Machine Learning-Based Authorship Identification in Web Fictions

Fangying Zhan
fzhan4@wisc.edu

Weijia Cao
wcao34@wisc.edu

Yuan Tian
tian74@wisc.edu

1. Introduction

In this project, we will focus on solving the issue of authorship identification implementing machine learning methods. By applying several different algorithms and techniques, we will try to maximize the accuracy of identifying the correct authors of web fictions, allowing the model to help identify anonymous articles. Moreover, our project will be devoted to classifying different authors' writing styles. Hence, it is plausible for us to compare writers, checking if some of them who readers formerly believed to be very similar in fact differ significantly in their wording, and if some of them that readers believed to apply totally different writing styles are actually more similar in their "stylistic features". That being the case, authorship identification can be implemented to examine the existence of plagiarism as well. Using fan fictions as our dataset also allows us to see whether these imitators' work successfully imitate the original authors.

1.1. Related Works

There have been many works in the past that focused on improving the author analysis, including identification, verification, and profiling. Multiple methods have been implemented to achieve the goal such as uses of univariate or multivariate measures that can reflect the style of a particular author, as well as statistical machine learning techniques. But all methods attempted were relatively flawed until Michal Rosen-Zvi presented the first systematic study of authorship identification by using enhanced version of LDA which has the ability to identify all hidden topics from large numbers of features and presents them as LDA topics, thus serving for dimensionality reduction and making it attractive for text analysis problems.[1]

There are people's works applying the author identification into practical use. A K-nearest neighbor (kNN) algorithm that is based on the authorship verification method was introduced by Halvani et al. for identifying authorship and the task of the PAN 2013 challenge.[2] Zamani described authorship identification as an important method to look for the existence of plagiarism in law and journalism.[3] An end-to-end digital investigation (EEDI) was proposed by Ding et al. for a visualizable evidence-driven approach (VEA) in order to smooth the way for cyber investigation.[4] Furthermore, the news authorship identification task was dealt with by Zhou and Wang in different levels, including author, article, word, and sentence, utilizing both deep and non-deep algorithms with the GloVe word vectors that they used as the pre-trained word vectors.[5]

1.2. Methodology

Authorship identification is one of the popular challenges from the field of Natural Language Processing (NLP), where the state of the art involves a variety of methods and techniques from machine learning and deep learning. With web fictions in English as raw data and the corresponding authors as class labels, our challenge falls in the category of supervised learning problems. The goal of our group project is to examine the use of different types classifiers in solving this supervised learning problem and compare the performances of different combinations of methods and tricks with a focus on the web fiction dataset. Specifically, classifiers that we mainly studied in STAT451, including kNN, C4.5 decision tree, Random Forest, and Multinomial Naive Bayes, will be involved in this project.

We would like to start examining the methods from the simplest. According to Zipf's Law, in natural languages, the rank-frequency distribution is exactly in an inverse relationship.[6] The ranks of word frequencies in human languages observably differs from person to person. From the standpoint of this empirical law, it might be possible to build classifiers simply based on word ranks and frequencies.

We will be incrementally adding text data preprocessing techniques into our framework. From the state-of-the-art NLP, there are many popular text preprocessing steps including removal of punctuations and stopwords, n-grams, lemmatization and stemming. We will attempt adding some of these steps into our framework and see how these will affect the overall model performance.

An important step of NLP is feature engineering, where text data will be transformed into numeric features that can be fitted into classifiers.[7] Two popular feature engineering techniques that we consider using are Tf-idf and Bag-of-words.

The problem of authorship identification is characterized especially by high dimensionality and low sample size. To tackle this problem, we will also be seeking dimensionality reduction techniques and see if that helps in improving the model performance. Our rough framework will be similar to Figure 1, which is a machine learning-based authorship identification approach proposed by Anwar, Bajwa and Ramzan.

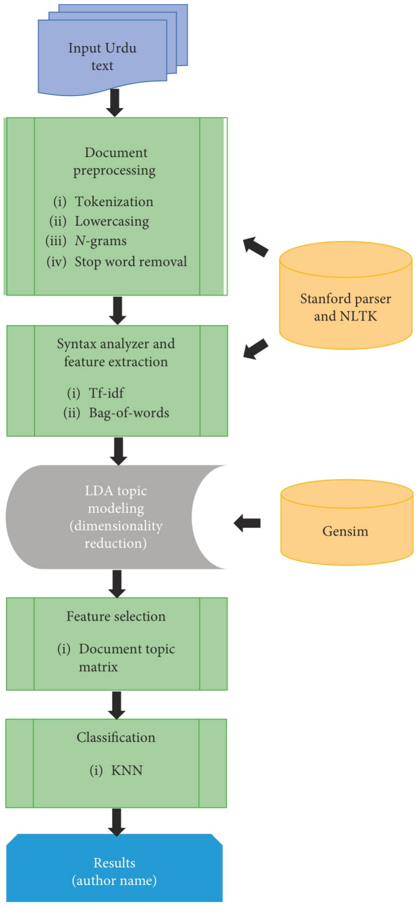

Figure 1. Architecture of text-based forensic analysis approach. Image source:[8].

2. Motivation

Besides eliciting great research interest in the domain of computational linguistics, authorship identification has some significance in many applied areas.

As the use of electronic devices and scientific technology is popularized in the contemporary world, new issues have risen in the area of forensic sciences. For example, it has become more and more knotty to detect criminals who write ransom notes and threatening letters, send libellous or personal attack messages and phishing e-mails, and distribute illegal contents online or in the real world, as people can now take advantage of anonymity or counterfeit their identities and addresses to avoid investigation[9], and words are more often typed rather than written nowadays, leaving traditional criminal investigation techniques, such as authentication of handwriting signatures, nullified. As a result, being able to identifying the authorship of pieces of writing from a known pool of candidates by simply analyzing the writing styles becomes rather meaningful in solving such conundrums. If such a technique can be proven valid and made full use of, it might even help save hostages, check terrorism, protect victims of cybercrimes from being further victimized, preserve intellectual property, or extract useful information for cracking criminal cases from traces in written contents. When we have a pool of possible authors of a piece of textual information, reliable authorship identification techniques can possibly significantly save the time and energy consumed to find the ultimate answer to the question.

Taking one step back from these heroic Sherlock Holmesish imaginations, we should note as well that authorship identification is far more versatile than merely solving crimes. For instance, authorship identification can also help find out most possible authors for literary works published anonymously. Literary works with anonymous authors have never been absent throughout the history of literature. Certainly, we do not want to go against the author's will if they intend to stay anonymous in order to reserve privacy, but the authorship for many literary works went unverifiable for various and complicated reasons. Victor Hugo published the first two editions of his novel The Last Day of a Condemned anonymously, more to tantalize his readers as a sales strategy. But imagine if he was accidentally unable to ever reveal his authorship, and, in the most extreme case-scenario, nobody else knew that it was his work, then we might miss the chance to get to know that the book was ”the most real and truthful of everything that Hugo wrote”, and perhaps authorship identification can come into use in this situation. In fact, the missing authorship of famous literary works does trouble those who do literature studies. Many Academic professionals have been arguing and making guesses about the author of the most popular Arthurian tales, *Sir Gawain and the Green Knight*, ever since the work was rediscovered in the 19th century, whereas the puzzle still remains unsolved except that most of them agree on the author's familiarity with French and French poetry, residency in or origin from a northern English province, and knowledge as well as great value placed in Christianity.[10] One can imagine how it would contribute to the literary world, if a possible author is found for this masterpiece, using statistically founded authorship identification techniques.

Further, authorship identification tells us not only if someone is likely to be the author of a work or not, but also helps us recognize the characteristics of different authors' writing styles, allowing us to do better in comparative literature and in studying different literary genres and literary movements and archiving literary works. It also helps us compare different author's written work to see if they are more similar to each other than they should be. Take what we will focus on in this project as an example. Performing authorship identification on snippets of fan fictions might help us understand if some fan fictions can truly imitate the writing of the original authors, if there is an issue of plagiarism in fan fictions, etc.

3. Evaluation

As we have introduced above, we will compare the performances of different approaches by implementing several classifiers (kNN, C4.5 decision tree, Random Forest, and Multinomial Naive Bayes) and incrementally adding preprocessing techniques in a way that we can interpret the methodology for authorship identification in web fictions. Among all the models, kNN will be set as a baseline.

We will compute training accuracy and validation accuracy to assess the model performance. We will also compute Precision, Recall and F1–Score for model performance analysis. Confusion matrix will also be constructed for interpreting the uniqueness of authors' writing styles.[7]

4. Resources

We plan to extract our textual data from FanFiction, the world's largest fanfiction archive and forum. Specifically, we plan to focus on fanfiction writers of one same original literary work. Ideally, we will find 10 productive fan authors of one same literary work and collect 1000-word long snippets from each of their 8 fan fictions and we will use these snippets to create our training dataset. This is only a tentative training size. We are currently not sure if around 80 snippets of fan fictions will constitute a large enough set. Hopefully we will have a better idea of the size we need in practice.

The computer hardware supporting this project will be each group member's laptop (CPU). The computational tools that we are all familiar with are Python through Jupyter notebook and the machine learning package Scikit-Learn. The package NLTK (Natural Language Took Kit), which is a popular tool for NLP, will also play an important role in our project.

5. Contributions

For the computational part of this project, Yuan will be responsible for data collection and exploratory analysis. Fangying will be responsible for the implementation of methods and models. Weijia will be responsible for model evaluation and the assessment of our experiments. For the writing tasks, Yuan will be responsible for the introduction and related works. Fangying will be responsible for the proposed methods and experiments. Weijia will be responsible for results, discussions and conclusions. We will be making sure that we work closely together and help each other out throughout the whole process.

References

[1] Imran Sarwar Bajwa Waheed Anwar and Shabana Ramzan. Design and implementation of a machine learning-based authorship identification model. *Hindawi*, 2019(7):14, 2019.

[2] Oren Halvani, Martin Steinebach, and Ralf Zimmermann. Authorship verification via k-nearest neighbor estimation. *Notebook PAN at CLEF*, 2013.

[3] Hamed Zamani, Hossein Nasr Esfahani, Pariya Babaie, Samira Abnar, Mostafa Dehghani, and Azadeh Shakery. Authorship identification using dynamic selection of features from probabilistic feature set. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 128–140. Springer, 2014.

[4] Steven HH Ding, Benjamin CM Fung, and Mourad Debbabi. A visualizable evidence-driven approach for authorship attribution. *ACM Transactions on Information and System Security (TISSEC)*, 17(3):1–30, 2015.

[5] L Zhou and H Wang. News authorship identification with deep learning. In *Conference and Labs of the Evaluation Forum, Portugal*, 2016.

[6] Zipf's law, Oct 2020.

[7] Navoneel Chakrabarty. A machine learning approach to author identification of horror novels from text snippets, Jan 2019.

[8] Waheed Anwar, Imran Sarwar Bajwa, and Shabana Ramzan. Design and implementation of a machine learning-based authorship identification model. *Scientific Programming*, pages 1 – 14, 2019.

[9] L. Srinivasan and C. Nalini. An improved framework for authorship identification in online messages. *Cluster Computing*, 22(5):12101 – 12110, 2019.

[10] BookRags. Sir gawain and the green knight author/context.