EVALUATING OBFUSCATION ATTACKS USING USER INTERACTION

SERGEY KRAMP

THESIS PROPOSAL
DATA SCIENCE & SOCIETY

STUDENT NUMBER

000000

COMMITTEE

dr. Chris Emmery

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

March 17th, 2023

WORD COUNT

X

EVALUATING OBFUSCATION ATTACKS USING USER INTERACTION

SERGEY KRAMP

1 PROJECT DEFINITION, MOTIVATION & RELEVANCE

Machine Learning and Natural Language Processing (NLP) has been used to collect information about users based on stylometric information (writing style). This could be sensitive information such as age, gender, political affiliation or sexual orientation. A growing filed of research has been concerned with protecting online users and their privacy against data collection methods based on stylometry (**empty citation**), adversarial stylometry. A sub-task of adversarial stylometry is the obfuscation, which involves changing written text to make it harder to be classified by a stylometric classifier while preserving its original meaning. Emmery et al. (**emmer2019obfuscation**) have proposed an obfuscation attack setup against a gender classifier. A part of that setup is a machine learning model that provides substitution suggestions for words that are likely to be informative for the classifier. In this work I will propose an external evaluation metric for the performance of this model that produces this substitutes. In particular by letting users interact with the suggestions and measuring the quality of the suggestions based on the user's feedback. The goal is to come up with an intuitive, model-agnostic metric that can be used to compare different

2 LITERATURE REVIEW

Provide a summary of what is known in the scientific literature about this problem. This summary should be based on at least five relevant **recent** sources and, if appropriate, some more classical sources. These recent sources need to satisfy the following requirements:

1. Recency – published in the last five years

2. Quality – published in scientific peer-reviewed journals or conference proceedings
3. Usefulness – they should help you frame the theoretical background of your project

Note that a complete literature review is not expected at this project stage, but the final report will expect it. Pay good attention to use **paraphrasing** instead of copying text; because of the limited space in this proposal, you should practice with summarising literature in your own words to avoid (accidental) **plagiarism**

To cite papers, copy paste BibTeX code[1] and put it in `references.bib`. After, you can cite some work (MacKay & Mac Kay, 2003) – using `\citep`. You can refer to the author of e.g. Minsky (1961) directly like using `\citet` If you just want to print the author names at the start of a sentence separate from the citation, you might want use `\citeauthor` when citing, like: In their seminal contribution, Ananny and Crawford provide evidence for ... (Ananny & Crawford, 2018). If you want to add pages you can use brackets in `\citep[][p. 5]{mackay2003information}`, which looks like: (MacKay & Mac Kay, 2003, p. 5). The first brackets can be used for *see*, and *e.g.* etc. If you want to cite multiple authors, simply comma-separate them (`\citep{minsky1961steps,mackay2003information}`) and it will aggregate them automatically (MacKay & Mac Kay, 2003; Minsky, 1961).

3 RESEARCH STRATEGY & RESEARCH QUESTIONS

Outline the concrete research strategy for the project, formulated as Research Questions that the thesis project will answer. What will be contributed to the literature by answering these research questions? Avoid very general statements ("is it feasible to ...") but try to formulate concrete research questions, split into sub-questions where appropriate. The RQs should follow logically from the problem statement combined with state-of-the-art to inform your research strategy.

Your strategy should address these elements that also appear in the evaluation rubric for the final thesis product (if already known at this point):

- Does the dataset/your target variable contain large class imbalances/non-normal distributions?
- Are there separable feature sets in the dataset(s)? Will additional features be generated? Which methods will be used for feature selection/ranking or model interpretability?

[1] Using e.g. the quote icon in GScholar, then BibTeX at the bottom.

- On which aspect will model comparison be implemented? Comparison of several tuned algorithms, comparison of different input datasets? What is the proposed method for validation and test set separation? Will you use resampling or other statistical methods to assess model comparisons?
- How will error patterns be analyzed? Are there interesting subgroups in the data for which bias/subgroup error analysis could be implemented?

(Sub-)RQs should specify which manipulations to the data, features, and/or algorithms are contrasted and should be specific rather than general statements (e.g., name the algorithms you are considering, instead of posing general questions like "which Machine Learning model..."). You can write a short motivation leading up to a subRQ, for example: "previous research has shown that a larger proportion of man failed in X compared to women. Therefore, model performance and error analysis will also be split according to gender".

4 METHODOLOGY AND EVALUATION

4.1 *Dataset Description*

Describe the dataset(s) you will use in your project (size, format, accessibility). Provide a rationale for why you are choosing these data. **If, at the point of proposal submission, you do not yet have your complete dataset (e.g., in a project with an external partner), there is a very real risk the project might not be completed in time. Prepare a plan B with your supervisor.**

4.2 *Algorithms and Software*

Describe what algorithms and software you plan to use in your project. Include a motivation for why you have chosen these implementations, with references to the literature.

4.3 *Evaluation Method*

Define how you will evaluate your results. For prediction problems (classification or regression), you will likely use standard techniques – they do not need to be explained in detail. How will you be able to judge the performance of competing models? Against what baseline methods(s) will

you compare your algorithm(s)? How do you plan to obtain ground-truth labeled data to measure accuracy, precision, recall, or some other metric? If you plan to use unsupervised techniques, provide information on how the clustering algorithm will be tested and how the model comparison will be implemented. Details for a cross-validation strategy or other out-of-sample evaluation should be included.

5 MILESTONES AND PLAN

Sketch out what you think will be the major intermediate milestones you need to achieve. Give a general idea of your planning.

REFERENCES

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

MacKay, D. J., & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*(1), 8–30.