Group 13

Authorship attribution based on Stylometry

Minutes of Meeting (MoM)

Week 1 February 8 February 11	1. Discussed various topics which can be pitched. 2. Ideas discussed: a. Sarcasm detection (Vivek) b. Authorship Attribution based on Stylometry (Sarvani) c. Coronavirus sentiment detection (Lyn)
Week 2 February 17	1. More ideas discussed a. An idea around creating a Chatbot (Piyush) b. Analysis of music lyrics to make new recommendations to users (Nikhil) c. Google QnA labelling on Kaggle (Lyn) d. On input, find similar questions like on QnA websites (StackOverflow)
Week 3 February 25	1. Finalized the topic for the essay after discussing with the professor (Feb 21) 2. Formed research questions 3. Discussed various possible datasets: ● Kaggle ● Pan Competitions ● Project Gutenberg 4. Found Parts of Speech features for the defined text corpus.
Week 3 March 4 March 8	1. Worked on literature review and writing various parts of the essay 2. Kaggle dataset is finalized. 3. Preliminary data analysis and feature extraction. 4. Discussion of possible features and extraction tools. 5. Intermediate results are presented.
Week 4 March 11	1. Discussion of Peer Reviews received. 2. Minor changes to the essay to address smaller feedbacks. 3. Piyush: "*We are gonna get the reviews back. How about we go authorship attribution on that*". Good idea. :-)
Week 5 March 18	1. Created a list of all the changes to be made in the final group document based on the reviews received 2. Looked into several approaches for extracting Text-based and meta-features using different libraries and tools. 3. Discussed valuable inputs received from discussion with other teams.

Week 6 March 25	1. Meta Feature-based approach implementation 2. Decided to use multiple classification models: Naive Bayes, Light GBM and XGBoost. 3. List of important features is selected among all.
Week 7 April 5	1. Content-based features extraction (finally used TF-IDF) 2. Getting feature importance from different classifiers. 3. Metadata and content-based approaches are implemented parallelly to get accuracy. 4. Decided to use at least two of the classification models from Meta-Features based approach 5. Agreed on Incorporating Ridge Classifier, Decision Tree and Logistic Regression also to Meta-feature based approach.
Week 8 April 13	1. Wrote the final essay 2. Formatted the essay in Latex 3. Generated visualizations for easy understanding 4. Got reviews on our final work from two other groups (Group 12 and Group 2), made changes to address the problems.