Data Mining for Text Attribution

Anonymous

ABSTRACT

Authorship attribution is the task of determining the author of a particular piece of text. The essence of this task lies in being able to represent the document in a way that captures the authors or individuals writing style.
As part of this project, we explore the task of attributing an author to a given corpus. To accomplish this, we collate our corpora from freely available online sources, unencumbered by copyrights. We then extract features from each of these documents to prepare our data-set. Next, we store our data in database management system of our choice. Further, we apply data mining techniques like classification and clustering to be to train a model to be able that learn from the data and confidently attribute a label to a new corpus. Finally we test our models performance, analyze which features contribute the most to this task and visualize the results.

1. INTRODUCTION

Authorship attribution is a well studied problem in the domain of Computational linguistics, Natural Language Processing, Text Mining and Information Retrieval. Interest in this field began in the later half of the 19th century primarily to be able to quantify writing style and have more objective justifications for verifying authors of disputed works by authors like Shakespeare and Bacon. The goal of this task is to be able to match a corpus to its author using some measure of similarity. Forensic linguistics and plagiarism detection are some of the common applications where such techniques are widely used.

To assemble our data-set, we collect corpora from freely available, not for profit online resources such as the gutenberg [1], Internet Archive [2] and Open Library [3]. We crawl these websites for books written by our chosen set of authors and download the books to use as raw data from which our data-set may be created. We now have now a set of *documents* whose meta-data like genre etc.. may have missing values, further the books also contain *noise* in the form of table of contents, licence information and chapter headings among others. Clearly we must devise a strategy to effectively be able to clean these books of such unwanted sentences or phrases which would ultimately affect the performance of our final models.
To be able to attribute an author to a paragraph, we must first break down our document into sentences and paragraphs. Clearly, we cannot simply feed this raw textual data to our model for training, instead we must find a way to be able to represent the corpus that captures the writing style of the author in a way that the model can *interpret*. To this end, we make the use of stylometry techniques to extract features like number of words, average word length, number of stop words, number of special characters, number of uppercase words, number of rare words, number of articles, number of verbs and number of nouns and number of characters among others.
Once our pre-processing and feature extraction steps are complete, we store this data into a database management system of choice. This would help us cleanly organize our organize our data-set and be able to map the raw text to its extracted features / attributes. Once we have our data managed, it would be easier to query records on the fly as needed by our model both for training and evaluation purposes.
Finally, now that we have our data-set assembled, cleaned and in a database, we can apply data mining techniques to depending on our application. Since we want to be able to attribute an author to an anonymous piece of text, then for classification we would treat the author name to be the label for the example. For clustering, we have choices based on which to cluster the data like the total number of authors in the data or the different genres of books in the data-set or even the time period when the books were published. But before we proceed with any information mining on our data-set, we must look at what our data looks like. This starts by understanding the descriptive statistics of the attributes, visualizing them and how they contribute towards our evaluation metric of choice. Finally, we perform a comparative analysis of the attributes before selecting classification or clustering algorithms to use.

In the upcoming sections we explore in detail the deliverables as described above paragraphs. In Section 2 we discuss our reason for choosing this project in the language processing domain and some applications of the same. In Section 3 we talk about some related work in the domain and discuss the current approaches in Section 4. Next in Section 5 we examine our project setup including collection, pre-processing, assembly into the database management component. The section further describes our attributes, their statistics, visualization and comparative analysis. We then discuss the mining techniques applied. Finally, in Section 6 we analyze their performance on the test set and comparison between them.

[1] `https://www.gutenberg.org`
[2] `https://archive.org/`
[3] `https://openlibrary.org`

2. MOTIVATION

Authorship is way of giving credit and has important implications. Other than the fundamental issue of integrity and honesty it implies responsibility for the work published. Although there are several approaches to authorship attribution, *stylometry* has been a popular technique in related literature. As a part of this paper we explore the technique for the purpose of authorship attribution. The main concept behind this statistical approach to authorship attribution is that each writer expresses themselves in a different way and thus have varying writing patterns. While this problem applies to more than just literature like law, forensics, security and academia, in our project we employ the technique for the purposes of attributing an author to piece of text in a book.

3. RELATED WORK

With the vast amount of textual information available in today and the advances in Natural Language Processing [8], there have been several studies in the text categorization domain. While there may be no direct way of determining the author of an anonymous corpus, there have been many successful solutions put forward. TODO Cite Machine learning in automatic text categorization.

The process of determining the correct author of a text involves several steps. Firstly, we need to select suitable style markers from the text such which may be extracted from its lexical, semantic and syntactic features. In [13] the authors explore the task of determining the most discriminatory features for attributing authors to newspaper text. In [1] the authors employ the use of Support Vector Machines (SVM) to determine the authors from German text using grammatical tags and word bi grams as features.

In [10] the authors use tri-gram word features with Markov Models which compares probabilities of occurrence of words based on the words that precede it. They employ several methods like extracting stylometric features like function word frequency and applying multiple discriminant analysis. Further they test their methods by applying it on the Federalist Papers whose authorship is partially disputed yet has scholarly consensus.

In [3] the authors discuss a few methods of authorship attribution on a data-set of Bengali literature using stylometric techniques. They collect a corpus of eight political writers and select style markers based on the statistical analysis of the extracted features. Next, they use a multilayer feedforward neural network and SVM classification model to be able to attribute an author to each document. Finally, they compare the results of two voting systems created by SVM and MLP classification.

As a part of [12] the author proposes a technique to caluclate metric called the Z score which would standardize the how a particular vocabulary in a text compares to the rest of the document. Next the author describes a simple authorship attribution scheme by defining an author profile using the distance measure between averaged text representations. Finally, the author evaluates their algorithm on a collection of about 5k newspaper articles in both English and Italian.

In the next section we discuss the current approaches employed in the task of authorship attribution and some of the techniques in the papers discussed above.

4. CURRENT APPROACHES

In [5] the authors work on the task of determining the author in the case of a large number of candidate authors. They use over 18,000 blog posts and collect the author data. Finally they use content, style tfidf and meta learning to determine authors.

In [6] the authors show how author attribution studies based on fewer number of authors leads one to overestimate the importance of features extracted from the corpus. The authors demonstrate the robustness of their proposed memory based approach for author attribution by using a corpus of 145 authors. In [7] focus on a large scale of authors as well using over 10,000 words per author. They use Bayesian multinomial logistic regression to in their one of k authors experiment. Further they test the topic author independence and the "odd man" out in the corpus.

As a part of [15] the author introduces a new technique for the linguistic profiling of authors which uses a large number of linguistic features as test profiles which are then compared against average profiles for groups of texts. The authors apply their technique to the Dutch Au- thorship Benchmark Corpus (ABC-NL1) as a test. While the author describes the use a variety of linguistic features such as syntax, semantics, pragmatics and vocabulary, he chooses to use only simple lexical features to demonstrate the efficacy of the proposed technique.

In their work in [4] the authors propose a new feature set set of k-embedded-edge subtree patterns that hold more syntactic information than previously proposed feature sets. Further the authors propose a new approach of directly mining them from a given set of syntactic trees. They go on to show that their proposed approach reduces the computational complexity of having syntactic structures as the feature set. Finally, the authors go on to demonstrate their approach on a real world data sets showing their approach to be more accurate than earlier studies.

In [14] the authors present an automatic approach to dealing with real world unrestricted textual data. They use a text processing tool to analyze the input text and use in addition to analysis dependent style markers in addition to the markers as output of the tool. The authors use no frequency counts or any other lexical features and show that with the proposed set of style markers, they are able to identify the authors of a weekly newspaper text using multiple regression. This is especially interesting since their approach is fully automated and requires no manual text sampling or processing.

The authors in [11] explore the use of word and character level sequence kernels to attribute authors to a short piece of text. They compare the performance of two probabilistic Markov chain based approaches using several configurations of the sequence kernels on a data set of 50 authors. In [2] the authors emphasize on constructing and visualizing the evidential traits in authorship attribution. The authors propose a vizualizable evidence driven approach to facilitate the work of cyber investigation. They evaluate their proposed mehtod on the real like Enron email data set and are able to achieve a higher accuracy when compared to traditional methods. Moreover, their output can be easily interpreted and vizualized as evidential traits.

In the next section, we describe our current project setup and discuss the methodologies employed to unmask authors.

5. PROJECT SETUP

5.1 Data Collection

As mentioned earlier, the textual data is available in abundance in today's world. However, data collection has become more challenging due to restrictive access to resources i.e. literacy, novels ,and books. After research we found a website gutenberg[4] where these resources are freely available unencumbered by copyright issues. From this website, we have identified and gathered some books for this project. Also, we have ensured to include mix contents and genres books for this project.

As part of data collection, we have gathered books for many authors, but will include below mentioned authors for now. In the next phase we aim to be able to include more authors and add more data for existing authors from the data that we have already collected. For this phase of the project, we perform a preliminary analysis with the following authors.

- Abraham Lincoln
- Andrew Lang
- Charles Darwin
- D H Lawrence
- Jacob Abbott
- Oscar Wilde

5.2 Preprocessing

As we are working with real-world, unstructured textual data, we need to be able to represent the data in a way that our models can interpret. We pre-process the collected data in order to follow a naming convention to name a record (i.e. books or novel) mentioned below to cleanly process the input file.

AuthorName_TitleOftheBook.txt

Also, it is important to analyze each authors books to remove extraneous information i.e. table of contents, letter name, and references present because these information does not contribute towards authorship attribution. Hence, these information is removed from the sample books and will be removed from all inputs books. Some examples of such information include chapter names and numbers, index and some other metadata which must be incorporated in other ways.

5.3 Data Statistics

Before conducting any form of analysis, it is important to know understand the collected data and the information contained in it so we can better use the data and visualize it. As part of this phase, we only conduct basic descriptive analysis which would be extended to author specific statistical analysis such as their descriptive statistics and word clouds to provide a visual representation of how a particular author uses words.
Some basic statistics about the data collected are as follows,
No of authors: 50
No of books: 3033

[4] www.gutenberg.org

5.4 Features extraction

As we are trying to identify the authorship of a text book, we need to analyze the writing style of the authors and find some pattern or features / attributes. We use the traditional stylometric approach to extract features from the corpus. As part of the initial work, we have identified some important features listed below. While this initial feature set is limited to mainly lexical features, we plan to extend the feature set to include syntactic and semantic representations of the text as well feature. Finally we would have a dataset which has the features extracted on a paragraph level.

- Paragraph length
- No of sentence
- Maximum sentence length
- Each word length
- Unique words in paragraph
- No of stops words in paragraph
- No of comma used in paragraph
- No of special character used
- No of upper case words
- No of Article used

5.5 Classification

To unmask the authors of a sample paragraph, we apply classification techniques such as the sklearn [9] implementation of Naive Bayes.

The Naive Bayes classification algorithm has following attributes.

Naive Bayes assumption:

$$P(w_1, w_2 \dots w_n | c_j) = \prod_i^n P(w_i | c_j)$$

Naive Bayes classifier:

$$\hat{c} = \mathbf{argmax}_{c_j \in C} \hat{P}(c_j) \prod_{w_i = 0 \in d}^{n} \hat{P}(w_i | c_j)$$

We are able to achieve an initial classification accuracy of about 47% which is significantly above the baseline which prove that our extracted features are actually contain useful information to predict the author of a short piece of text. As part of the next phase we would include more classification techniques like SVM, decision trees and Random Forest to name a few.

5.6 Clustering

In this phase we perform K Means clustering on the data and found the most optimal K which corresponded to the number of authors in our data as expected.

K Means can be mathematically represented as

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - \mu_i^{(t)} \right\|^2 \leq \left\| x_p - \mu_j^{(t)} \right\|^2 \forall j, 1 \leq j \leq k \right\} \quad (1)$$

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \quad (2)$$

With just our basic features, we are able to a higher than baseline accuracy of about 30%. In the next phase we would explore on how to improve this accuracy and also other clustering techniques like Density based clustering.

6. RESULTS & FUTURE WORK

As part of this phase result, with the basic features/attributes extracted as mentioned above, the Naive Bayes model gives 47% of accuracy for sample of 5 books for each author.

Also, The Kmeans algorithm implemented as part of the project gives about 30% of accuracy.

For each section we describe the future work that would be involved which mainly includes adding more data, better feature representation and exploring other mining techniques as well.

7. REFERENCES

[1] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1–2):109–123, May 2003.

[2] S. H. H. Ding, B. C. M. Fung, and M. Debbabi. A visualizable evidence-driven approach for authorship attribution. *ACM Trans. Inf. Syst. Secur.*, 17(3), Mar. 2015.

[3] A. S. Hossain, N. Akter, and M. S. Islam. A stylometric approach for author attribution system using neural network and machine learning classifiers. In *Proceedings of the International Conference on Computing Advancements*, ICCA 2020, New York, NY, USA, 2020. Association for Computing Machinery.

[4] S. Kim, H. Kim, T. Weninger, J. Han, and H. D. Kim. Authorship classification: A discriminative syntactic tree mining approach. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 455–464, New York, NY, USA, 2011. Association for Computing Machinery.

[5] M. Koppel, J. Schler, S. Argamon, and E. Messeri. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 659–660, New York, NY, USA, 2006. Association for Computing Machinery.

[6] K. Luyckx and W. Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, page 513–520, USA, 2008. Association for Computational Linguistics.

[7] D. Madigan, A. Genkin, D. Lewis, S. Argamon, D. Fradkin, L. Ye, and D. Consulting. Author identification on the large scale. 01 2005.

[8] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490, 2012.

[10] T. J. Putnins, D. J. Signoriello, S. Jain, M. J. Berryman, and D. Abbott. Advanced text authorship detection methods and their application to biblical texts. In *SPIE Micro + Nano Materials, Devices, and Applications*, 2006.

[11] C. Sanderson and S. Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, page 482–491, USA, 2006. Association for Computational Linguistics.

[12] J. Savoy. Authorship attribution based on specific vocabulary. *ACM Trans. Inf. Syst.*, 30(2), May 2012.

[13] J. Savoy. Feature selections for authorship attribution. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, page 939–941, New York, NY, USA, 2013. Association for Computing Machinery.

[14] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic authorship attribution. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, page 158–164, USA, 1999. Association for Computational Linguistics.

[15] H. van Halteren. Linguistic profiling for authorship recognition and verification. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 199–206, Barcelona, Spain, July 2004.

APPENDIX

A. APPENDIX

The project and setup guidelines can be found at `https://github.com/ratuljain/csci720-proj`.