Project Status

CSCI544: Applied Natural Language Processing
Group 44

1 Project Goals/Milestones

Goal-1: Selection of Models and Evaluation:
Select the open-source models along with their APIs to access them. Then perform evaluations on them to see how other models perform and stand with respect to the 'hurtful'ness score.

Goal-2: Fine-tuning of the models:
a) **Prompts-based**: By using prompts to let it not/reduce generating hurtful completions.
b) **Using a custom 'Non-hurtful' Dataset**: Prepare by pre-processing a dataset with no/less hurtful words. Then training the models with that dataset so their distribution could shift towards not/less generating the hurtful completions. Finally, perform evaluations to check their hurtful score.

Goal-3 [Extended Goal]: A separate model for Evaluations: Extend the existing evaluation strategy by having another model to return whether a completion was a hurtful completion and sentence overall, so as to better generalize the results beyond the chosen HurtLex lexicon's limited categories in the paper.

Novelty of the Project:
We discussed that, by accomplishing the above tasks, the novelty in our project is that we are using other recent models to perform evaluations and see the current trend of hurtfulness in the models. Furthermore, we take a step ahead by fine-tuning them so as to reduce the generation of hurtful words in the completions, instead of just reporting the hurtfulness. In the process of this, we would try taking different approaches to fine-tuning them. Additionally, the consideration of getting evaluations done through another model is another step in the direction of generalizing the results beyond the scope of a particular lexicon used in the paper.

Design Consideration:
We discussed not going in the direction of increasing language support as English is widely used and all models support it while other models may not support another language and also, as it does not look necessary enough to consider investing efforts in at the moment.

2 Tasks completed

In this section detailing the completed tasks of this research project, the following activities have been undertaken:

- **Model Selection**: The team collectively curated a set of 2-3 models per member from the Hugging Face model repository.

- **Performance Evaluation**: Subsequently, the chosen models were subjected to "honest" evaluations, as outlined in the respective research papers, and their performance metrics were recorded.

- **Fine-Tuning Experiments**: In addition to the evaluation phase, an essential part of this study involved the fine-tuning of select models. This was achieved through the implementation of two distinct approaches:
  - **Fine-Tuning with Prompts**: Initial fine-tuning experiments were conducted utilizing prompt-based techniques, using multiple prompts, aiming to enhance the models' performance.
  - **Fine-Tuning with Additional Datasets**: We prepared multiple datasets by pre-processing them and removing the hurtful words to further fine-tune the models. So far, we have used reviews and the IMDB dataset.

- **Evaluation of Fine-Tuned Models**: Subsequent to the fine-tuning procedures, evaluations were performed to gauge the performance improvements achieved through these interventions.

Presently, the project is in the stage of fine-tuning the remaining models, with the aim of achieving comprehensive insights into the effectiveness of these strategies in optimizing natural language processing models.

Model Name	Params	Evaluation	Prompt-based Eval.	Dataset-based Eval.	Performed By
Bert-base	109 M	0.00138	0.0359	0.00828	Sarthak
Google-Muril-base	17 Lang.	0.01104	0.0220	0.0110	Sarthak
hateBERT	110 M	0.03176	0.06215	0.02348	Vidit
google-electra-base	33M	0.01519	0.07458	0.01657	Vidit
secBERT	84 M	0.09668	0.16022	0.18646	Vidit
bart-base	110M	0.00856	0.01491	0.00082	Anupam
distilroberta-base	82 M	0.01436	0.00027	0.04392	Anupam
Albert-base	11.8 M	0.0607	0.11187	0.05110	Utkarsha
ClinicalBERT	1.2B Diseases	0.13259	0.135359	0.1325	Utkarsha
LessSexistBert	110 M	0.015	0.11187	0.0096	Harsh
xlm-roberta-base	278 M	0.031	0.05273	0.02019	Harsh

3 Risks and Challenges

We mainly faced two major challenges:

First, we tried using and performing the evaluations upon GPT-3 however, it's accessible only through Open AI and the number of request per minute are limited and further paid, so we couldn't use it.

Second, some models were very computationally heavy with >700M parameters taking very long time to run once.
To mitigate these challenges, as planned, we considered the models not very computationally expensive i.e., having model parameters <=300M parameters. We were successfully able to select 11 such computationally feasible models and proceed further.
Furthermore, our next task/focus, as discussed with our project advisor, is to try using other datasets for fine-tuning purposes. We have planned to preprocess the datasets and remove hurtful words from them and then use them for fine-tuning task in order to get even lower hurtful completion scores.

4 Individual Contributions

Project implementation-wise, as per the above results' table, models selected, evaluations performed, along with the fine-tuning task are listed per individual. Documentation-wise and presentation-wise, one of us prepared the project proposal draft, two of us prepared the project status draft, two of us created the slides, and two of us presented, while all of us edited all docs together at the end to finalize them.

Hence, the work and the tasks have equally been divided and worked upon individually so far.

5 References

[1] Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2398–2406, Online. Association for Computational Linguistics.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186.

[4] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

[5] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017.

Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

[6] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

[7] Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In Proceedings of the 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, pages 1–6. CEUR-WS.

[8] https://github.com/MilaNLProc/honest.