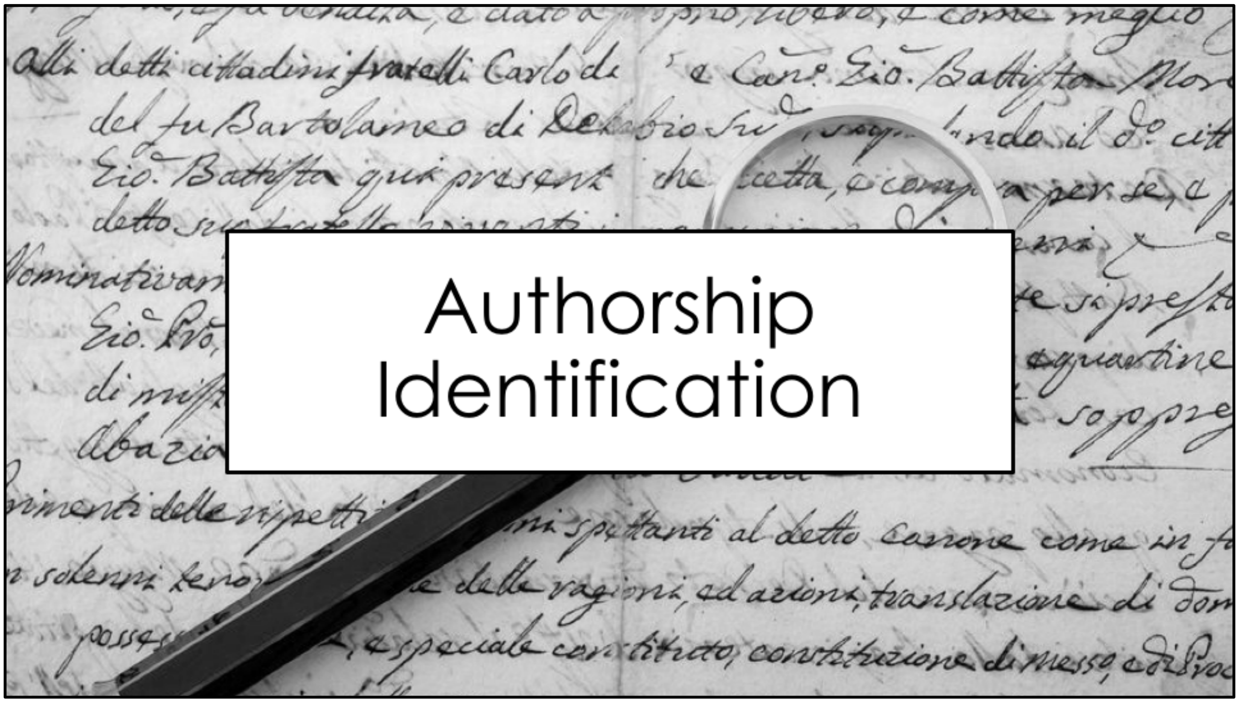
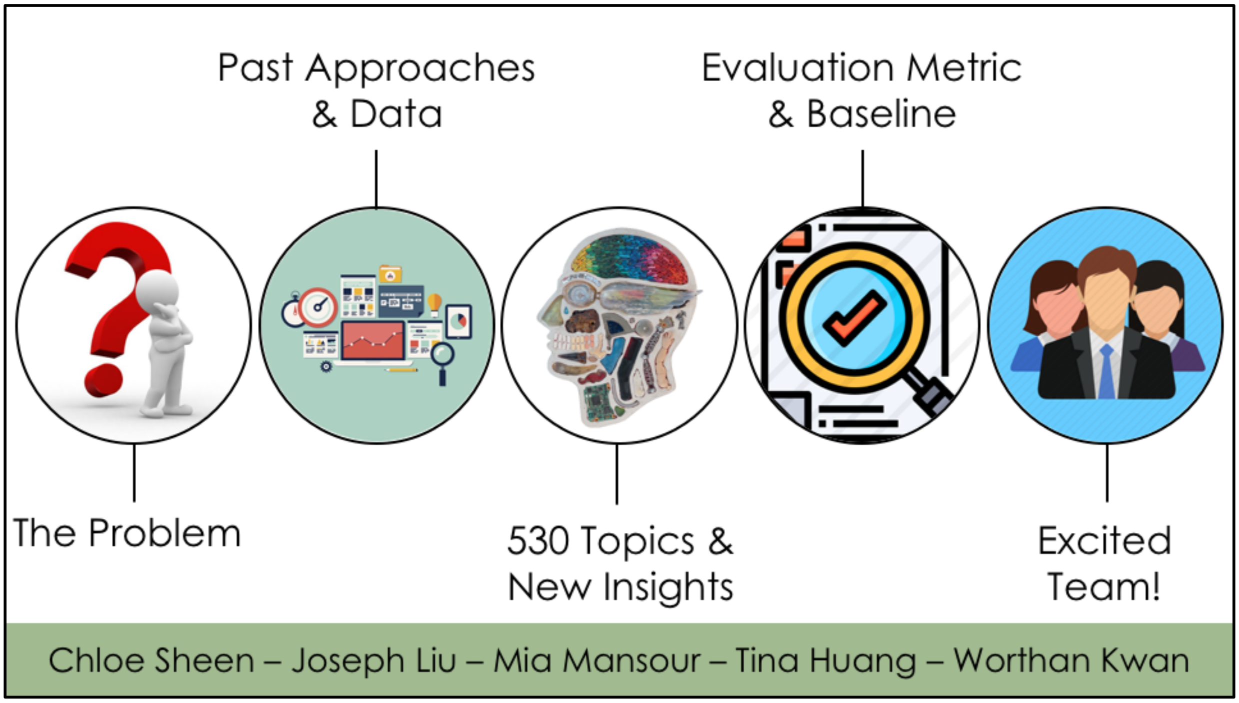
Say you have a piece of text – an interesting anonymous article or a mysterious message. You don’t know who wrote it, but you have a list of possible candidates and it’s up to you to use what you learned in 530 and find out who’s the most likely author.



Authorship Identification

This is what we call Author Identification – the study of linguistic and computational characteristics of the written documents of individuals.



The Problem

Past Approaches & Data

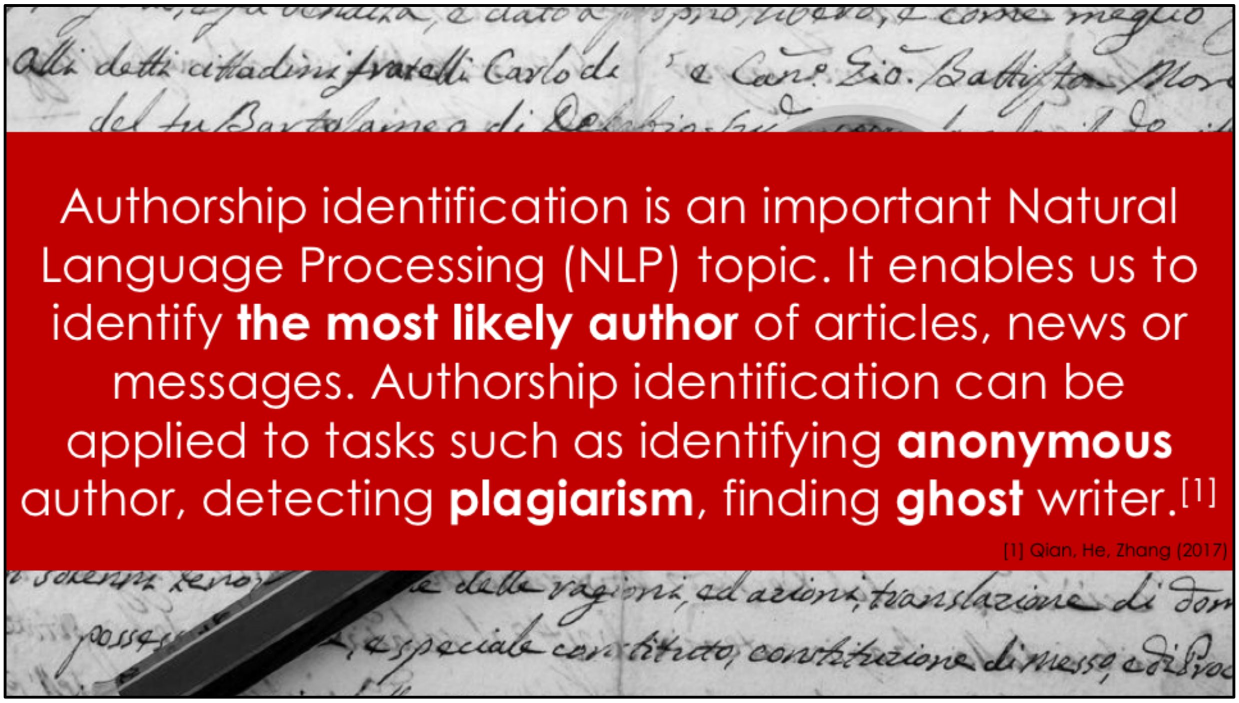
530 Topics & New Insights

Evaluation Metric & Baseline

Excited Team!

Chloe Sheen – Joseph Liu – Mia Mansour – Tina Huang – Worthan Kwan

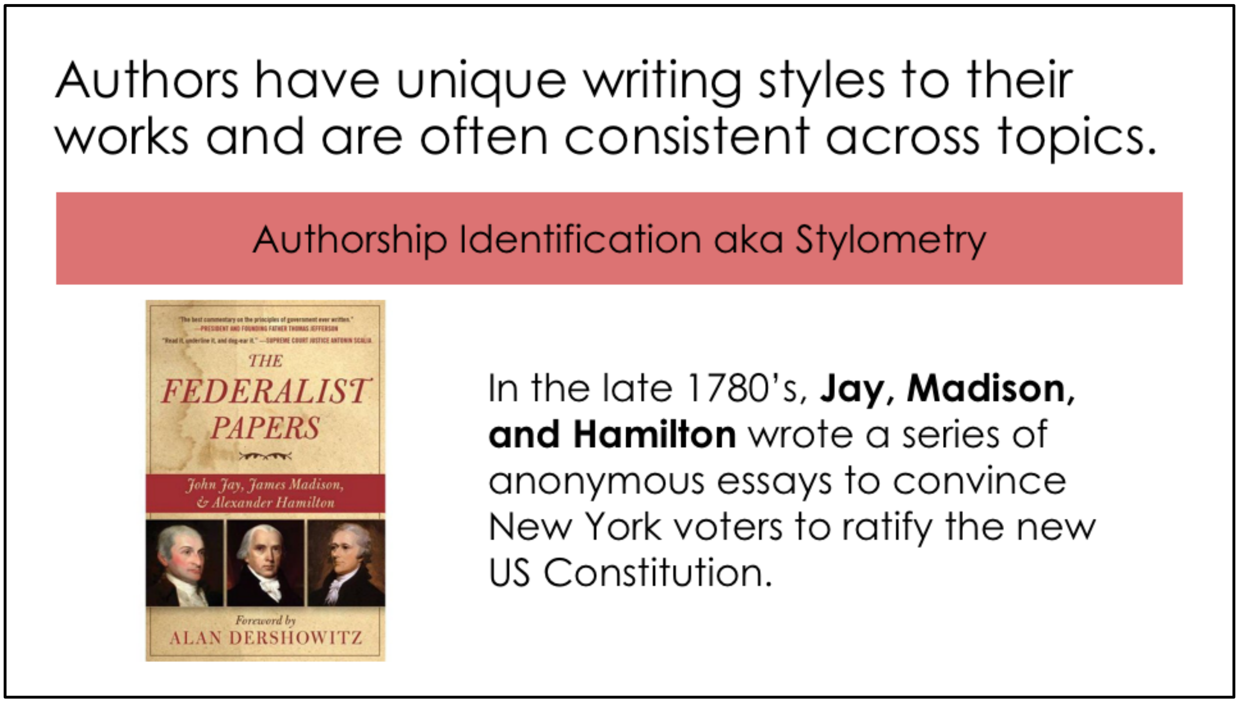
My team will be presenting our project on Authorship Identification. We will first start by diving into what the problem is. We will then present some part approaches we came across during our literature review and we'll present some data sources that can be used for the task. We will explain how some 530 topics were very useful in creating our model and how we learned new things on the topic. We will then explain the evaluation metric used, and the baseline performance. Finally, we will tell you exactly how much our team is exciting to be working on authorship identification.



Authorship identification is an important Natural Language Processing (NLP) topic. It enables us to identify **the most likely author** of articles, news or messages. Authorship identification can be applied to tasks such as identifying **anonymous** author, detecting **plagiarism**, finding **ghost** writer.[1]

[1] Qian, He, Zhang (2017)

We begin by identifying the problem. Authorship identification is an important Natural Language Processing (NLP) topic. It enables us to identify the most likely author of articles, news or messages. Authorship identification can be applied to tasks such as identifying anonymous author, detecting plagiarism, finding ghost writer.



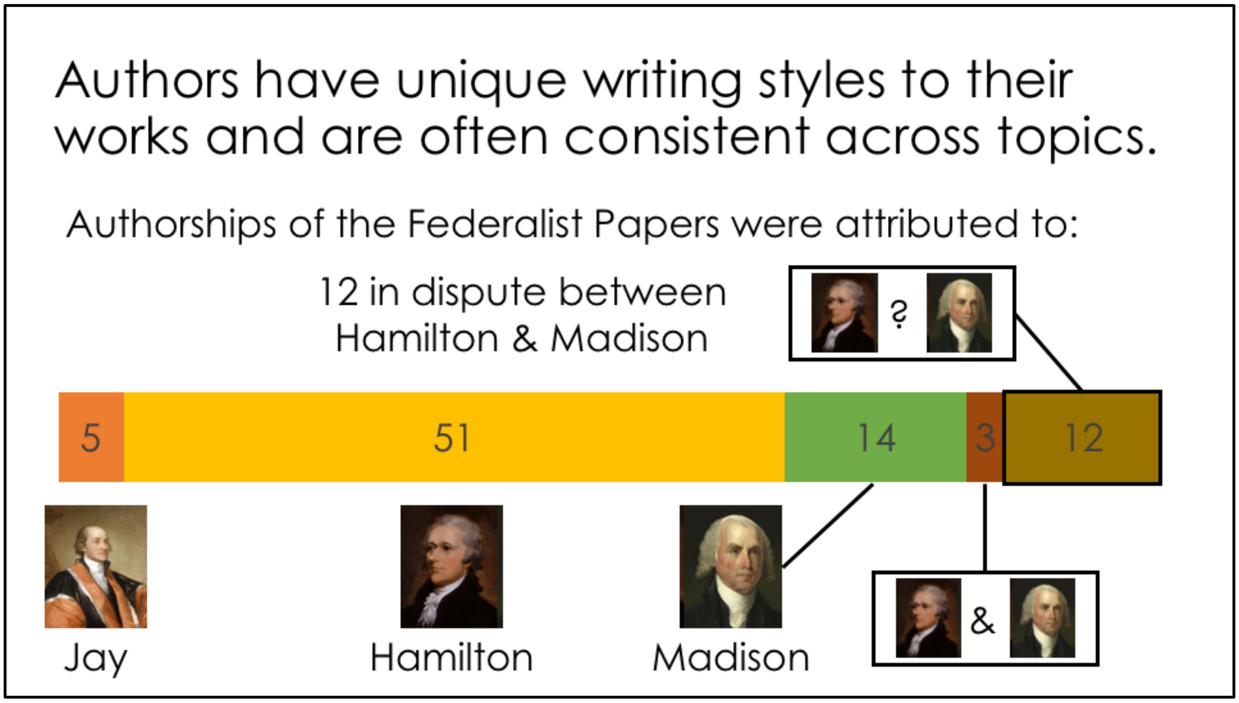
Authors have unique writing styles to their works and are often consistent across topics.

Authorship Identification aka Stylometry

"The best commentary on the principles of government ever written."
—PRESIDENT AND FOUNDING FATHER THOMAS JEFFERSON
"Read it, underline it, and dog-ear it." —SUPREME COURT JUSTICE ANTONIN SCALIA
THE
FEDERALIST
PAPERS
John Jay, James Madison,
& Alexander Hamilton
Foreword by
ALAN DERSHOWITZ

In the late 1780's, **Jay, Madison, and Hamilton** wrote a series of anonymous essays to convince New York voters to ratify the new US Constitution.

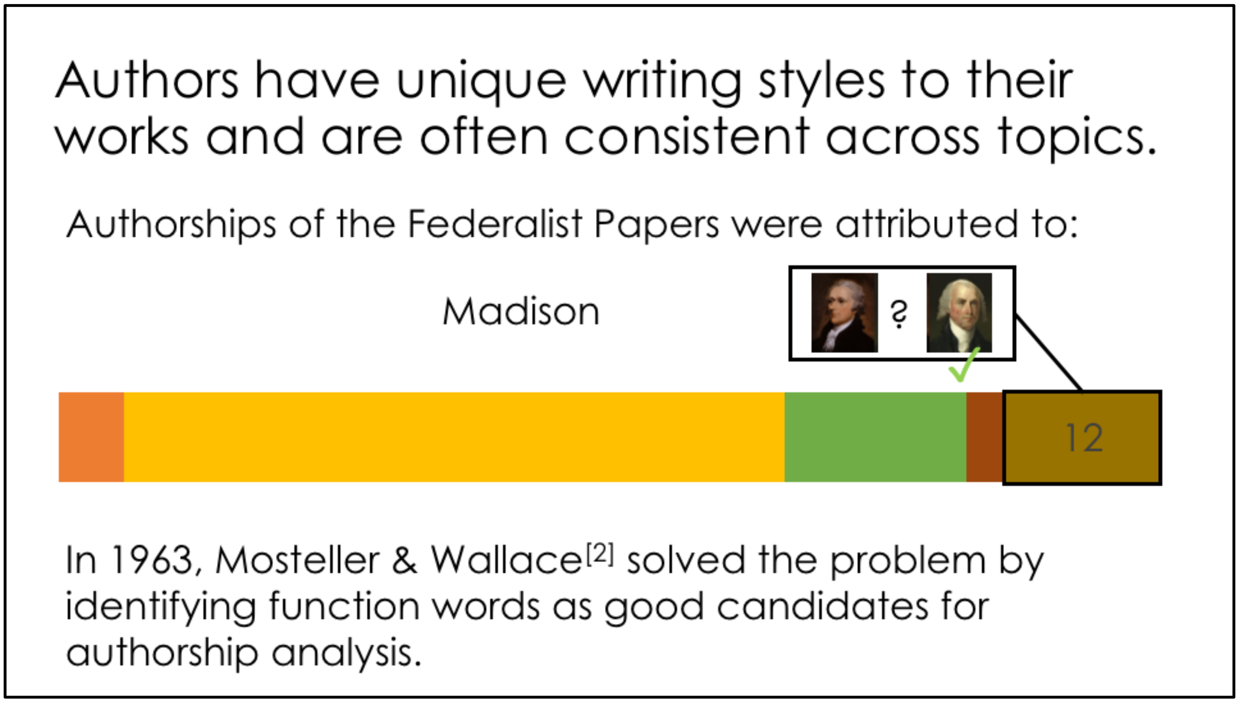
When first learning about authorship identification, we came across the term "stylometry". Authors have unique writing styles to their works and are often consistent across topics. That means, no matter what topic I choose to write about, my writing style can be detected as my unique "identity". Authorship identification has been a topic of interest since the 18th century. In the late 1780's, Jay, Madison, and Hamilton wrote a series of anonymous essays to convince New York voters to ratify the new US Constitution. These are the Federalist papers that I am sure many of you have heard about.



Authors have unique writing styles to their works and are often consistent across topics.

Authorships of the Federalist Papers were attributed to:

Many scholars have identified the exact author of the Federalist Papers. 5 of them were authored by Jay. 52 of them were authored by Hamilton. 14 were authored by Madison. 3 were co-authored by Hamilton and Madison. However, 12 of them were in dispute between Hamilton and Madison. In 1788, it was impossible to tell which of the two was the true author of those 12 mysterious papers.

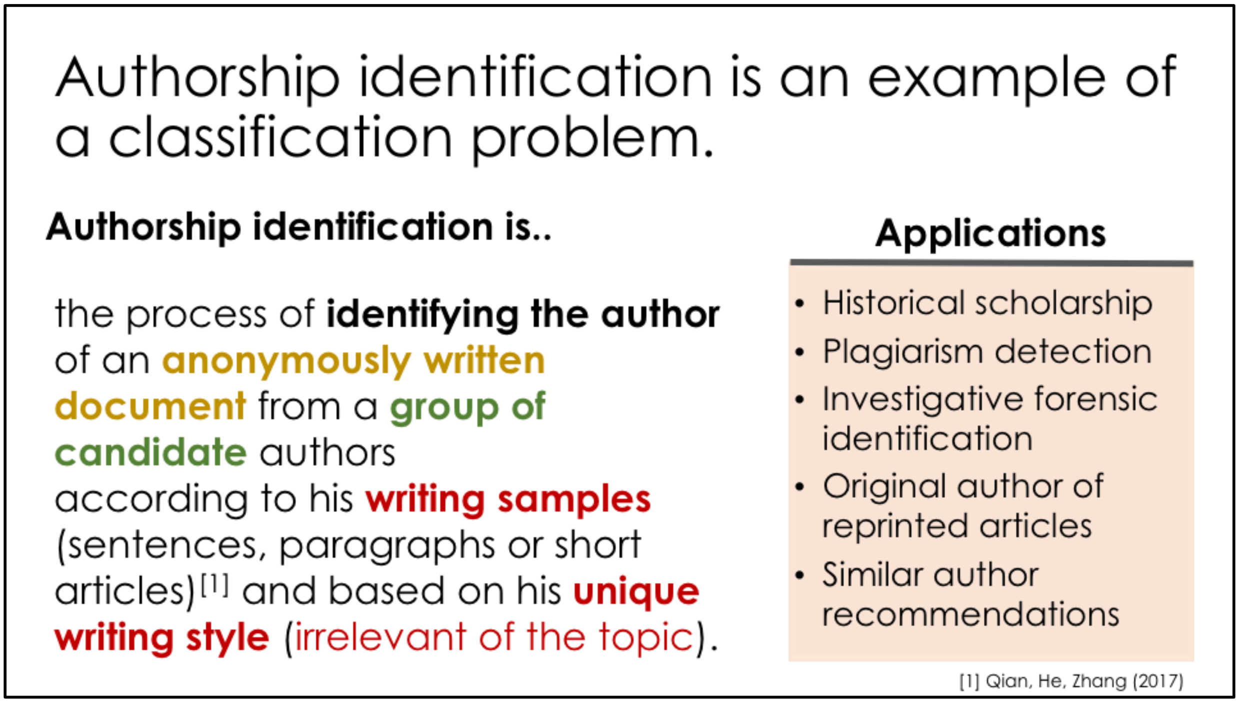


Authors have unique writing styles to their works and are often consistent across topics.

Authorships of the Federalist Papers were attributed to:

In 1963, Mosteller & Wallace[2] solved the problem by identifying function words as good candidates for authorship analysis.

In 1963, Mosteller & Wallace solved the problem by identifying function words as good candidates for authorship analysis. Using these statistical tools, Mosteller & Wallace were able to pinpoint that Madison is the author of the 12 last Federalist Papers. This is only one example of how authorship identification began. This topic is now expanding to many current state-of-the-art real-world applications.



Authorship identification is an example of a classification problem.

Authorship identification is..

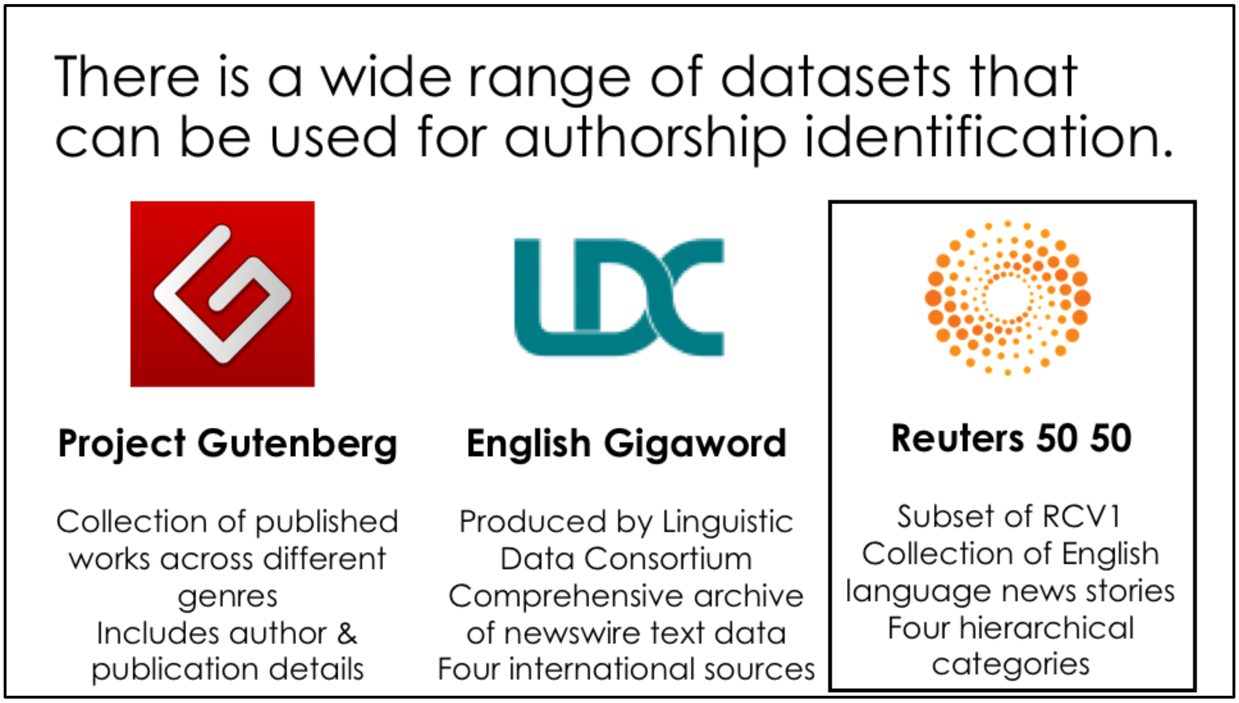
the process of **identifying the author** of an **anonymously written document** from a **group of candidate** authors according to his **writing samples** (sentences, paragraphs or short articles)[1] and based on his **unique writing style** (irrelevant of the topic).

Applications

- Historical scholarship
- Plagiarism detection
- Investigative forensic identification
- Original author of reprinted articles
- Similar author recommendations

[1] Qian, He, Zhang (2017)

As a recap, Authorship identification is essentially an example of a classification problem. It is the process of identifying the author of an anonymously written document from a group of candidate authors according to his writing samples (sentences, paragraphs or short articles)[1] and based on his unique writing style (irrelevant of the topic). Authorship identification can be used for a wide variety of applications including but not limited to historical scholarship, plagiarism detection, investigative forensic identification, determining who the original author of reprinted articles is, recommending authors to readers based on similar authors whose style they enjoy, etc.



There is a wide range of datasets that can be used for authorship identification.

Project Gutenberg

Collection of published works across different genres
Includes author & publication details

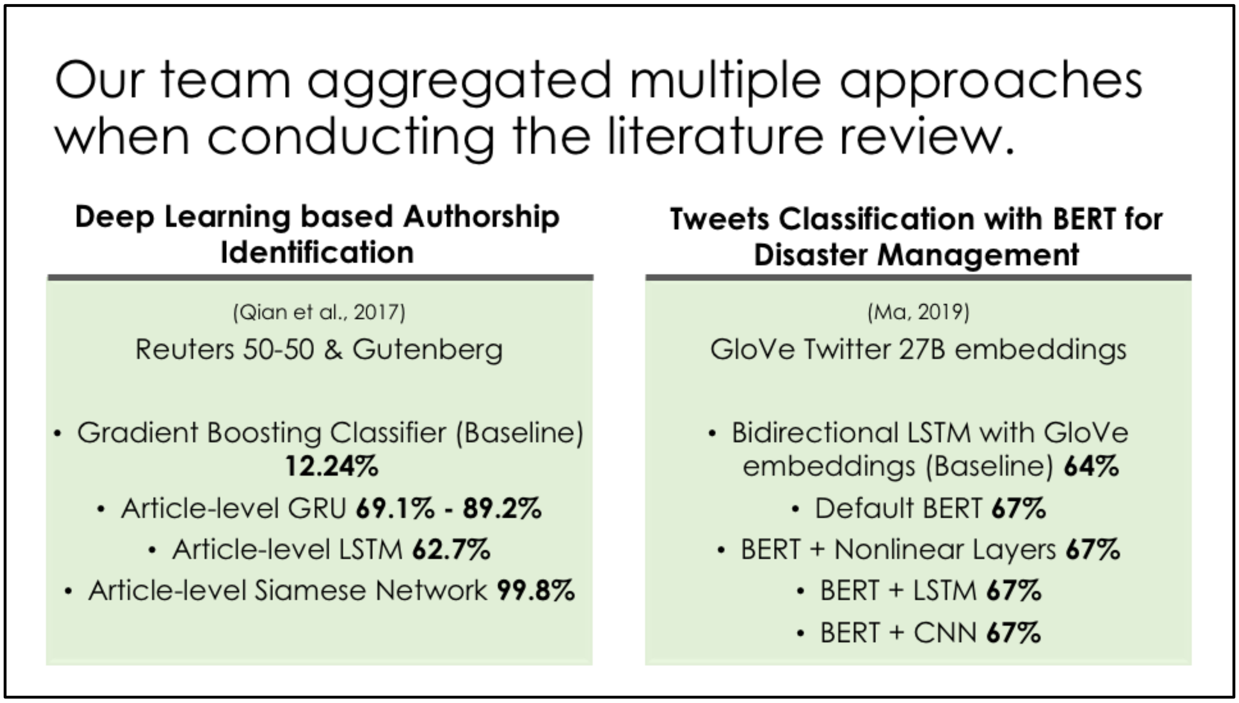
English Gigaword

Produced by Linguistic Data Consortium
Comprehensive archive of newswire text data
Four international sources

Reuters 50 50

Subset of RCV1
Collection of English language news stories
Four hierarchical categories

There is a wide range of datasets that can be used for authorship identification. We identify three common data sources throughout papers on authorship identification. The first is Project Gutenberg. Project Gutenberg is a collection of published works across many different genres. Each plain text entry has a label of the author, the genre and additional publication details. The second is the English gigaword corpus. English gigaword is produced by the Linguistic Data Consortirum or LDC. This is a comprehensive archive of newswire text data from four international sources: Agence France Press English Service(afe), Associated Press Worldstream English Service(apw), The New York Times Newswire Service(nyt), and The Xinhua News Agency English Service(xie). The third is the Reuters 50 50 dataset. This represents a subset of RCV1, a collection of over 800,000 English language news stories dating from August 20, 1996 to August 19, 1997 that have been made available by Reuters, Ltd.. Reuters 5050 is made up of four hierarchical categories: Corporate/Industrial (CCAT) Economics (ECAT) Government/Social (GCAT) and Markets (MCAT). Our team will be using the Reuters 50 50 dataset to train and test our model.



Our team aggregated multiple approaches when conducting the literature review.

Deep Learning based Authorship Identification

(Qian et al., 2017)

Reuters 50-50 & Gutenberg

- Gradient Boosting Classifier (Baseline) **12.24%**
- Article-level GRU **69.1% - 89.2%**
- Article-level LSTM **62.7%**
- Article-level Siamese Network **99.8%**

Tweets Classification with BERT for Disaster Management

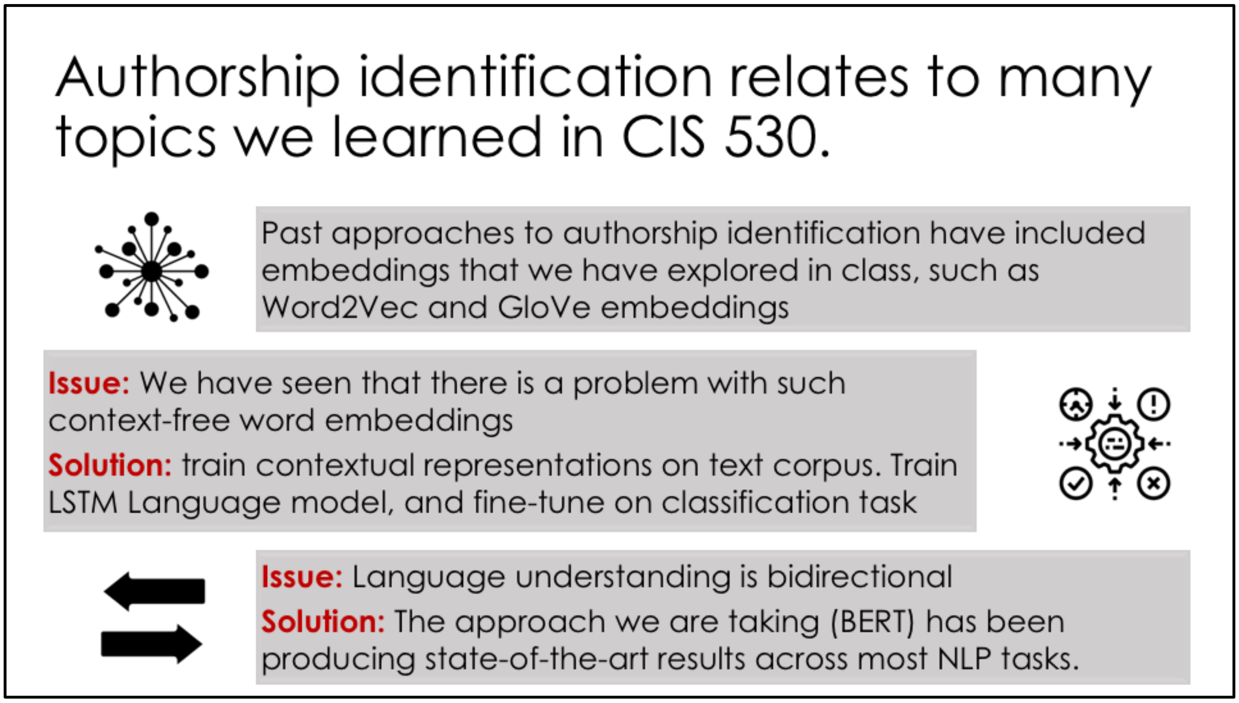
(Ma, 2019)

GloVe Twitter 27B embeddings

- Bidirectional LSTM with GloVe embeddings (Baseline) **64%**
- Default BERT **67%**
- BERT + Nonlinear Layers **67%**
- BERT + LSTM **67%**
- BERT + CNN **67%**

Our team aggregated multiple approaches when conducting the literature review.
In the first paper Deep Learning based Authorship Identification, the datasets used were both the Reuters 50-50 and Project Gutenberg. The three major models were a GRU (Gated Recurrent Unit) network, LSTM (Long Short Term Memory) network, and a Siamese network. The baseline as presented in this paper with just the Gradient Boosting Classifier on the C50 dataset achieved only an accuracy of 12.24%, which they were able to bring up to an impressive 99.8% using an article-level Siamese network on both datasets, used for verification.

In the second paper, which applies deep learning techniques to address Tweets classification problems in the disaster management field, the authors tested different combinations to achieve a marginally better result of 67% accuracy using BERT with a convolutional neural network.



Authorship identification relates to many topics we learned in CIS 530.

Past approaches to authorship identification have included embeddings that we have explored in class, such as Word2Vec and GloVe embeddings

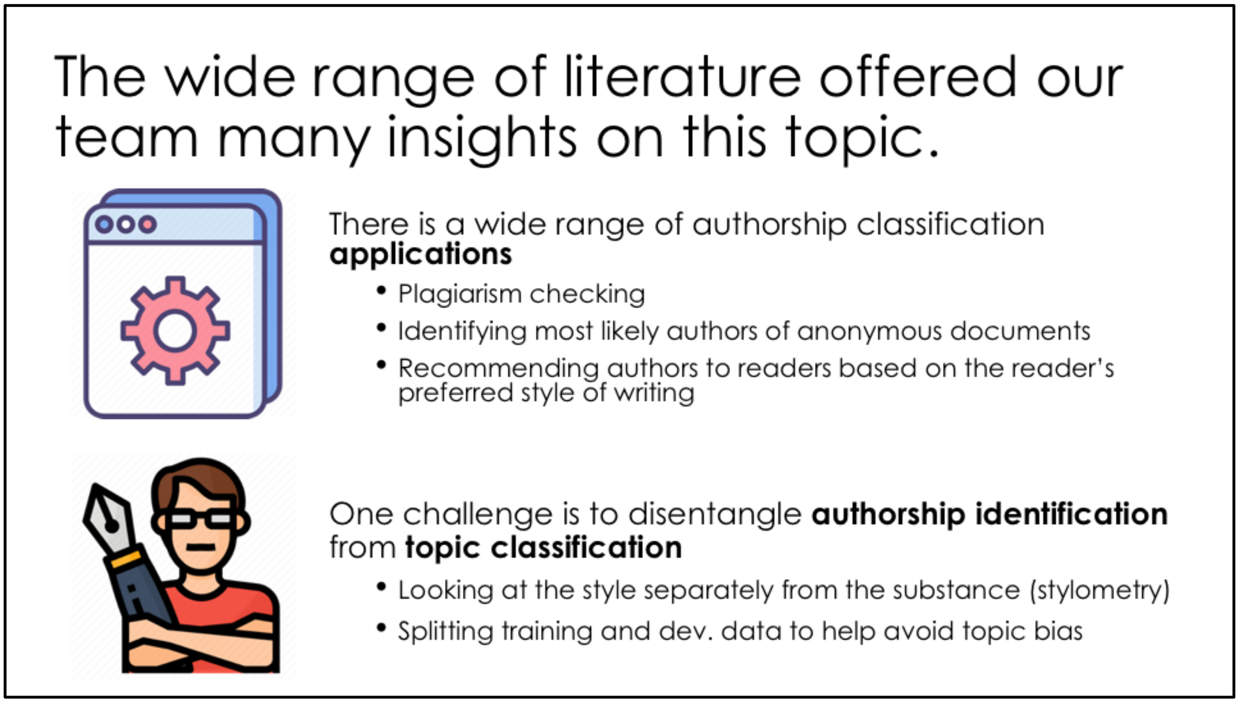
Issue: We have seen that there is a problem with such context-free word embeddings

Solution: train contextual representations on text corpus. Train LSTM Language model, and fine-tune on classification task

Issue: Language understanding is bidirectional

Solution: The approach we are taking (BERT) has been producing state-of-the-art results across most NLP tasks.

Authorship identification relates to many topics we learned in CIS 530.
Past approaches to authorship classifications utilized Word2Vec embeddings or using the bag-of-words model. As mentioned in lectures, we have seen that there is a problem with such context-free word embeddings, which was then solved by training contextual representations on text corpus (such as using LSTM and fine-tuning). With BERT, which can be used as an encoder to produce sentence embeddings, and to produce context-based word embeddings, these issues will be taken care of.



The wide range of literature offered our team many insights on this topic.

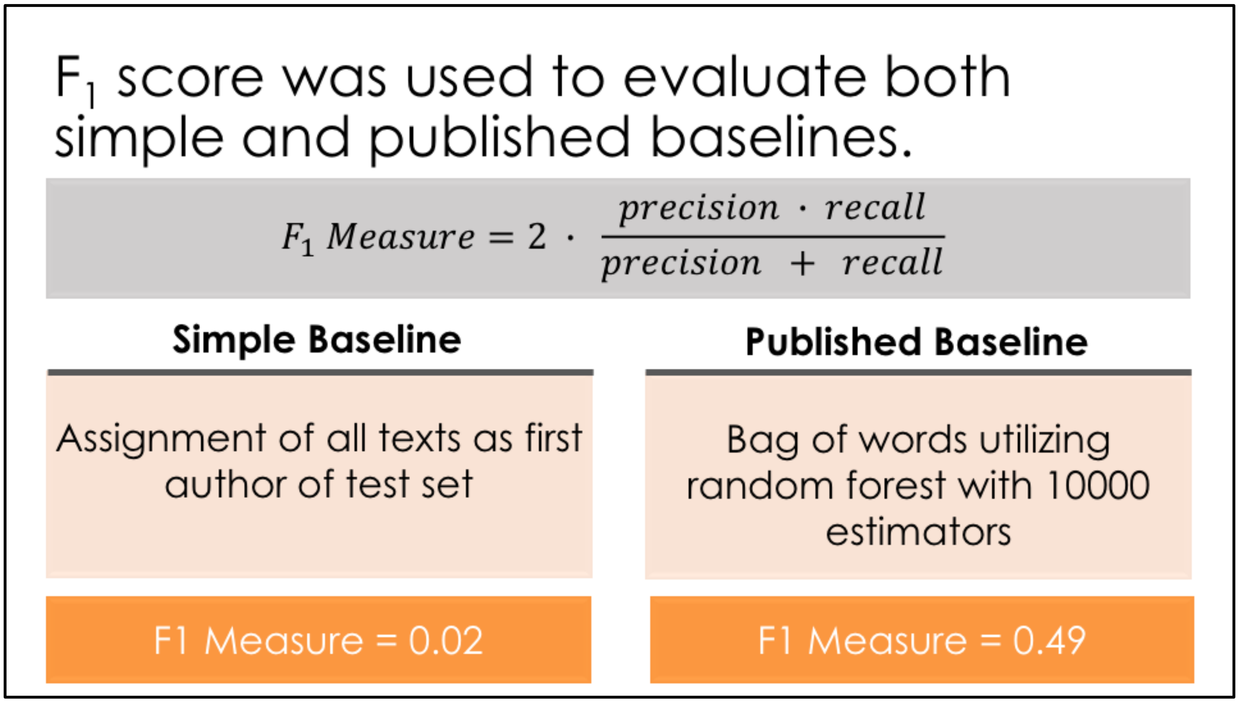
There is a wide range of authorship classification **applications**

- Plagiarism checking
- Identifying most likely authors of anonymous documents
- Recommending authors to readers based on the reader's preferred style of writing

One challenge is to disentangle **authorship identification** from **topic classification**

- Looking at the style separately from the substance (stylometry)
- Splitting training and dev. data to help avoid topic bias

The wide range of literature offered our team many insights on this topic. In particular, we noticed the wide application of a seemingly narrow topic of authorship classification, such as identifying most likely authors of documents, plagiarism checking, and as a new way for recommending authors to readers based on the reader's preferred style of writing.
We also noted that there was a need to make sure to disentangle authorship identification and topic classification while conducting our research.



F_1 score was used to evaluate both simple and published baselines.

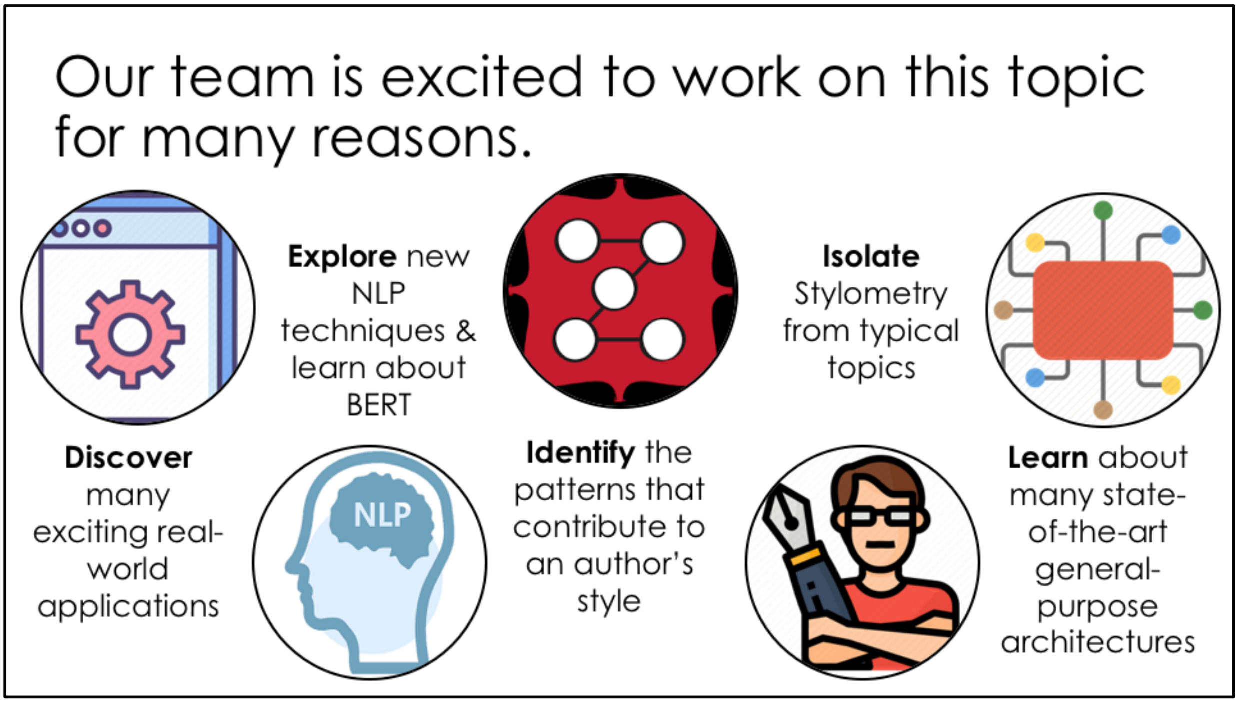
$$F_1\ Measure = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Simple Baseline	Published Baseline
Assignment of all texts as first author of test set	Bag of words utilizing random forest with 10000 estimators
F1 Measure = 0.02	F1 Measure = 0.49

Our evaluation metric is the F1 score. Our simple baseline assigns all the texts in the test set to the first author of the featured on the test set. This yielded the F1 score of 0.02.

We used the baseline described by Garibay et al. 2015 linked: http://ceur-ws.org/Vol-1391/72-CR.pdf. The published baseline is a bag of words model utilizing random forest with 10000 estimators.

Our published baseline performed at F1 of 0.49.



Our team is excited to work on this topic for many reasons.

Explore new NLP techniques & learn about BERT

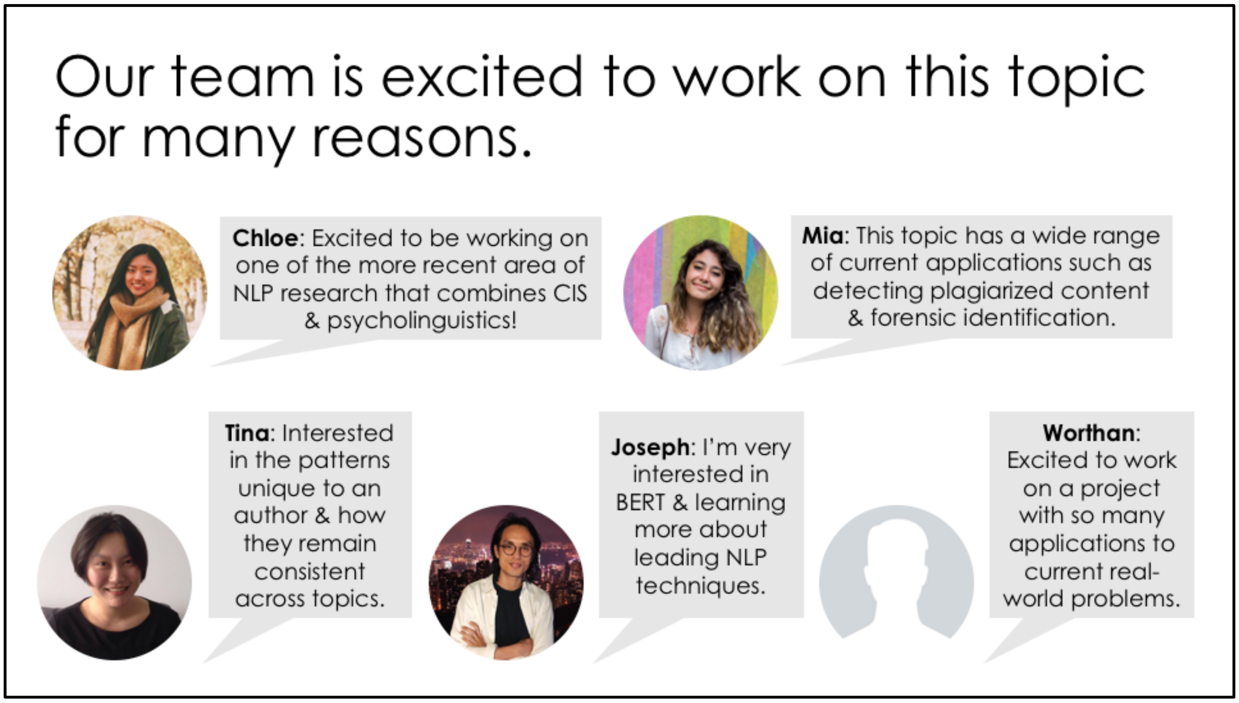
Isolate Stylometry from typical topics

Discover many exciting real-world applications

Identify the patterns that contribute to an author's style

Learn about many state-of-the-art general-purpose architectures

Our team is very excited to work on this topic for many reasons. Authorship identification allows us to discover many exciting real-world applications – as cool as forensic analysis! It will allows us to explore new NLP techniques we haven't yet learned in 530, but also explore BERT even further and learn about many state-of-the-art genera—purpose architecture. The topic in itself is very fascinating: our team is curious to understand what patterns contribute to an author's unique style in writing. It will be interesting to isolate and disentangle stylometry from the typical topcis the author usually addresses.



Our team is excited to work on this topic for many reasons.

Chloe: Excited to be working on one of the more recent area of NLP research that combines CIS & psycholinguistics!

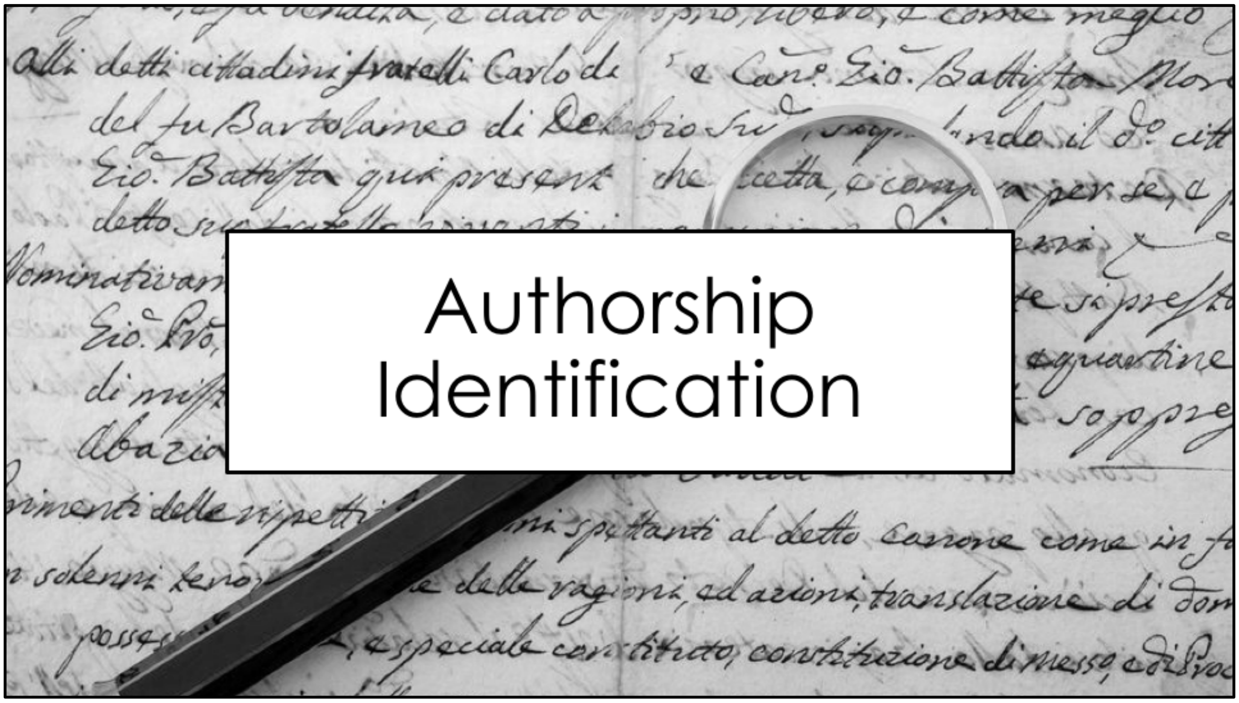
Mia: This topic has a wide range of current applications such as detecting plagiarized content & forensic identification.

Tina: Interested in the patterns unique to an author & how they remain consistent across topics.

Joseph: I'm very interested in BERT & learning more about leading NLP techniques.

Worthan: Excited to work on a project with so many applications to current real-world problems.

We are all equally motivated to be working on this project – and for many reasons. Chloe is excited to be working on one of the more recent area of NLP research that combines CIS & psycholinguistics. Like her, Joseph is very interested in BERT & learning more about leading NLP techniques. Mia and Worthan are motivated by the importance of the applications. They are excited to work on a project with so many applications to current real-world problems. This topic has a wide range of current applications such as detecting plagiarized content & forensic identification. And finally, Tina is really interested in the patterns unique to an author & how these remain consistent across topics. All in all, authorship identification will be a fun project once we start seeing how our model performs for different styles of writing!



Authorship Identification

Our team is looking forward to concluding our authorship identification project. Authorship identification is an important Natural Language Processing (NLP) topic that enables us to identify the author of an anonymously written document based on their unique writing style. Looking forward to diving in interesting literature!