School of Computing and Information Systems
The University of Melbourne
COMP90049, Introduction to Machine Learning, Semester 2 2022

Assignment 3: Toxicity Classification in Online Comments

Released: Friday, September 9th 2022.

Due: **Stage I**: Friday, October 7th 5pm
Stage II: Wednesday, October 12th 5pm

Marks: The Project will be marked out of 30, and will contribute 30% of your total mark.

1 Overview

Nowadays, having discussion online is very common. However, due to the harassment and abuse online, some people may express their opinions in an offensive, rude or disrespectful manner and post **toxic comments**, which are defined as *comments that contain obscene, identity attack, insult, threat or otherwise likely to make people uncomfortable*. In this assignment, you will develop and critically analyse models for classifying the **toxicity of comments**. That is, given a comment, your model(s) will classify whether the comment is *toxic* or not. You will be provided with a data set of comments that have been labelled with their toxicity. In addition, each comment is labelled with identities mentioned in the comment. There are 24 annotated identities, which can be grouped into 5 different categories (see Sec. 3 for more details). You may use this additional identity information to investigate whether your models work equally well for comments mentioning different identities. The assessment provides you with an opportunity to reflect on concepts in machine learning in the context of an open-ended research problem, and to strengthen your skills in data analysis and problem solving.

The goal of this assignment is to **critically assess** the effectiveness of various Machine Learning algorithms on the problem of determining a comment's toxicity, and to **express the knowledge that you have gained in a technical report**. The technical side of this project will involve applying appropriate machine learning algorithms to the data to solve the task. There will be a Kaggle in-class competition where you can compare the performance of your algorithms against your classmates.

The focus of the project will be the report, formatted as a short research paper. In the report, you will demonstrate the knowledge that you have gained, in a manner that is accessible to a reasonably informed reader.

2 Deliverables

Stage I: Model development and testing and report writing (due October 7th 5pm):

1. One or more programs, written in Python, including all the code necessary to reproduce the results in your report (including model implementation, label prediction, and evaluation). You should also include

a README file that briefly details your implementation. *Submitted through Canvas.*

2. An anonymous written report, of 2000 words ($\pm 10\%$) **excluding** reference list. Your name and student ID should **not** appear anywhere in the report, including the metadata (filename, etc.). *Submitted through Canvas/Turnitin.*
3. Predictions for the toxicity of comments submitted to the Kaggle[1] in-class competition described in Sec. 6.

Stage II: Peer reviews (due October 12th 5pm):

1. Reviews of two reports written by your classmates, of 200-400 words each. *Submitted through Canvas.*

3 Data Sets

You will be provided with a *training* set of comments, labeled with a toxicity (target label) and identity labels a *development* set with the same labels which you should use for model selection and tuning; a *test* set with no target (but identity) labels, which will be used for final evaluation in the Kaggle in-class competition; and an *unlabelled* data set providing additional comments with no labels at all, which you may use for semi- or unsupervised learning approaches.

Data format All data sets are provided as "csv" files. Each row in the csv files corresponds to one instance. It contains the comment representation (raw text data or extracted features), its target toxicity label (train and dev only) and its identity labels (train, dev and test only).

Target Labels

These are the labels that your model should predict (y). In the provided data set, each comment is labelled with one of two possible toxicity values:

- 1, which denotes the comment is toxic.
- 0: which denotes the comment is non-toxic.

Identity Labels

There are 24 identities from 5 different categories annotated in each comment. The details of the identity labels are shown in Tab. 1. The identity labels provide additional identity information mentioned in the comment. They should *only* be used to evaluate models on specific subgroups of comments, but *not* be predicted (and probably not used as features, although you can discuss this in your report). Each of the 24 identity labels is labelled with one of the two possible values:

- 1, which denotes presence of the identity in the comment.
- 0, which denotes that the comment does not mention this identity

[1] https://www.kaggle.com/

Table 1: Annotated Identities in the Comments

Category	Identities
Religion	Atheist, Buddhist, Christian, Hindu, Jewish, Muslim, Other religion
Race or ethnicty	Asian, Black, White, Latino, Other race or ethnicity
Sexual Oreintation	Bisexual, Hetrosexual, Homosexual gay or lesbian, Other sexual orientation
Gender	Male, Female, Transgender, Other gender
Disability	Physical disability, Intellectual or learning disability, Psychiatric or mental illness, Other disability

3.1 Features

To aid in your initial experiments, we have created different **feature representations** from the raw comments. You may use any subset of the representations described below in your experiments, and you may also engineer your own features from the raw comments if you wish. The provided representations are

1. Raw The raw comments represented as a single string, e.g.,

"Read the whole article.... nowhere does it mention Vera Katz..... woman mayor of portland..... "

2. TFIDF We applied term frequency - inverse document frequency pre-processing to the comments for feature selection. In particular, we (1) removed all stopwords and (2) only retained the 1000 words in the full raw comment data set with highest TFIDF values. As a result, each comment is now represented as a 1000 dimensional feature vector, each dimension corresponding to one of the 1000 words. The value is 0 if the word did *not* occur in the comment, and the word's TFIDF score if the word occurs in the comment. Note that most values will be 0.0 as comments are short. E.g.,

[0.0, 0.0, 0.0, 0.0, . . . 0.998, . . . 0.0]

A 1000-dimensional list of numbers — TFIDF score of word in comment — Word not in comment

You can learn more about TFIDF in [4]. The file `tfidf_words.txt` contains the 1000 words with highest TFIDF value, as well as their index in the vector representations. You may use this information for model/error analysis.

3. Embedding We mapped each comment to a 384-dimensional embedding computed with a pre-trained language model, called the Sentence Transformer [3][2]. These vectors capture the "meaning" of each comment so that similar comments will be located closely together in the 384-dimensional space. E.g.,

[0.037218336 0.000965205 -0.05178352 0.016464263 0.021155884 0.05699353 0.019121459 . . .]

a 384-dimensional list of numbers

[2]https://pypi.org/project/sentence-transformers/

4 Tasks

Stage I

You should formulate a research question (two examples provided below), and develop machine learning algorithms and appropriate evaluation strategies to address the research question.

You should *minimally* implement and analyse in your report one baseline, and at least two different machine learning models. **N.B.** We are more interested in your *critical analysis* of methods and results, than the *raw performance* of your models. You may not be able to solve the research question, which is perfectly fine, but you should analyse and discuss your (possibly negative) results in depth.

1. Research Question

You should address **one** research question in your project. We propose two research questions below, for inspiration but you may propose your own. Your report should clearly state your research question. Addressing more than one research question does **not** lead to higher marks. We are more interested in your *critical analysis* of methods and results, than the coverage of more content or materials.

Research Question 1: Does Unlabelled data improve comment toxicity classification?

Various machine learning techniques have been (or will be) discussed in this subject (Naive Bayes, 0-R, clustering, semi-supervised learning); many more exist. These algorithms require different levels of supervisions: some are supervised, some unsupervised and some combine both strategies. Develop machine learning models that leverage different amounts of supervision, using the `train` and `unsupervised` data sets. You may also want to experiment with different feature representations (Sec 3.1). You are strongly encouraged to make use of machine learning software and/or existing libraries in your attempts at this project (such as `sklearn` or `scipy`). What are the strengths and weaknesses of the different machine learning paradigms? Can you effectively combine labelled and unlabelled training data?

Research Question 2: Exploring Bias in comment toxicity classification

Use the following five identities: **"Christian"**, **"Muslin"**, **"Female"**, **"Homosexual gay or lesbian"** and **"Male"** to separately select five subsets of comments from the development set, each subset of comments mentioning one of the five identities. Compare different models and/or feature representations in their performance on the five subsets of comments separately. Can your models work equally well for all groups? Critically analyse the gap, and try to explain it in the context of the concepts covered in this subject. Can you adapt your models to close the performance gap? How? *Note:* your grade does not depend on your success in closing the performance gap. Interesting, failed attempts with in-depth analyses are perfectly acceptable.

2. Feature Engineering (optional)

We have discussed three types of attributes in this subject: categorical, ordinal, and numerical. All three types can be constructed for the given data. Some machine learning architectures prefer numerical attributes (e.g. k-NN); some work better with categorical attributes (e.g. multivariate Naive Bayes) – you will probably observe this through your experiments.

It is **optional** for you to engineer some attributes based on the `raw` comment dataset (and possibly use them instead of – or along with – the feature representations provided by us). Or, you may simply select features from the ones we generated for you (tfidf, and embedding).

3. Evaluation

The objective of your learners will be to predict the classes of unseen data. We will use a **holdout strategy**: the data collection has been split into three parts: a training set, a development set, and a test set. This data will be available on the LMS. We will use Accuracy as the evaluation metric.

To give you the possibility of evaluating your models on the test set, we will be setting up a **Kaggle In-Class competition**. You can submit results on the test set there, and get immediate feedback on your system's performance. There is a Leaderboard, that will allow you to see how well you are doing as compared to other classmates participating on-line.

4. Report

You will submit an **anonymised** report of 2000 words in length ($\pm 10\%$), **excluding** reference list. The report should follow the structure of a short research paper, as discussed in the guest lecture on Academic Writing. It should describe your approach and observations, both in engineering (optional) features, and the machine learning algorithms you tried. Its main aim is to provide the reader with **knowledge** about the problem, in particular, **critical analysis of your results and discoveries**. The internal structure of well-known machine learning models should only be discussed if it is important for connecting the theory to your practical observations.

- Introduction: a short description of the problem and data set, and the research question addressed
- Literature review: a short summary of some related literature, including at least two relevant research papers of your choice. Other options include [6], [5] and [2], which are provided in the Reference list of this document. You are encouraged to search for other references, for example among the articles cited within the papers referenced in this document.
- Method: Identify the newly engineered feature(s), and the rationale behind including them (Optional). Explain the methods and evaluation metric(s) you have used (and why you have used them)
- Results: Present the results, in terms of evaluation metric(s) and, ideally, illustrative examples
- Discussion / Critical Analysis: Contextualise** the system's behavior, based on the understanding from the subject materials as well as in the context of the research question.
- Conclusion: Clearly demonstrate your identified knowledge about the problem
- A bibliography, which includes [1], as well as references to any other related work you used in your project. You can use any citation style, *as long as you are consistent* throughout your report.

** Contextualise implies that we are more interested in seeing evidence of you having thought about the task and determined reasons for the relative performance of different methods, rather than the raw scores of the different methods you select. This is not to say that you should ignore the relative performance of different runs over the data, but rather that you should think beyond simple numbers to the reasons that underlie them.

We will provide LaTeXand RTF style files that we would prefer that you use in writing the report. Reports are to be submitted in the form of a **single PDF file**. If a report is submitted in any format other than PDF, we reserve the right to return the report with a mark of 0.

Your name and student ID should **not** appear anywhere in the report, including any metadata (filename, etc.).

If we find any such information, we reserve the right to return the report with a mark of 0.

Stage II

During the reviewing process, you will read two anonymous submissions by your classmates. This is to help you contemplate some other ways of approaching the Project, and to ensure that every student receives some extra feedback. You should aim to write 150-300 words total per review, responding to three *'questions'*:

- Briefly summarise what the author has done in one paragraph (50-100 words)
- Indicate what you think that the author has done well, and why in one paragraph (50-100 words)
- Indicate what you think could have been improved, and why in one paragraph (50-100 words)

5 Assessment Criteria

The Project will be marked out of 30, and is worth 30% of your overall mark for the subject. The mark breakdown will be:

Report Quality: (26/30 marks)
You can consult the marking rubric on the Canvas/Assignment 3 page which indicates in detailed categories what we will be looking for in the report.

Kaggle: (2/30 marks)
For submitting (at least) one set of model predictions to the Kaggle competition.

Reviews: (2/30 marks)
You will write a review for each of two reports written by other students; you will follow the guidelines stated above.

6 Using Kaggle

The Kaggle in-class competition URL will be announced on LMS shortly. To participate do the following:

- Each student should create a Kaggle account (unless they have one already) using your Student-ID
- You may make up to 8 submissions per day. An example submission file can be found on the Kaggle site.
- Submissions will be evaluated by Kaggle for accuracy, against just 30% of the test data, forming the public leaderboard.
- Prior to competition close, you may select a final submission out of the ones submitted previously – by default the submission with highest public leaderboard score is selected by Kaggle.
- After competition close, public 30% test scores will be replaced with the private leaderboard 100% test scores.

7 Assignment Policies

7.1 Terms of Data Use

The data set is derived from the resource [1]:

> Jigsaw/Conversation AI. Jigsaw Unintended Bias in Toxicity Classification: https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification.

The original target and identity labels of the dataset are numerical values from 0 to 1. We use 0.5 as the threshold to transform all values to either 0 or 1 for this project. More details of the dataset can be found from the above link. In your report you need to briefly describe the datset and this reference **must** be cited and put in the bibliography. We reserve the right mark of any submission lacking this reference with a 0, due to violation of the Terms of Use.

Please note that the dataset is a sample of actual data posted to the World Wide Web. As such, it may contain information that may be considered profane, vulgar, or offensive for toxicity classification. We would ask you, as much as possible, to look beyond this to the task at hand.

Changes/Updates to the Project Specifications

We will use Canvas announcements for any large-scale changes (hopefully none!) and Piazza for small clarifications. Any addendums made to the Project specifications via the Canvas will supersede information contained in this version of the specifications.

Late Submission Policy

There will be **no late submissions** allowed to ensure a smooth peer review process. Submission will close at **5pm on October 7th**. For students who are demonstrably unable to submit a full solution in time, we may offer an extension, but note that you may be unable to benefit from the peer review process in that case. A solution will be sought on a case-by-case basis. Please email Hasti Samadi with documentation of the reasons for the delay.

Academic Misconduct

For most people, collaboration will form a natural part of the undertaking of this project. However, it is still an individual task, and so reuse of ideas or excessive influence in algorithm choice and development will be considered cheating. We highly recommend to (re)take the academic honesty training module in this subject's Canvas. We will be checking submissions for originality and will invoke the University's Academic Misconduct policy[3] where inappropriate levels of collusion or plagiarism are deemed to have taken place.

[3] http://academichonesty.unimelb.edu.au/policy.html

References

[1] Jigsaw/Conversation AI. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification. Accessed: July, 2022.

[2] Pallam Ravi, Greeshma S Hari Narayana Batta, and Shaik Yaseen. Toxic comment classification. *International Journal of Trend in Scientific Research and Development (IJTSRD)*, 2019.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[4] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[5] Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*, 2018.

[6] Sara Zaheri, Jeff Leath, and David Stroud. Toxic comment classification. *SMU Data Science Review*, 3(1):13, 2020.