – cut off where the development accuracy dips for the first time. Note also that the features were ordered by Information Gain, so e.g. the *top k* n-grams are the most discriminative *k* n-grams within the most frequent.

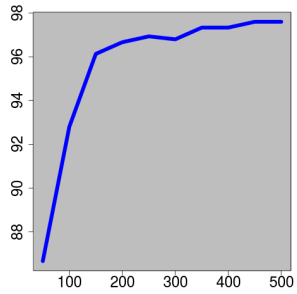
[1] The complete works of these three authors are available from `http://www.nltr.org/`.

[2] A random baseline would achieve only 33% on the same data.

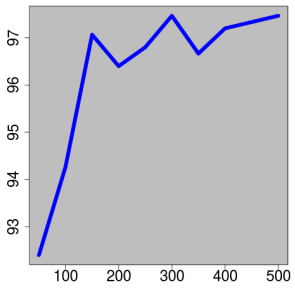
[3] Available at `http://www.isical.ac.in/~fire/data/stopwords_list_ben.txt`.

Feature Representation	Feature Category	J48	NB	SMO
Binary (Presence/Absense)	Stop words	89.73	96.40	96.00
	Word unigrams	92.80	**97.60**	98.40
	Word bigrams	67.87	73.47	73.87
	Word trigrams	36.80	40.00	40.00
	Character bigrams	84.53	96.40	96.40
	Character trigrams	79.60	93.07	92.27
Tf (Term Frequency)	Stop words	92.93	95.73	97.60
	Word unigrams	94.93	97.47	98.80
	Word bigrams	69.20	74.13	74.40
	Word trigrams	36.80	39.07	40.67
	Character bigrams	90.93	97.33	98.67
	Character trigrams	85.07	94.27	96.93
Tfidf (Term Frequency	Stop words	92.53	95.87	97.73
Inverse Document Frequency)	Word unigrams	**95.20**	**97.60**	**98.93**
	Word bigrams	65.87	74.13	74.00
	Word trigrams	36.80	40.00	40.67
	Character bigrams	91.47	97.33	98.40
	Character trigrams	85.07	94.93	97.93

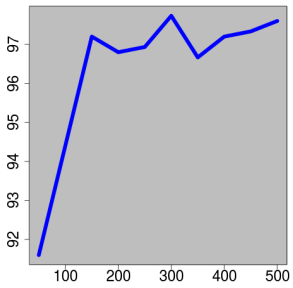
Table 2: Percentage accuracy of three classifiers when trained on the training set and tested on the development set. For each category of feature, 500 most frequent were used. For stop words, there were 355. Best number in each column is boldfaced.



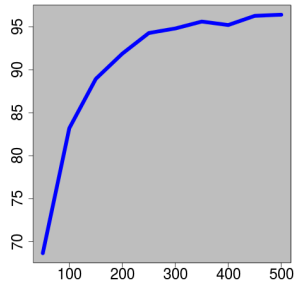
(a) Word unigrams (binary)



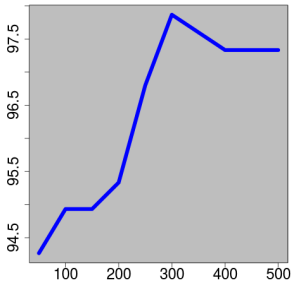
(b) Word unigrams (tf)



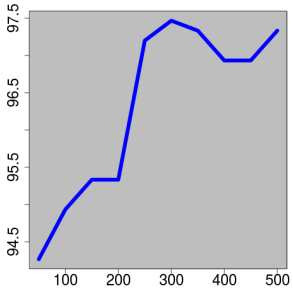
(c) Word unigrams (tfidf)



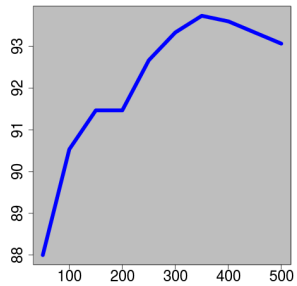
(d) Character bigrams (binary)



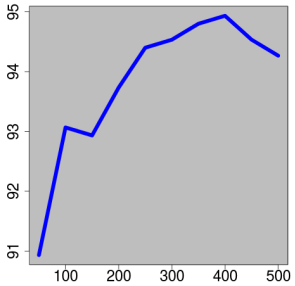
(e) Character bigrams (tf)



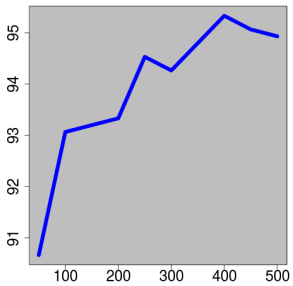
(f) Character bigrams (tfidf)



(g) Character trigrams (binary)

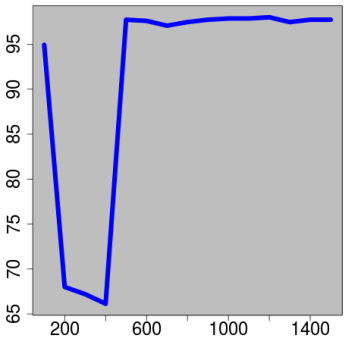


(h) Character trigrams (tf)

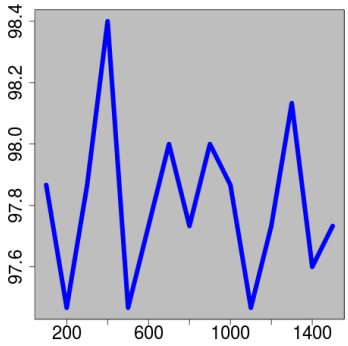


(i) Character trigrams (tfidf)

Figure 1: Impact of number of features on accuracy. X-axis is Number of Features, and Y-axis is Percentage Accuracy on the development set (can be viewed in grayscale).



(a) Training on train set



(b) Training on train + development set

Figure 2: Learning curves. X-axis is Number of Training Samples, and Y-axis is Percentage Accuracy on the **test set** (can be viewed in grayscale).

6 Learning Curve

With the feature set and classifier now optimized on the development data, we re-trained the model on train set (1,500 instances) and train + development set (2,250 instances), and measured accuracy on the **test set** that was untouched so far. In both cases, we obtained 97.73% test accuracy – thereby showing the viability of our approach on completely untouched held-out data. To be noted is the fact that this high test accuracy is similar to the high development accuracy we obtained in Section 5. This is because our samples were drawn from the same *universe* of authors. Furthermore, our test accuracy is superior to the state-of-the-art (84% reported by Chakraborty (2012)), and more reliable because we worked with a much larger sample of passages than (Chakraborty, 2012) and (Das and Mitra, 2011), and because we followed a more rigorous experimental paradigm by splitting our data into three parts and selecting the model on the development set.

We next asked the following question: *Can we reduce the amount of training data, and still get the same (or better) performance?* To answer this question, we plotted two learning curves, shown in Figure 2. Figure 2a shows the case when we train on the training data only, and test on the test data. Figure 2b shows the case when we train on training + development data, and test on the test data. In both cases, we varied the number of training samples from 100 to 1,500 in steps of 100.[4]

We empirically observed that the best test accuracy of 98.4% was obtained for 200 training instances + the development set (the first spike in Figure 2b). In general, the performance almost always stayed within a tight band between 95 and 99%, thus indicating the validity of our approach, and (relative) insensitivity to the number of training examples. We would like to recommend that 500 training examples should be good enough for practical applications.

[4]For Figure 2b, the development set part was fixed; only the training samples were varied.

7 Feature Ranking

We next investigated the most discriminative features among Bengali stop words. The top 20 stop words are shown in Table 3, ordered by their Information Gain (IG) on the training set. Note that apart from pronouns such as তা, এ, কে, and যে, we also obtained do-verbs such as করি, করিয়া, and করিতে in the top ranks. This is an interesting finding.

Further, we show the term frequency of the features in the last three columns of Table 3, grouped by authors. Note that in all cases, Bankim Chandra's passages contain many more of the stop words, indicating that the passages are longer and more complex (as mentioned in Section 3). Among Rabindranath and Sarat Chandra, the variations are less systematic. Sometimes Sarat Chandra has more occurrences of a particular word, sometimes Rabindranath.

8 Conclusion

We presented the first large-scale study of Authorship Attribution in Bengali. As part of our study, we constructed a corpus of 3,000 literary passages from three eminent Bengali authors. On our balanced dataset, we performed classification experiments, and reached state-of-the-art test accuracy of 98% using character bigrams (tf) as features and Naive Bayes classifier. We further showed how performance varied on held-out data as the number of features and the number of training samples were altered. In most cases, we obtained a range of accuracy values between 95 and 99%. We analyzed the most discriminative features (stop words) and showed that the passages from one of our authors (Bankim Chandra) was longer and more complex than others. To the best of our knowledge, our study is the first reliable attempt at Authorship Attribution in Bengali, especially because prior studies had very limited training and test data. As future work, we would like to extend our approach to other forms

Rank	Word	PhTr	Meaning	IG	F_R	F_S	F_B
1	ই	/ee/	–	0.898	9249	8572	42637
2	করি	/kori/	I do	0.864	2268	2398	14174
3	তা	/taa/	then/that	0.852	4899	3815	17559
4	এ	/ay/	this/these	0.851	5595	4537	18624
5	কে	/kay/	who	0.843	4640	3186	14045
6	না	/naa/	no	0.828	4442	4321	19162
7	যা	/jaa/	that/which	0.792	2035	1778	10172
8	কি	/ki/	what	0.776	2152	3085	9912
9	যে	/jay/	that/which	0.762	2779	1946	10074
10	বা	/baa/	or	0.748	4446	4361	16436
11	পর	/por/	after/other	0.677	1332	1323	6738
12	তাহা	/taahaa/	that	0.672	1364	1402	6634
13	জন	/jon/	person/people	0.660	958	646	4657
14	করিয়া	/koria/	having done	0.657	1031	1245	5306
15	ও	/o/	and/also	0.639	2294	2408	8669
16	এই	/ey/	this	0.629	1055	752	4441
17	নাই	/naai/	no/not	0.601	413	425	3157
18	করিতে	/koritey/	to do	0.600	461	380	3215
19	হইতে	/hoitey/	to be	0.578	502	343	2989
20	সে	/shay/	he/she	0.578	2812	2128	7028

Table 3: Feature ranking of most discriminative Bengali stop words (by Information Gain). PhTr = phonetic transcription (approximate); IG = information gain (on training set); F_R, F_S, and F_B denote term frequency of the feature in the training set for Rabindranath, Sarat Chandra and Bankim Chandra, respectively.

of text, such as blogs, news articles, tweets, and online forum threads.

References

Victoria Bobicev, Marina Sokolova, Khaled El Emam, and Stan Matwin. 2013. Authorship Attribution in Health Forums. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 74–82. INCOMA Ltd. Shoumen, Bulgaria.

Dasha Bogdanova and Angeliki Lazaridou. 2014. Cross-Language Authorship Attribution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Tanmoy Chakraborty. 2012. Authorship Identification Using Stylometry Analysis in Bengali Literature. *CoRR*, abs/1208.6268.

Suprabhat Das and Pabitra Mitra. 2011. Author Identification in Bengali Literary Works. In Sergei O. Kuznetsov, Deba P. Mandal, Malay K. Kundu, and Sankar K. Pal, editors, *Pattern Recognition and Machine Intelligence*, volume 6744 of *Lecture Notes in Computer Science*, pages 220–226. Springer Berlin Heidelberg.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Siladitya Jana. 2015. Sister Nivedita's influence on J. C. Bose's writings. *Journal of the Association for Information Science and Technology*, 66(3):645–650.

Patrick Juola. 2006. Authorship Attribution. *Found. Trends Inf. Retr.*, 1(3):233–334, December.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60(1):9–26, January.

Robert Layton, Paul Watters, and Richard Dazeley. 2010. Authorship Attribution for Twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*, pages 1–8, July.

Kim Luyckx and Walter Daelemans. 2008. Authorship Attribution and Verification with Many Authors and Limited Data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 513–520, Manchester, UK, August. Coling 2008 Organizing Committee.

Frederick Mosteller and David L. Wallace. 1963. Inference In An Authorship Problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Sibansu Mukhopadhyay, Tirthankar Dasgupta, and Anupam Basu. 2012. Development of an Online Repository of Bangla Literary Texts and its Ontological Representation for Advance Search Options. In *Workshop on Indian Language and Data: Resources and Evaluation Workshop Programme*, page 93. Citeseer.

S. Nagaprasad, T. Raghunadha Reddy, P. Vijayapal Reddy, A. Vinaya Babu, and B. VishnuVardhan. 2015. Empirical Evaluations Using Character and Word N-Grams on Authorship Attribution for Telugu Text. In Durbadal Mandal, Rajib Kar, Swagatam Das, and Bijaya Ketan Panigrahi, editors, *Intelligent Computing and Applications*, volume 343 of *Advances in Intelligent Systems and Computing*, pages 613–623. Springer India.

A. Jamal Nasir, Nico Görnitz, and Ulf Brefeld. 2014. An Off-the-shelf Approach to Authorship Attribution. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 895–904. Dublin City University and Association for Computational Linguistics.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86, Portland, Oregon, USA, June. Association for Computational Linguistics.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship Attribution of Micro-Messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891. Association for Computational Linguistics.

Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. 2012. Authorship Attribution with Author-aware Topic Models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–269, Jeju Island, Korea, July. Association for Computational Linguistics.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship Attribution with Topic Models. *Volume 40, Issue 2 - June 2014*, pages 269–310.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556, March.

Andreas van Cranenburgh. 2012. Literary authorship attribution with phrase-structure fragments. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 59–63, Montréal, Canada, June. Association for Computational Linguistics.