Applied algorithms for detecting ghost writing in high school assignments

Dept. of Computer Science
University of Copenhagen

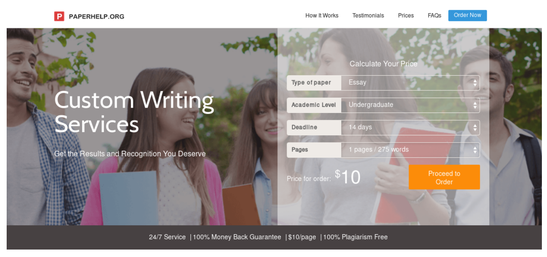
Stephan Lorenzen (lorenzen@diku.dk), Department of Computer Science, University of Copenhagen

Introduction

The Problem:

- Students cheat in high school. In particular, 7% of students in the U.S. admits to have handed in assignments done by others [1].

- Assignments may have been done by other students, or they may have been bought from a *paper-mill* (example right).



- We refer to this problem as *ghost-writing*.

- In Danish high schools, there have been a rise in cases of ghost-writing recently [2]. Normal copy-paste plagiarism software does not work, since the assignments are original work.

- The goal is to combat ghost-writing by automatic writing style analysis, as a warning tool for teachers. Achieving high specificity is important.

The data:

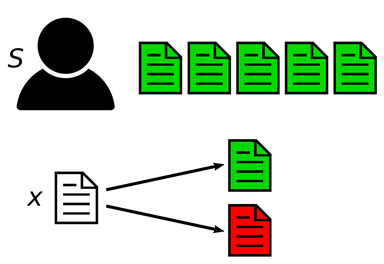
- We cooperate with the company *MaCom*, who provides the Lecture Manegement System *Lectio* to 90% of Danish high schools.

- MaCom has data for more than 150,000 students, with more than 15 million assignments, across all high school subjects.

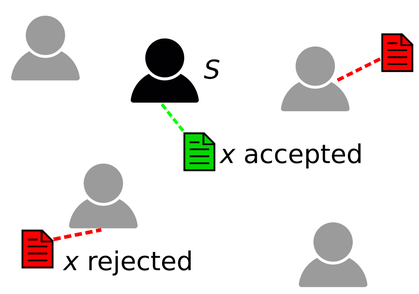
Approach

Basic ideas:

- **Authorship Verification (AV)**: *Use previous assignments handed-in by student S to verify authorship of new assignment x handed in by S*.



- Solve AV by solving **Authorship Attribution (AA)**: *Given n students with previous assignments and a new assignment x, attribute x to one of the n students*. Include student *S* in group, and accept if *x* is attributed to *S*.



Textual features:

- Average word length
- Average sentence length
- Character n-grams
- Word n-grams
- Etc.

Methods:

- Distance based
- Random forests

Preliminary results*

*Results computed in cooperation with MSc student Kenneth Jürgensen

Experiment setup:

- A subset of the full Lectio data is used, consisting of 41567 Danish essays for 3268 students (see table).

	Training	Test
Students	28784	12783
Texts per student	12.7	12.6
Avg. text length	5335	5379

- Each student has a positive and a negative test.

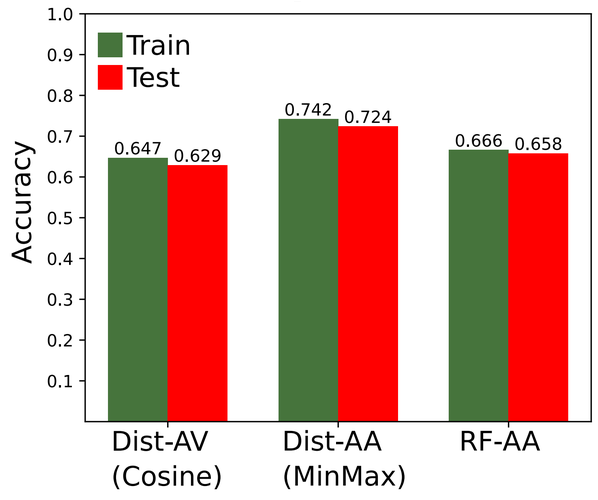
- The data is split into training and test data as shown in the table.

Performance:

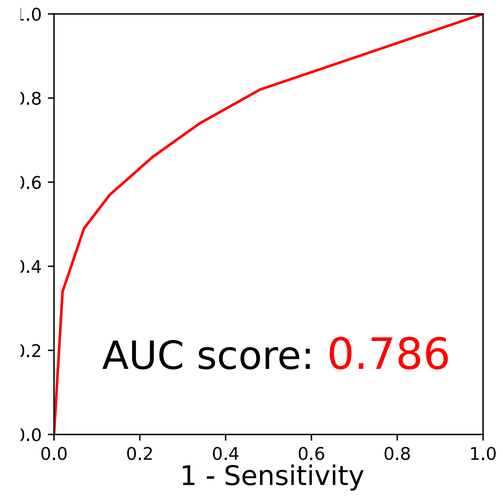
- Distance based Authorship Attribution (Dist-AA) with a limited number of students achieves best result: 72.4%, with specificity 0.822.

- In generel, approaches based on Authorship Attribution outperforms Authorship Verification approaches.

Method performance



ROC (Dist-AA)



Conclusion

- Initial experiments show that determining authorship in the Lectio data is indeed feasible.

- The experiments indicate that methods based on Authorship Attribution perform better.

Future work:

- Include full Lectio data for the Authorship Attribution approaches.

- Tuning methods to other high school subjects, such as math.

- Improving specificity.

References

[1] D.L. McCabe, Cheating among college and university students: A North American perspective. *Int. J. for Educational Integrity*, 1(1), 2005.

[2] DR, Elever bruger ghostwritere til eksamen. *DR.dk* (2016). Available here: www.dr.dk/nyheder/indland/elever-bruger-ghostwritere-til-eksamen