CS4248 Natural Language Processing Assignment 2: Solving fact checking

Distributed on 14 Feb 2025

Due in Canvas `Assignments` by 13 Mar 2025 11:59 PM SGT

This assignment contributes 10 percent of your final mark for the class, and is graded out of a rubric of 100 points.

Integrity Note. *All three assignments in CS4248, including this Assignment 2 (A2), are strictly **individual** assignments. These assignments are all crafted such that you, as an individual student, can accomplish the assignment on your own, without needing to consult others or AI tools. Explicitly, this means:*

- ***NO** discussion with fellow students, peers, seniors, or alumni about any part of the assignment.*
- ***NO** use or consultation of AI tools on the assignment. This includes coding utilities such as GitHub Copilot, Cursor, etc.*

*Inevitably, as AI-enhanced tools are being added to many software packages, you may find it difficult to avoid the implicit use of AI. Please ask us in the **Assignment 2** discussion forum on Canvas – public discussion on the forum is sanctioned, as long as you do not give away answers. If you are unsure, ask your tutorial leader or contact us directly via email. While we encourage students to learn from one another, there are appropriate times for such collaboration. However, these assignments are **not** such an appropriate time. We have made these guidelines **explicit** (no sharing/consultation/discussion) to remove any ambiguity about what is acceptable.*

Since this assignment is a common Natural Language Processing task and uses public datasets, there are (undoubtedly) solutions posted somewhere. Under the NUS Code of Conduct, you must follow class policy in working on this individual assignment. When in doubt of whether an action would constitute a violation of policy, please ask teaching assistants or either Min or Chris by email, **before** *attempting the action.*

Failure to follow these guidelines may result in prosecution that can take months to resolve and may result in significant penalties to your academic career.

1 Instruction

The task for Assignment 2 is to attempt to solve a fact-checking task. In this fact-checking task, you need to classify whether a text is an important fact (labelled 1), an unimportant statement (labelled 0), or a non-factual statement (labelled -1).

1. *Sign-in or Sign-up to Kaggle using your **NUSNET email** If you have an existing Kaggle account under a different name, please create a new account using your NUSNET email.*
2. *Go to `https://www.kaggle.com/competitions/cs-4248-fact-checking-2420/` and join the competition by this link `https://www.kaggle.com/t/e141b2a9e63b4292b81f1bee468843ff`.*
3. *Change your team name to your student number.*
4. *Read the description, understand the data, and the evaluation metric.*
5. *Start coding and have fun :)*

*Please be aware that due to Kaggle's restriction, you can only submit up to **20 submissions** per day.*

You can use the optional skeleton code in Canvas Files for this assignment. Note, the skeleton code is optional, you may or may not use it as you like. It is also okay if you want to submit the code with `.ipynb` format.

2 Submission

*You need to submit **a single .zip file**, named using your full name and NUSNET ID in the following format: **A2_YourName_YourNUSNETID.zip** (e.g.* `A2_BobSmith_e01234567.zip`*), which contains your assignment write-up (up to 4 pages in total, one column, please do not use too small fonts, no smaller than 9 pts) and the code that can reproduce your final (highest scoring) Kaggle submission. We encourage to just write the key description within 2-3 pages in your write-up. All files should be stored within a directory named for your student number. Your report should be named* `A2_YourName_YourNUSNETID_writeup.pdf` *and your code should be in Python and the main file should be named* `A2_YourName_YourNUSNETID_assignment2.py`*. Please use Python 3.11.x and explain your code dependencies (including the library version) and the instruction to run it in your report. Both files are to be uploaded to Canvas* `Assignments` *before 13 Mar 2025 11:59 PM SGT.*

3 Grading criteria

1. *Model [20%]*
   You should explain the reason behind the choice of the machine learning algorithm and hyper-parameters that are used to train your model. You do not need to explain how the algorithm works, but should explain any deviations from any standard model that you used or re-implemented. Please do use the algorithms we learned in this module. That is, you are limited to Naive Bayes, Logistic Regression, and simple Neural Networks. The advanced deep learning models (e.g. Transformer, RNN) are not allowed for a fair comparison. You need to try and compare more than a single model, either varying in the choice of algorithm, features tried, or hyper-parameter values. Then in the report, you need to explain why some choices worked better than others through an explanation (not just presentation).

2. *Pre-processing [10%]*
   Explain all pre-processings that you have tried (including those that do not work), and how they affect the model. Also, you can use any library to do pre-processing, but, again, please use the classifier we suggested above.

3. *Feature Engineering [30%]*
   Explain all feature engineering that you have tried (including those that do not work), and how they affect the model. You can use pre-trained word vectors for feature extraction like Stanford GloVe, or NLTK for some other features. But, you are not allowed to fine-tune a pre-trained classifier like BERT (it is perfectly fine if you don't know what this means yet). It may also be helpful to encode linguistic structures as features, given the small size of the dataset — good prior information may help.

4. *Analysis [30%]*
   Make an ablation study (analysis of how each part of your model contributes to the final score), or make a baseline model for empirical comparison. There are two types of analysis expected in the report, first the comparison between different types of the same kind, second is the overall effect of each part in the final model. The first is supposed to be explained in each section, while the second is in the analysis section. You do not have to do the analysis for all possible combinations of features and pre-processing step (since it will be exponential). You can do the analysis either incrementally, grouped into subsets, or a combination of both. Prioritize the analysis on the crucial features and pre-processing step. What is important is that we can see your critical thought process in developing the model, and your decision is based on clear evidence.

5. *Clarity [10%]*
   The report should be written in a clear manner and contains all of the necessary information to document and replicate your work.

At the bottom of your report, you must include the text of the two statements below in your submitted work and digitally sign your homework using your Student Number (starting with `A...`*; N.B., not your NUSNET email identifier). Make sure you have attached this statement to your submission either in written or typed form.*

1A. Declaration of Original Work. *By entering my Student ID below, I certify that I completed my assignment independently of all others (except where sanctioned during in-class sessions), obeying the class policy outlined in the introductory lecture. In particular, I did not discuss the problems and my solutions in this assignment with any student, alumnus, or AI tool. The work I submit is solely my own work.*

1B. Exception to the Class Policy. *I did not follow the CS4248 Class Policy in doing this assignment. This text explains why and how I believe I should be assessed for this assignment given the circumstances explained.*

Signed, [Enter your A... Student ID here]

2. References *Note, this is an optional section, if you referred to other resources. I give credit where credit is due. I acknowledge that I used the following websites or contacts to complete this assignment (but please note that many uses of Web search and detailed discussion are not allowed:*

- *Sample. Website 1, for following mathematical proofs.*
- *Sample. My friend, A0123456X, whom helped me figure out the course deadlines*