IIITH at SemEval-2021 Task 7: Leveraging transformer-based humourous and offensive text detection architectures using lexical and hurtlex features and task adaptive pretraining

Tathagata Raha, Ishan Sanjeev Upadhyay, Radhika Mamidi, Vasudeva Varma
IIIT Hyderabad, India,
{tathagata.raha, ishan.sanjeev}@research.iiit.ac.in
{radhika.mamidi, vv}@iiit.ac.in

Abstract

This paper describes our approach (IIITH) for SemEval-2021 Task 5: HaHackathon: Detecting and Rating Humor and Offense. Our results focus on two major objectives: (i) Effect of task adaptive pretraining on the performance of transformer based models (ii) How does lexical and hurtlex features help in quantifying humour and offense. In this paper, we provide a detailed description of our approach along with comparisions mentioned above.

1 Introduction

Humour is an important part of human conversation. It has a social function as well and can play an important role in group cohesiveness(Ziv, 2010). Hence humorous content is also found on various social media websites. While there has always been a fine line between funny and offensive humour, the anonymity, distance and isolation provided by being online can increase instances of offensive or controversial humour being posted online. (Weitz, 2017)

In this task, we have presented a transformer based approach combined with lexical and hurtlex feature sets to quantify humour and offense of a piece of text.

We achieved an F1 score of 0.959 in the humor classification task and 0.592 in the humor controversy task. For the regression tasks, we achieved a RMSE score of 0.541 and 0.488 in the humor regression and offense regression task respectively.

2 Related work

There have been many attempts made at computational humour detection. In this section, we briefly describe other work in this area. In this approach(Blinov et al., 2019), the authors have used universal language model fine-tuning method for humour recognition. Convolutional neural networks (CNN) have also been used for this task by (Chen and Soo, 2018) whereas (Weller and Seppi, 2019) used transformers to classify humour.

There has also been a lot of shared tasks and workshops related to computational humour. One of them is SemEval-2020 Task 7: Assessing Humor in Edited News Headlines(Hossain et al., 2020) where Zhang(Zhang et al., 2020) used bidirectional neural networks with an attention mechanism and incorporated lexical features to assess humour in edited news headlines.

There has been a lot of work done on hate speech and offensive speech detection as well. CNN's and gated recurrent units (GRU) have been used for this task (Zhang and Luo, 2018). Recurrent neural networks combined with user-related information have also been used for hate speech detection in Twitter Data (Pitsilis et al., 2018) whereas multilingual transformer architectures were leveraged by (Ghosh Roy et al., 2021) to detect hostile content in English, Hindi and German.

3 Task and dataset overview

The task(Meaney et al., 2021) is divided into 4 sub-tasks.

1. **Humour detection**: This is a binary classification task where the model needs to predict if the text is humorous or not where the values are either 0 and 1.

2. **Humour Rating**: This is a regression task where the model needs to rate how humorous the text is where the value can vary between 0 to 5.

3. **Controversy detection**: This is a binary classification task where the model needs to classify text as controversial or not if it has been classified as humorous. It can be either 0 or 1.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1221–1225
Bangkok, Thailand (online), August 5–6, 2021.

4. **Offense Rating**: This is a regression task where the model needs to rate how offensive the text is. It can vary between 0 to 5.

   The dataset for the tasks was provided by The workshop organizers. It consisted of 10,000 sentences. 8,000 sentences were provided for training and 1,000 for validation. The remaining 1,000 were used for testing. Each row consisted of a unique identifier,the text and the label values of "is_humor", "humor_rating","humor_controversy" and "offense_rating".

4 Methodology

4.1 Hurtlex features

HurtLex(Bassignana et al., 2018) is a lexicon of offensive, aggressive, and hateful words in over 50 languages which is further categorized into 17 categories. Identifying these kinds of words can potentially help in offensive content detection. Also, in some cases, a humorous piece of text might contain such a word to denote humour. We have also experimented with this feature for humour classification and regression task.

4.2 Lexical features

The structure of humorous and offensive texts can be a bit different from normal texts. We have leveraged a lexical feature set that would help us capture that information and distinguish humorous and offensive texts. The set of lexical features are:

- Counting the total number of letters, punctuation, upper case letters and numbers within the text.
- Identifying the presence of any named entity. For detecting named entities, we have used the AllenNLP named entity recogniser[1] which uses pretrained GloVe vectors for token embeddings and a GRU encoder.(Peters et al., 2017)
- Detecting the presence of interrogation by identifying '?' symbol or any WH-word
- Detecting the number of personal pronouns and what kind of personal pronouns they are: first-person, second-person or third-person. For detecting the personal pronouns, we have used a pre-defined list of personal pronouns.

[1] https://demo.allennlp.org/named-entity-recognition/named-entity-recognition

4.3 Sentence embeddings

For generating the sentence embeddings, we have experimented with 4 different pre-trained transformer models: bert-base-uncased(Devlin et al., 2018), roberta-base(Liu et al., 2019), google/electra-base-discriminator(Clark et al., 2020) and xlnet-base-cased(Yang et al., 2019). Initially, we finetuned each of the pre-trained models for each task and made predictions on the validation set. On the basis of the performance, we have selected one pre-trained model to proceed to our final setup4.5. For the binary humour classification, humour regression and offensive regression task, we have selected roberta-base. On the other hand, google/electra-base-discriminator gave the best performance for humour controversy task.

4.4 Task adaptive pretraining

In the paper (Gururangan et al., 2020), we can see the benefits of continued pretraining of pre-trained transformer models on unlabelled task-specific data or Task Adaptive Pretraining (TAPT) before fine-tuning them on a downstream task like text classification. This paper (Raha et al., 2021) showcases the gains attributed to further pre-training of the IndicBERT(Kakwani et al., 2020) model for hostility detection in Hindi. We have experimented with the same approach for all our downstream tasks where a pretrained transformer model(roberta-base for humor classification, regression and offensive regression) is further pretrained on training data with the masked language modelling (MLM) objective. In our results 5, we have shown the benefits gained from task adaptive pretraining for each task. Note that task adaptive pretraining was not done on google/electra-base-discriminator for the humour controversy classification.

4.5 Final setup

In this subsection, we outline our final architecture from the set of input features to the final label generation for each task.

At first, we have generated the set of lexical features and the hurtlex features on both training, validation and testing data. For generating the hurtlex features, we have used the featurizer in hurtlex

Setting	Task 1a (F1-Score)	Task 1b (RMSE)	Task 1c (F1-Score)	Task 2 (RMSE)
TRANS	0.944	0.572	**0.592**	0.522
TRANS + LEX	0.956	0.547	0.521	0.524
TRANS + HURT	0.949	0.570	0.347	**0.488**
TRANS + LEX + HURT	**0.959**	**0.541**	0.375	0.505

Table 1: Results on the Validation split for each task with and without hurtlex and lexical features. TRANS refer to transformer embeddings, LEX refer to lexical features and HURT refers to hurtlex features. Task 1a refers to humour classification, Task 1b refers to humour regression, Task 1c refers to humour controversy and Task 2 refers to the offensive regression task. For Task 1a, 1b and 2 we have used the TAPT roberta-base and for task 1c we have used pre-trained google/electra-base

Github repository [2]. We have used Pytorch(Paszke et al., 2019) [3] and Pytorch Lightning as our primary deep-learning framework [4]. For our pre-trained transformer models, we chose the roberta-base [5] and google/electra-base-discriminator[6] as a part of HuggingFace's Transformers library. For performing the task adaptive pretraining(TAPT) on downstream tasks, we have used AllenAI's implementation of Task Adaptive Pretraining[7]. The roberta-base model was further pretrained on MLM objective for 100 epochs with the other hyperparameters being set to their default values. For all the transformer architectures, we have set the maximum sequence length to 128. As this is a classification task, we have used the embeddings of [CLS] as the transformer representation of the whole sentence.

Finally, the embeddings generated from the transformer models are concatenated with hurtlex features and lexical features to form the final vector representation for a particular text. For optimization, we have used the Adam (Kingma and Ba, 2017) optimizer where the learning rate was set to 1e-5 and a dropout (Srivastava et al., 2014) with the probability of 0.1. We updated weights based on cross-entropy loss values for the classification tasks and Mean Squared Error for the regression tasks. A dense multi-layer perceptron serves as the final binary classifier head or regression head. The model weights were saved and evaluated on the development set at the end of every epoch and the finetuning continued for 10 epochs. We have reported the scores of the models that yielded the best F1 score on the development set and used them to further predict on the test set. We have also experimented with or without considering the hurtlex and lexical features to showcase the gains or losses attributed to them.

Task	Without TAPT	With TAPT	Gains
Task 1a (F1-Score)	0.933	0.944	0.011
Task 1b (RMSE)	0.616	0.572	0.044
Task 2 (RMSE)	0.579	0.522	0.057

Table 2: Results on the Validation split for each task with and without Task Adaptive Pretraining(without considering the lexical and hurtlex features). Task 1a refers to humour classification. Task 1b refers to humour regression and Task 2 refers to the offensive regression task.

5 Results

The gains attributed to task adaptive pretraining of roberta-base on the humour classification is shown in table 2. We can see that continued pretraining of roberta-base has improved the model performances significantly.

In table 1, we can see the results of inclusion and exclusion of the lexical and hurtlex features for each task. We notice that lexical and hurtlex features do contribute to the performance of humour classification. Combining hurtlex features and lexical features with transformer embeddings have improved the results of both humour classification and humour regression task. For offensive regression, the hurtlex features played an important role while lexical features degraded the performance. This is probably because the lexical features were curated for the identification of humour. For the

[2] https://github.com/valeriobasile/hurtlex

[3] pytorch.org/

[4] https://www.pytorchlightning.ai/

[5] https://huggingface.co/roberta-base

[6] https://huggingface.co/google/electra-base-discriminator

[7] github.com/allenai/dont-stop-pretraining

Task	Score	Rank
Task 1a (F1-Score)	0.9616	14
Task 1b (RMSE)	0.5263	5
Task 1c (F1-Score)	0.6242	6
Task 2 (RMSE)	0.4772	23

Table 3: Results on the test split for each task and their respective ranks on the leaderboard during the evaluation phase. Task 1a refers to humour classification, Task 1b refers to humour regression, task 1c refers to humor controversy and Task 2 refers to the offensive regression task.

humour controversy, excluding lexical and hurtlex features gave the best results. This might be because textual features played much more important role than lexical and hurtlex features.

In table 3, we report the results obtained on the test set during the evaluation phase and the rank of our models on the official leaderboard[8]. We used the best performing models on the validation set to achieve those results.

Overall, this work shows how task adaptive pretraining can improve model performance for downstream tasks and the role of hurtlex and lexical features for humor and offensive detection.

6 Conclusion

All the experiments performed above were done with default hyperparameters(unless explicitly mentioned) due to resource constraints. The model performances could have improved if we could search for optimal hyperparameters using cross validation. Furthermore, the regression tasks could improve if we could use an ensemble of the best performing models for our final predictions.

References

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

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4027–4032, Florence, Italy. Association for Computational Linguistics.

[8] http://smash.inf.ed.ac.uk/tasks_results/hahackathon_results.html

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection. *arXiv e-prints*, pages arXiv–2101.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. Semeval-2020 task 7: Assessing humor in edited news headlines. *arXiv preprint arXiv:2008.00304*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

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*.

J.A. Meaney, Steven R. Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani,

Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and R. Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Tathagata Raha, Sayar Ghosh Roy, Ujwal Narayan, Zubair Abid, and Vasudeva Varma. 2021. Task adaptive pretraining of transformers for hostility detection. *arXiv preprint arXiv:2101.03382*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Eric Weitz. 2017. Editorial: Humour and social media. *The European Journal of Humour Research*, 4:1.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Tiantian Zhang, Zhixuan Chen, and Man Lan. 2020. ECNU at SemEval-2020 task 7: Assessing humor in edited news headlines using BiLSTM with attention. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 995–1000, Barcelona (online). International Committee for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter.

Avner Ziv. 2010. The social function of humor in interpersonal relationships. *Society*, 47:11–18.