CE807 – Assignment 2 - Final Practical Text Analytics and Report

Student id: 2200367

Abstract

Nowadays, everybody is using internet especially social media. However with the increase usage of social media and it's easy to access nature anywhere from the world, there is a concerning rise in offensive speech. As a result, various kinds of research are being done all over the world to detect offensive speech with the goal of stop spreading hate in the digital world. We are going to discuss current state of art methods and what is it's limitations and how we are going to tackle it.

1 Materials

Here are the materials for the assignments.

- Code
- Google Drive Folder containing models and saved outputs
- Presentation

2 Model Selection (Task 1)

2.1 Summary of 2 selected Models

2.1.1 TimeLMs

TimeLMs is pretrained model based on RoBERTa. It is trained on 154 million tweets and every three months it is trained on more tweets to keep it up to date.

2.1.2 Multilayer Perceptron

Multilayer Perceptron(MLP) performs quite well around various tasks especially classifications. Offensive language detection is no exception.

2.2 Critical discussion and justification of model selection

2.2.1 TimeLMs

In 2019, Google researchers make revolution at the field of natural language processing with the introduction of Bert(Tenney et al., 2019). Bert got state of art performance in 11 criteria. It changed the whole field.

Bert was trained in self supervised way in large corpus. Corpus contains wikipedia and google book 15 percent text were masked and trained to itself to predict the masked word. As a result it could predict bidirectional words. It can predict which word it is going to get next and which word could be previous to it. Bert has understanding of senetence structure. So it can form sentence. It can do various tasks like summary, similarity between two sentences and so on. Using berts understanding of the language it can be finetuned with small dataset to get better result than traditional language models.

Within next few months, Facebook AI researchers discovered that Bert could me improved even more(Liu et al., 2019). They found out that Bert is under trained. They also trained it with bigger batches and more data. Their objective was to have more dynamic model. All those changes brought more improvement around the table. They called their new better and bigger model RoBERTa.

With the evoluation of Bert and RoBERTa, it was clear that pretrained models are the future. However, there are few drawbacks that come with those models. Pretrained models are trained in very large corpus which is very costly and time consuming. Furthermore, they are also harmful for environment as training those models require many gpus and electricity which can pollute environment.

Training large model is not economically feasible nor environmentally. That is where comes transformers. Transformers will keep the base version pretrained model and we can directly use the model or train the model to finetune it. We can transfer the knowledge of base Roberta to learn new tasks. It always better if the base model is trained on similar kind of texts.

As we are detecting offensive tweets, we need

a model that is trained on tweet so it has the understanding of tweet. TimeLMs is trained on 154 millions tweets(Loureiro et al., 2022a). Another issue addressed by TimeLMs is that language model need to be updated. Before 2019, there was COVID. So, model trained before 2019 will have no idea what COVID is. Similar things could happen to offensive language. Both language and context of language is changing. TimeLMs research paper shown impact of model overtime(Loureiro et al., 2022b). New tweets have less accuracy on old models and it become worse over time. That is why TimeLMs is trained with new data every 3 months. Currently, to our knowledge TimeLMs is the state of art algorithm to detect offensive language.

2.2.2 Multilayer Perceptron

When comes to traditional algorithms, MLP always perform well. We can also see that in offensive language detection(Hajibabaee et al., 2022). Pretrained models are quite big and not always suitable. For portable device like IoT and mobile, LLM are not feasible. Mobile giant like apple focusing on running neural engine which can compute mlp neural network quickly. For realtime softwares, mlp is also valuable. So, considering accuracy and performance, MLP is good choice.

3 Design and implementation of Classifiers (Task 2)

In the below table you can see the distribuition of dataset.

Dataset	Total	% OFF	% NOT
Train	12313	33%	67%
Valid	927	33%	67%
Test	860	28%	72%

Table 1: Dataset Details

Final F1 score of both models

Model	F1 Score
TimeLMs	0.811
MLP	0.748

Table 2: Model Performance

4 Data Size Effect (Task 3)

You need to use Table 5 6 and 4 to compare the model's output and provide exciting insights. Note that you can't have same examples in all tables.

Data %	Total	% OFF	% NOT
25%	3078	34%	66%
50%	6156	33%	67%
75%	9234	33%	67%
100%	12313	33%	67%

Table 3: Train Dataset Statistics of Different Size

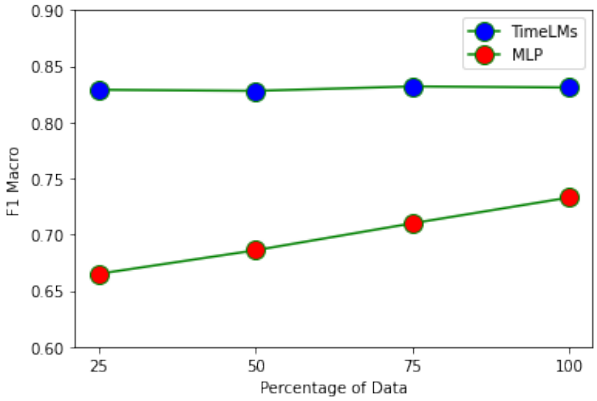

Figure 1: Validation Score on Models based on Different data sizes.

5 Summary (Task 4)

5.1 Discussion of work carried out

5.2 Lessons Learned

Pretrained models need less data compared to traditional methods.

6 Conlusion

In the future I want to balance the imbalance dataset. Also want to try various hyper parameters on TimeLMs models.

References

Parisa Hajibabaee, Masoud Malekzadeh, Mohsen Ahmadi, Maryam Heidari, Armin Esmaeilzadeh, Reyhaneh Abdolazimi, and H James Jr. 2022. Offensive language detection on social media based on text classification. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0092–0098. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022a. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Example	GT	M1(100%)	M2(100%)
1. @USER Do you get the feeling he is kissing @USER behind so he can humiliate him later?	OFF	OFF	NOT
2. : Take note of how to protest with civility. This is why socialism (aka communism) will never win. It is inherently evil and unsustainable. URL	OFF	NOT	NOT
3. Always smack URL	NOT	OFF	NOT
4. #HIAC Damn Matt Hardy and Randy Orton put on one hell in a cell match!! Woooo!!! I hope he is okay!!	NOT	OFF	OFF
5. #ChristineBlaseyFord is your #Kavanaugh accuser... #Liberals try this EVERY time... #ConfirmJudgeKavanaugh URL	OFF	NOT	NOT

Table 4: Comparing two Model's using 100% data:

Example	GT	M1(25%)	M1(50%)	M1(75%)	M1(100%)
1. #SesameStreet #BertandErnie the conversation about this is so unattractive. The creator says he didn't create a character gay and he is being attacked.	OFF	NOT	OFF	OFF	OFF
2. @USER #FakeNewsMedia is the #EnemyOfThePeople! They want to control YOUR minds by controlling what information you get and do not get. The #1stAmendment may protect #elitist #propaganda but it does not legitimize it or make it palatable to an engaged public. URL	NOT	OFF	OFF	OFF	OFF
3. @USER @USER Awwww she is so stinking cute! How old is she now?	OFF	NOT	NOT	OFF	OFF
4. #CarbonTax works and will put tons more money BACK INTO Canadian pockets. .@USER .@USER .@USER .@USER Part of why Trudeau won in 2015 was campaigning on a national carbon tax. Conservatives died then. They want another death in 2019...we will oblige. URL	OFF	NOT	NOT	OFF	NOT
5. #BeckyLynch is beautiful one of the few women in wrestling that absolutely need no work done at all. She's absolutely beautiful just the way she is. Doesn't need giant boobs or a fake booty. @USER is just simply amazing #HIAC	NOT	OFF	NOT	OFF	NOT

Table 5: Comparing Model Size: Sample Examples and model output using Model 1 with different Data Size

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022b. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Example	GT	M2(25%)	M2(50%)	M2(75%)	M2(100%)
1. #SilsilaBadallteRishtonKa tag is filled with appreciation of USER only.... Never watched this serial but I always wanted appreciation for Aditi... She is underrated since starting of the show... She is worth of all this appreciation.. üëèüëèüëè	NOT	NOT	NOT	OFF	OFF
2. @USER @USER @USER @USER @USER @USER @USER @USER @USER I,Äôm shocked to learn human traffickers had guns. Some probably illegal too. CA needs more gun control. But don,Äôt worry about the actual crime. The pic of black guns is worse.	OFF	NOT	NOT	NOT	OFF
3. *gets all the bitches*	OFF	NOT	OFF	OFF	OFF
4. Amazon will ship live Christmas trees to your door ,Äî but will they stay and put on the damn lights? URL	OFF	NOT	OFF	OFF	OFF
5. @USER 'Cult like' caps. Hmm where else do we see cult like caps #MAGA ü§Øü§Ø	NOT	NOT	NOT	OFF	NOT

Table 6: Comparing Model Size: Sample Examples and model output using Model 2 with different Data Size

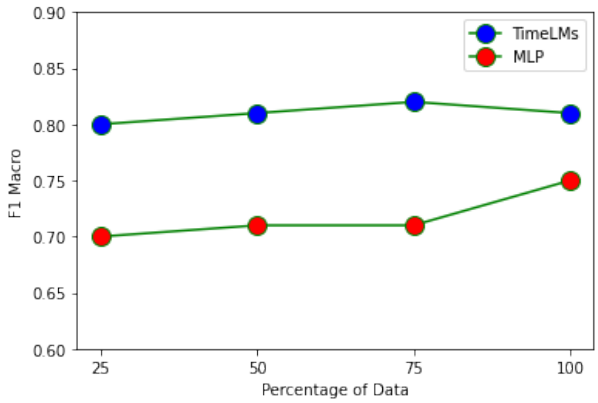

Figure 2: Test Score on Models based on Different data sizes.