Deanonymizing Quora Answers

- Author identification of anonymous quora answers using Deep Learning techniques in Natural Language Processing

Project Report CS299

– HARSHIKA (1601CS14)
– ARUNIKA YADAV (1601CS56)

IIT Patna
Computer Science and Engineering

Abstract

Quora is a widely used knowledge sharing website where users can ask/answer questions with the option of anonymity. Some people abuse this privilege by posting bad questions or answers anonymously -- that aren't intended to provide better knowledge, but rather are intended to annoy, provoke, and/or to make a largely rhetorical point. Quora , currently , does not have a system to deanonymize it's answers . Hence it is vulnerable to cyber crime ,plagiarism and other such content related crime .Thus ,in order to curb the anonymity which is useful sometimes and offer a better experience to the users on such a large platform , we offer this solution. The work finds applications to several other tasks like Forensic Linguistics, email spam detection, identity tracing in cyber forensics etc.

Objective

We hope to achieve significant precision on the task of identifying users from their writings with the end-goal of recognizing the authors of these anonymous answers. Previous work indicates that writing style harbors essential cues about authors and we believe that deep learning is a powerful tool to extract such features to distinguish between the various writing styles that people have.In this project, we will tackle this problem at different levels, with different deep learning models and on different datasets.

Background Reading

There has been fair amount of interest in author-identification in previous NLP works with most of the work focussing on manually engineered features:-

1. Comparing Frequency and Style-Based Features for Twitter Author Identification:

   Examines author identification in short texts focussing on messages retrieved from Twitter to determine the most effective feature set for recognizing authors look at Bag-of-Words and style-marker features and use SVMs for the classification task

2. A Comparative Study of Language Models for Book and Author Recognition:

   Evaluates similarity between documents and authors showed that syntactic features are less successful than function words for author attribution

3. A Survey of Modern Authorship Attribution Methods
   Discusses how this scientific field has been developed substantially taking advantage of research advances in areas such as machine learning, information retrieval, and natural language processing over the past few decades.

Approach

In this section, the approaches of authorship identification are elaborated in the following order: datasets, data pre-processing and models.

Datasets

To generate a small-version of the problem we deal with , we have collected dataset from kaggle.com . These corpus has already been used in author identification experiments. 50 authors of texts labeled with at least one subtopic of the class CCAT (corporate/industrial) were selected. That way, it is attempted to minimize the topic factor in distinguishing among the texts. The training corpus consists of 2,500 texts (50 per author) and the test corpus includes other 2,500 texts (50 per author) non-overlapping with the training texts.In our work, we re-organize these 5000 texts into non overlapping 8:2 ratio as train set to test set.

Data Pre-processing

Firstly , we have imported the dataset in the form of text file itself (unlike the regular way of importing it in the form of .csv file using the panda library). Then we replaced all the ‘newline’ characters with a space in order to make the full answer as a whole chunk of words rather .Then we cleaned the data which includes removing punctuations , converting into lower case ,removing the stop words and stemming.

Models

The labels are hidden for a fraction of answers for every author to test our final model and the remaining dataset is used to train on the task of author attribution. We will use machine learning models on engineered features as well as deep learning models for the task.

.

Naive Bayes Model : (Done)

Currently , we have used this model to train our datasets. It is one of the feature extraction algorithms for text which ignores grammar and order of words. We defined a collection of strings called a corpus. Then we used the CountVectorizer to create vectors from the corpus(collection of strings). This made the data ideal for splitting the dataset into a ratio of 8:2 to be taken as the training data and test data respectively.

Once the Naive Bayes algorithms were trained, the classifier was testing using randomly selected excerpts from the testing texts.This particular study resulted in 69% accuracy in classifying documents as being written by a particular author. It should be noted that this is only a preliminary study and serves merely to prove the feasibility of work in the area .

Future Works :

Style marker features Model

The following commonly used style marker features from previous works in author-identification can be used for the classification task :

1. Number of words in the answer
2. Fraction of words that were punctuations
3. Average word length
4. Standard deviation of word length
5. Number of sentences in the answer
6. Average Sentence length
7. Number of digits in the answer

Fingerprint identification Model

It is a well-known technique in forensic sciences. The basic idea of identifying a subject based on a set of features left by the subject actions or behavior can also be applied to other domains. Identifying text authorship based on an author fingerprint is one such application. Finger-printing has mainly been used in the past for plagiarism detection.

Word Frequency model

Each answer is modeled by a feature vector of the length of the vocabulary set and contains the counts for each word in the vocabulary set for that answer The vocabulary set is varied by incrementally adding more tokens in the order of decreasing frequency, using the complete vocabulary works best. The term

frequency-inverse document frequency (also called tf-idf), is a well known method to evaluate how important is a word in a document. tf-idf is a very interesting way to convert the textual representation of information into sparse features.

LSTM with mean-pooling

The LSTM model is basically a recurrent neural network with memory units, allowing the cells to remember or forget its previous state, as needed .Large datasets can be easily handled through this. And we are still reading into it.

Proposed Timeline :

As we are already reading about the various models that we mentioned above , hence we propose to implement one model per week while keeping in mind the accuracy component of the project.

Link to our project :

https://github.com/harshika-arya/Author-Identifier