Misclassifications: The Missing Link

Abitha Thankaraj
Computer Science & Engineering
National Institute of Technology
Calicut, Kerala 673601
Email: abitha95@gmail.com

Aparna Jairaj Nair
Computer Science & Engineering
National Institute of Technology
Calicut, Kerala 673601
Email: aparna1196@gmail.com

Vasudevan N.,
Vinod Pathari
Computer Science & Engineering
National Institute of Technology
Calicut, Kerala 673601
Email: vasudev2020@gmail.com
Email: pathari@nitc.ac.in

***Abstract*—The notion of style is pivotal to literature. The choice of a certain writing style moulds and enhances the overall character of a book. Stylometry uses statistical methods to analyze literary style. This work aims to build a recommendation system based on the similarity in stylometric cues of various authors. The problem at hand is in close proximity to the author attribution problem. It follows a supervised approach with an initial corpus of books labelled with their respective authors as training set and generate recommendations based on the misclassified books. Results in book similarity are substantiated by domain experts.**

I. Introduction

Beyond the essentials of spelling, grammar and punctuation, writing style serves as a literary element that describes how an author uses words to establish mood, imagery and meaning in a text. The authors of this work attempt to model a book recommendation system, based on the similarities between various authors in their style of writing.

Several recommendation systems that exist for books depend primarily on customer reviews, genre, customers' purchase history and occasionally on recently read, viewed or bought books. These reviews do not take into account the writing style of the author. The motivation behind this is drawn from the need for a recommendation engine that suggests books most similar to the one selected, by profiling authors based on their writing styles in different books across various genres in the corpus, rather than suggesting books that fall into the same genre.

The rest of the paper is organized as follows. Section II gives an overview of the problem. The subsequent section details the previous work done in the authorship attribution problem. Section IV outlines in detail the design of the system and the work plan followed to model it. The results obtained from the same are tabulated and analyzed in Section V. The last section, meaningful deductions are made from our proposed model and concludes the paper.

II. Problem Definition

To model a recommendation engine that suggests books most similar to the one selected, by categorizing books solely based on the writing style employed which is effectively captured by character n-grams.

The problem can be viewed as a modification of the authorship attribution problem. When a book by a certain author gets misclassified as another, this research attempts to derive meaning from this *error* in classification and observe if it suggests that both authors may share similarity in writing style.

III. Related Works

A survey of the various automated approaches to attributing authorship, examining their characteristics for both text representation and text classification is presented by Efstathios Stamatatos[7]. Different types of stylometric features relevant to the Authorship Attribution problem, namely: lexical, character, semantic, syntactic and application specific are proposed. Under lexical features, word-grams model (word frequencies) and the bag-of-words model were suggested as the baseline. Character n-grams have also proven as an effective measure to quantify an authors writing style. Apart from feature extraction and selection, the profile-based and instance-based machine learning algorithms are compared to reveal that it is easier to capture and represent various kinds of stylometric features through an instance-based approach.

The robustness of authorship attribution based on character n-gram features under cross-genre and cross-topic conditions is studied by Efstathios Stamatatos[8]. It has been demonstrated that the most effective stylometric features are function words (the most frequent words in the training set) and character n-grams, though the combination of several feature types typically enhances the performance of an attribution model.

Upendra Sapkota et al.[6] portray the power of character n-grams as highly efficient features that capture information about affixes and punctuation. The observation can be attributed to character n-grams capturing information regarding: lexical content, syntactic content, and style (by means of punctuation and white spaces). The authors preferences for particular patterns of punctuation are captured effectively by character n-grams.

John Houvardas and Efstathios Stamatatos[4] propose a variable-length n-gram approach to distinguish authors by experimenting over a subset of the new Reuters corpus. The proposed method involves obtaining variable length n-grams

978-1-5090-6367-3/17/$31.00 ©2017 IEEE

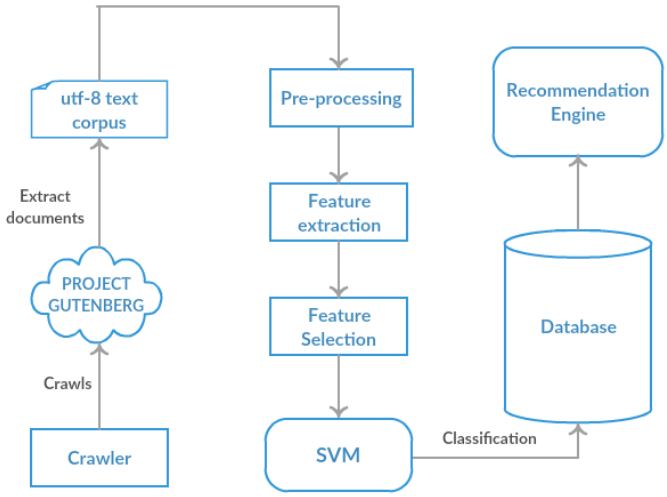

Fig. 1. Design - Recommendation System

from the corpus (3, 4 and 5-grams; as studies prove they provide the best results) and selecting a subset of these features. Feature selection is performed by comparing each n-gram with similar n-grams (either longer or shorter than itself) and then keeping only the dominant n-grams. Then, a Support Vector Machine (SVM) is trained using the reduced feature set.

The above mentioned papers have helped us identify character n-grams as the highly impactful feature in solving the authorship attribution problem. Further studies in the variations in the length of n-grams help us select a subset of these features most effectively.

IV. Design and Implementation

The system is designed as follows.

A. Data Collection

The data set is obtained by crawling the Project Gutenberg website[2] and obtaining e-books in plain text UTF-8 format. The collected works of the top five authors in Project Gutenberg are selected for this purpose.

B. Data pre-processing

The downloaded text files are stripped of insignificant information, such as details about the publisher, Project Gutenberg etc. to ensure the presence of relevant text only.

C. Feature Extraction

It is challenging to capture and quantify the stylistic choices of an author. The most frequently repeating patterns encountered in texts by a particular author are observed. It is preferable to capture patterns formed subconsciously by the author and remain stable over text length (measures like tf-idf) as proposed by Efstathios Stamatatos[8]. Several measures that have been proposed for the same include sentence/word length, vocabulary richness measures, function word frequencies, character n-gram frequencies, syntactic and semantic-related measures. In several independent studies, it has been demonstrated that character n-grams , function words (defined as the set of the most frequent words of the training set), and word-grams are among the most effective stylometric features.

Function words are ill suited for our problem as it captures an insignificant amount of meaningful information when building a recommendation system.

As word-grams tend to lean towards the thematic angle of the text (which would serve more useful in a genre classification problem), only character n-grams are studied for stylometric cues since they capture both thematic as well as stylistic information.

Assuming all the available texts fall into the same genre (in our case, classics), this property of character n-grams can be viewed as an advantage since they yield a richer representation of an authors choice of words for a specific theme[8].

- **Character N-Gram Method**
  Character n-grams are independent of the language used and rely on certain syllables or a combination of them to be repeated. These can be used to characterize an authors literary style. Character n-grams are well suited to our problem since they provide a unique combination of lexical, content specific and syntactic stylometric features. Lexical (ex: frequency of character n-grams), content specific (ex: frequency of certain words) and syntactic (ex : punctuation marks, function words) aspects of the stylometric features extracted were exploited in this stage.
  So far, character n-grams have been widely used to solve the authorship identification problem owing to their ability to successfully capture subtle modulations in the lexical, syntactical, and structural level.

D. Feature Selection

Feature-set used in the model has high dimensionality. In order to ensure that only the relevant features are trained, feature selection has been performed.

- **Top k features**
  Character/word n-grams with frequencies that lie within the top 500 frequencies are separated to be used as features.
  Term frequency-Inverse document frequency is used to transform text into feature vectors. Term frequency is the raw frequency of a term in a document[5].

$$\text{TF(t)} = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

Inverse document frequency is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term[5].

$$\text{IDF(t)} = \frac{\log_e(\text{Total number of documents})}{\text{Number of documents with term t in it}}$$

The TF-IDF score is computed as the product of these two quantities[5].

$$\text{TF-IDF(t)} = \text{TF(t)} * \text{IDF(t)}$$

- **Threshold method**
  A threshold limit is specified for the TF-IDF score. If the TF-IDF score falls below the threshold value, it is set to zero.

E. Supervised Learning

A support vector machine with a linear kernel (LinearSVC) is trained on the selected feature set using scikit-learn libraries[1]. The SVM is implemented by training a one-vs-rest classifier for each class.

1) Training: The training corpus consists of books labelled with their respective authors.

2) Testing: Testing is carried out in two phases. The first phase emphasizes the effectiveness of our model in solving the authorship attribution problem. The second phase hones in on the misclassifications used to build our recommendation system.

1) Phase 1 - Using only known authors
   - K-fold cross-validation A 7 fold cross-validation is performed on the dataset to evaluate the accuracy of our model. This involves splitting our dataset into 7 equal sized subsamples first. Of the 7 subsamples, 6 are used to train the SVM and the last subsample is used as validation data to test the model. This validation process is repeated 7 times so as to make sure each fold serves as validation data exactly once. The results from this cross-validation process are then averaged to estimate the accuracy of the model.
   - Subsampling method In this phase, individual cases of the k fold cross validation are studied so as to evaluate the SVM's accuracy in classifying according to the known author.
2) Phase 2 - Addition of books by unknown authors
   After verifying that the model effectively classifies author styles, a test set is created by adding books by authors who are not found in the training set previously and attempt to classify these books into the existing labels. Further, meaningful conclusions are drawn from the same.

F. Recommendation engine

Each author has a distinctive style which may develop throughout the author's career. For example, an author could have radically different styles of writing in the early and later stages of his career. This recommendation system aims to capture similarity in writing styles as opposed to a specific author or genre. This can lead to the similarity in written work by different authors which can be detected by our system. Thus, the misclassified elements in the author identification problem are used to build our recommendation system.

A random test - train split is selected from the given set of books. In the initial training of the model which is done on the training set, it is assumed that an author has only one specific writing style and label each book with its author's style. When the classification system is run on the testing data, the results show some surprising yet similar books (by different authors) that have not been connected before.

G. Item-item collaboration filtering

Item-item collaborative filtering is a form of collaborative filtering for recommendation systems that works on the similarity between items, which is derived from their respective user rating. This is done in two stages. First, the pair-wise similarity between all pairs of books in the system is calculated and a model is built based on the same. In this work, the distance from the hyper-planes generated by the SVM is used as a similarity metric. Then, these similarities are compared and analyzed to produce a list of recommendations.

H. Reordering recommendations

The order in which the recommended books are shown has been determined by the similarity of the writing style between the two books.The relevance of the book in the context of the writing style is measured as the distance from the hyper-plane for that particular class of writing style, the one with the shortest distance from the hyper-plane being the most similar.

V. Results and Analysis

Feature extraction using strategic leave-one-out method yields results that indicate that **character n-grams** are the most efficient feature to be used in out specific problem. This is further reinforced by the coefficients generated by the SVM classifier.

Corpus Description

The break-up of the corpus that has been used is shown below in Table I

TABLE I
Corpus Description

Group	Author	Number of books
0	Arthur Conan Doyle	21
1	Charles Dickens	23
2	Jane Austen	7
3	Joseph Conrad	27
4	R.L. Stevenson	25

K-Fold cross validation

Leave one out K fold cross validation (7 folds) of the same corpus has been tabulated below in Table II. Precision has been employed as a measure of accuracy.

TABLE II
K- Fold Cross Validation Results

Precision	Recall	F-Score
0.9598875	0.9609375	0.9627875

The high precision denotes that each authors writing style gets classified effectively.

Of the k-folds generated, detailed results of one fold is given below for analysis.

TABLE III
PRECISION RESULTS

Group 0	Group 1	Group 2	Group 3	Group 4
1.	0.8571	1.	1.	0.8571

Weighted precision : 0.9285

TABLE IV
RECALL RESULTS

Group 0	Group 1	Group 2	Group 3	Group 4
1.	0.75	1.	1.	0.9230

Weighted recall : 0.9285

TABLE V
F-SCORE RESULTS

Group 0	Group 1	Group 2	Group 3	Group 4
1.	0.8	1.	1.	0.8889

From the above tables III. IV and V, it is inferred that author 1 (Charles Dickens) and author 4 (R.L Stevenson) have been misclassified as each other. This leads us to believe that they share similar writing styles[3].

Addition of unknown authors

Books by authors not included in the original corpus were taken as the test set. The training set taken was the original corpus used. The SVM was trained on this training set. The results of the testing tabulated below denote the similarity in writing styles found in books of unknown (to the SVM) authors and authors whose writing styles have been learned by the SVM.

Table VI shows how a book by a new author is classified into the 5 existing groups in the training set.

TABLE VI
RESULTS ON CLASSIFICATION OF UNKNOWN AUTHORS

Book title	Original Author	Group
The Mysterious Affair at Styles	Agatha Christie	0
The Secret Adversary	Agatha Christie	3
Biographical notice of Ellis	Charlotte Bronte	2
The Clue of the Twisted Candle	Edgar Wallace	3
The Daffodil Mystery	Edgar Wallace	0
The Angel of Terror	Edgar Wallace	0
Ivanhoe	Sir Walter Scott	4
Jane Eyre	Charlotte Bronte	4
Rob Roy	Sir Walter Scott	4
Shirley	Charlotte Bronte	4
The Lady of the Lake	Sir Walter Scott	4
The Professor	Charlotte Bronte	4
The Talisman	Sir Walter Scott	1
Villette	Charlotte Bronte	4
Waverly	Sir Walter Scott	1
Wuthering Heights	Emily Bronte	1

From Table VI, it can be observed that Biographical notice of Ellis and Acton Bell by Charlotte Bronte has been labelled as being similar to Jane Austen. However, the remaining books by Charlotte Bronte have been labelled as R. L. Stevenson.

Further, two out of three works of the author Edgar Wallace have been classified as Arthur Conan Doyle and the third book as Joseph Conrad. The similarity with Arthur Conan Doyle may be attributed to the general theme of detective fiction. Books by Agatha Christie, another popular detective fiction themed author get misclassified as Arthur Conan Doyle and Joseph Conrad, further cementing our assumption that detective themed novels follow a similar writing style. Joseph Conrad, although of the adventure genre (a genre closely related to detective mysteries) would be a good surprise in the recommendation system.

Books by Sir Walter Scott have been classified as both Charles Dickens and R. L. Stevenson, this identified similarity is justified as all three authors share a distinct preference for the Scottish dialect as established by Anna Faktrovich.[3]

VI. CONCLUSION

On experimenting with various literary features in an attempt to study their contribution to an author's writing style, it is observed that character n-grams (followed by word-grams and function words) are the most effective in characterizing an authors literary profile. From the results of classification based on this feature, it is inferred that books of unknown authors mapped to known authors implies a similarity in their writing habits. Using this concept, a recommendation engine that generates a list of recommendations in a slightly different fashion as compared to the existing ones.

The proposed recommendation system relies solely on the content of the book as opposed to the user reviews as in popular search engines today, thus resulting in unbiased recommendations.

REFERENCES

[1] Documentation of scikit learn : Svm library. http://scikit-learn.org/stable/modules/svm.html. Accessed: 02-05-2017.

[2] Project gutenberg. http://www.gutenberg.org. Accessed: 02-04-2017.

[3] Anna Faktorovich. *Rebellion as Genre in the Novels of Scott, Dickens and Stevenson*. McFarland, 2013.

[4] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 77–86. Springer, 2006.

[5] Daniel Jurafsky. Speech and language processing: An introduction to natural language processing. *Computational linguistics, and speech recognition*, 2000.

[6] Upendra Sapkota, Steven Bethard, Manuel Montes-y Gómez, and Thamar Solorio. Not all character n-grams are created equal: A study in authorship attribution. In *HLT-NAACL*, pages 93–102, 2015.

[7] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[8] Efstathios Stamatatos. On the robustness of authorship attribution based on character n-gram features. *JL & Pol'y*, 21:421, 2012.