Style Change Detection

Anthony Odinukwe
odinuk02@ads.uni-passau.de
University of Passau

Shine Jose
jose01@ads.uni-passau.de
University of Passau

Subin Sam Varghese
samvar01@ads.uni-passau.de
University of Passau

1 INTRODUCTION

Style change detection is identifying the number of authors who have written a document by checking their style change in a given context[2]. It uses the idea of stylometry, according to which every author can have a distinct writing pattern[3]. There has been lots of research done on authorship attribution over the years. However, in this project work, we will focus more on identifying a single author from a small portion of text [5]. A proposed three different approaches will be used to carry out analysis on the dataset [6]. These approaches include: Stylometry based approach, content based approach and topic based approach. To identify features, we presented a supervised approach [7] which combines a TF.IDF representation of document features to predict an ensemble of diverse models

According to the project PAN2020[1], The goal of the style change detection task is to identify text positions within a given multi-author document at which the author switches. They provided a document and asked participants to answer the following questions.

- Is the given document written by multiple authors?
- For each pair of consecutive paragraphs in the given document: is there a style change between these paragraphs?

At the end of this project, we will be able to use machine learning algorithm [5] to achieve the following task

- whether documents are written by a single person from his writing style by extracting features from the content of documents.
- if it is written by multiple authors, then we try to predict the exact paragraph where the style change has occurred

2 APPROACH

There are lots of approaches which have been taken over the years to achieve the proposed project. However, given the evaluation framework, which consists of combining two independent measurements, we will build a system that treats them separately by applying a two-step pipeline. The first step is to separate single-author from multi-author documents, and the second step we try to predict the exact paragraph where the style change has occurred in multi-author documents. In our approach, we will divide the whole document into several paragraphs, and cluster the paragraphs based on writing style. Then the feature extraction module is then conducted on each paragraph after splitting. Furthermore, we will calculate the TF-IDF scores of the words in each segment, and use it for the creation of the term-document matrix.

The phases of the project can be categorized into three parts as follows:

- Analysis of Data
- Model Implementation
- Evaluation

The flowchart in Figure 1 below describes the workflow of the project.

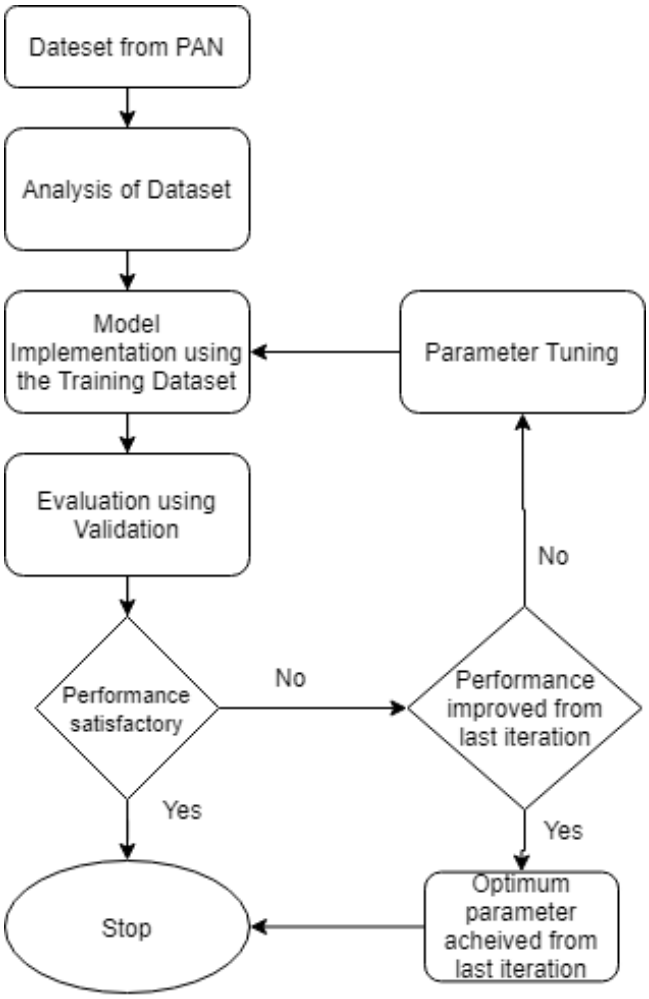

Figure 1: flowchart for the project

2.1 Analysis of Data

To develop our project we have to train and test our algorithm,for that we make use of a dataset which contains different documents in text file and corresponding labels in json file.Afterwards the documents are further stored in two different folders named 'dataset-narrow' and 'dataset-wide'. The data is provided by PAN [3] which

[1] https://pan.webis.de/clef20/pan20-web/style-change-detection.html

conducts event research based on digital forensic and stylometry. The "dataset-wide" contains files from large sets of subject areas such as travel, philosophy, etc, On the other hand "dataset-narrow" contains files from a relatively narrow set of subject matter mostly related to technologies.

The features for this project will be estimated from the text files by using approaches from the paper [7]. Some features to be evaluated are:

2.1.1 Sentence Length and Word Length. Percentage of short sentences and long sentence.Percentage of Words with length more than 6 and percentage of words with length 2 or 3.Average length of words and sentences in the text.

2.1.2 Percentage of POS tags. Using POS tags from NLTK extract percentage of Noun,Pronouns,Adverbs,Verbs,Adjective,prepositions, interjections etc.

2.1.3 Percentage of Frequent words. Using the stop words from NLTK find the frequency of frequent words.

2.1.4 Readability. Using the python package Textstat we can find the readability of the text.It has measurement like Flesch reading ease,Coleman-Liau index,difficult words etc which can be a good feature for obtaining stylometeric features.

More analysis will be carried out to identify the features which features will be more suitable for our implementation.

2.2 Model Implementation

The project is divided into 2 tasks.The Task 1 will be predicting whether the document is written by a single person or multiple persons while the Task 2 will be predicting the location of style change if the document is written by multiple people.The plan is to achieve Task 1 by using different machine learning algorithm like Logistic Regression, SVM, Decision Tree, Random Forest etc,.. on the selected stylometric features.In addition to the previous algorithms we also implements LightBGM algorithm [2] which is a gradient boosting framework that uses tree based algorithms.The LightBGM based approach were used to achieve high accuracy in style change detection task at PAN 2018[7]. For Task 2, two different approaches are introduced.First approach[4] is finding the distance between paragraphs to all the other paragraphs in the document using a vector created by the features identified, and using KNN to identify the different clusters in the document where 'k' is the number of authors collaborated.In the second approach [1] the document is segmented into different paragraphs and for each paragraph a vector is created by the selected stylometric feature.Using these vectors as input to the Gradient Boosting Regression Trees to identify the clusters.

2.3 Evaluation

To evaluate an author's style change detection, we will first define the appropriate measures "performance measures" which were defined in the framework proposed by the PAN committee competition on style change detection: recall, precision, granularity, and overall score. The model will be evaluated by the F1-score measure for each document. The two tasks are evaluated independently based on the obtained accuracy measures. For task 1, we will compute the average F1-score value across all documents and for task 2, we will use the micro-averaged F1-score across all documents. Our two datasets will be evaluated independently and the resulting F1-scores for the two tasks will be averaged across the two datasets. In addition to accuracy and macro f1 score, we evaluate our classifiers using the Ordinal Classification Index (OCI) , as proposed by the PAN committee. The Ordinal Classification Index measures how much the result diverges from the ideal prediction and how inconsistent the classifier is concerning the relative order of the classes.

[2] https://lightgbm.readthedocs.io/en/latest/

REFERENCES

[1] Mikhail P Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, and Vadim V Strijov. 2016. Methods for Intrinsic Plagiarism Detection and Author Diarization.. In *CLEF (Working Notes)*. 912–919.

[2] Peter Müller. 2019. Style Change Detection.

[3] Ifrah Pervaz, Iqra Ameer, Abdul Sittar, and Rao Muhammad Adeel Nawab. 2015. Identification of Author Personality Traits using Stylistic Features: Notebook for PAN at CLEF 2015.. In *CLEF (Working Notes)*.

[4] Abdul Sittar, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. 2016. Author Diarization Using Cluster-Distance Approach.. In *CLEF (Working Notes)*. 1000–1007.

[5] Efstathios Stamatatos. 2009. Intrinsic plagiarism detection using character n-gram profiles. *threshold* 2, 1,500 (2009).

[6] Eva Zangerle, Maximilian Mayerl, Michael Tschuggnall, Günther Specht, Martin Potthast, and Benno Stein. 2020. PAN20 Authorship Analysis: Style Change Detection. https://doi.org/10.5281/zenodo.3660984

[7] Dimitrina Zlatkova, Daniel Kopev, Kristiyan Mitov, Atanas Atanasov, Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2018. An ensemble-rich multi-aspect approach for robust style change detection. *CLEF 2018 Working Nots of CLEF* (2018).