IDENTIFYING • WHO • SAID • A • LINE

RESEARCH QUESTION

PROBLEM

Identifying authorship of text is currently used to combat phishing or unrightful authorship claims, but mostly on large documents

WHY

Identifying authorship of smaller, everyday sentences would benefit chatbots or personal assistents

HOW

Make text readable for machines and apply classification model to this representation, compare their outcome

HAND-CRAFTED

TF-IDF

Count word occurrence

LEARNED

EMBEDDING

Capture context; match words to meaning

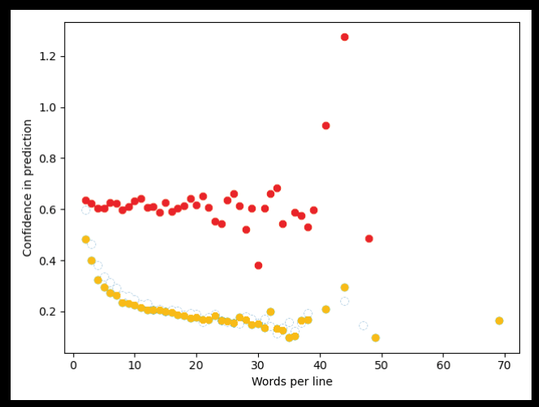
PRELIMINARY RESULTS

F·R·I·E·N·D·S

70,000 lines
6 people
7 word/line average

TECHNIQUE	ACCURACY
GUESSING	17%
TF-IDF	28%
EMBEDDING	23%

TF-IDF's accuracy is highest, but it is also more confident that its mistakes are correct



FUTURE

Understand the reason why the classification model chooses a character

Analyse the effect of different preprocessing techniques

Check the minimum words per sentence needed for achieving good accuracy

Check whether complexity of words has influence on performance

PROJECT INFO

Thomas van Tussenbroek
CSE3000
20/05/2020

David Tax; Arman Naseri Jahfari;
Tom Viering; Stavros Makrodimitris

TU Delft