CS4248 Natural Language Processing Assignment 2: Solving fact checking

Distributed on 17 Feb 2024
Due in Canvas `Assignments` by 14 Mar 2024 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. *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.*

As for AI tools — such as ChatGPT and Co-pilot — we follow the NUS policy on Guidelines on the Use of AI Tools for Academic Work: `https://libguides.nus.edu.sg/new2nus/acadintegrity` *. You are encouraged to use them to consult for NLP knowledge, but you must acknowledge and submit your chat history alongside your write-up, see Declaration of Original Work. However, you are not allowed to use it directly for prediction or for feature engineering, as this subverts the purpose of our assignment.*

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 (labeled 1), an unimportant statement (labeled 0), or a non-factual statement (labeled -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/cs4248-fact-check-2320/` and join the competition by this link `https://www.kaggle.com/t/b036cd3d20f04baf906c20af6a5415cb`.

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 with your student number (`A01234567.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 `A01234567X/writeup.pdf` and your code should be in Python and the main file should be named `A01234567X/assignment2.py`. Please use Python 3.6 and explain your code dependencies (including the library version) and the instructions to run it in your report. Both files are to be uploaded to Canvas `Assignments` before 14 Mar 2024 11:59 PM SGT.

3 Grading criteria

The grading consists of 70 points from the report and 30 points from the competitive level of the Kaggle final submission. The write-up grading criteria consists of:

3.1 Report

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 steps (since it will be exponential). You can do the analysis either incrementally, grouped into subsets, or a combination of both. Prioritize the analysis of 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.

3.2 Competitive Level

1. Competitive Performance [100%]

   Performance will come in the form of performance relative to a simple baseline — *tf.idf* based Naïe Bayes classification. [‼] Although submitters are welcomed to use Kaggle to validate their performance on the leaderboard, we will no longer be using the leaderboard for grading.

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 am allowed to discuss the problems and solutions in this assignment, but have waited at least 30 minutes by doing other activities unrelated to class before attempting to complete or modify my answers as per the Pokémon Go rule.*

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 did not refer to other resource. 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*

- *I utilised an AI tool, X6543210A* `http://x6543210A.example.com/` *with the following prompts. These were used to accomplish initial brainstorming and to help improve the language of the final report submission.*

Prompt	Output	Purpose
...	...	...
...	...	...
...	...	...
...	...	...