RUTGERS

Leveraging BERT and Convolutional Neural Networks to Aid in Identifying Hate Speech Amongst Online Gaming Communities

Malav Modi - (mhm114)
Shyam Patel - (spp128)
Praveen Sakthivel - (ps931)

Department of Computer Science
Rutgers University

Topic Agenda:

RUTGERS

- I - **Introduction**
- II - **Previous State of Knowledge**
- III - **Challenges**
- IV - **Main**
- V - **Evaluations**
- VI - **Conclusion**

Introduction:

RUTGERS

- Due to the recent pandemic, there has been a massive surge in the number of people who remain active in various online communities
- Hate Speech has been a long standing problem that has plagued the gaming community
- Due to huge imbalances between players and moderator, manual moderation is largely ineffective
- An automatic moderator that can accurately and efficiently detect hate speech would be useful in helping control this issue

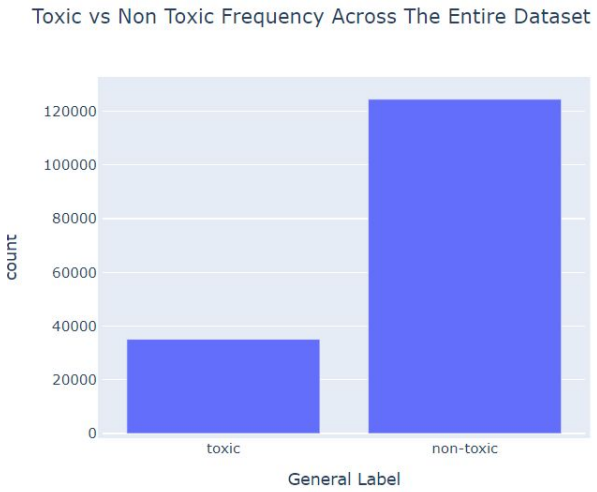
Previous State of Knowledge:

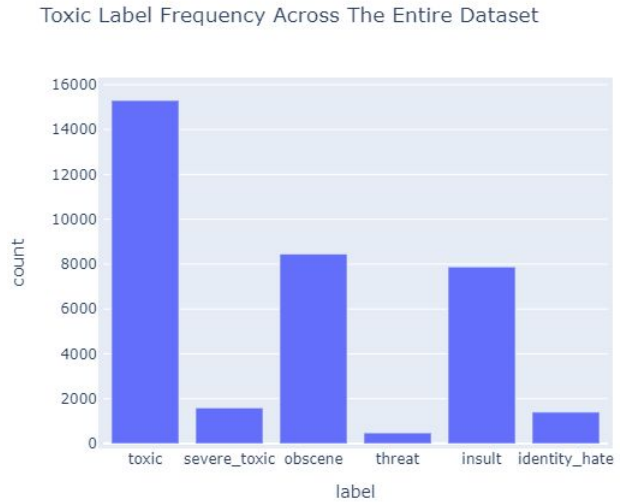
RUTGERS

- Kaggle Dataset
- Previous attempts:
  - Naive Bayes + Support Vector Machine
  - Bidirectional LSTM
  - Logistic Regression

Challenges:

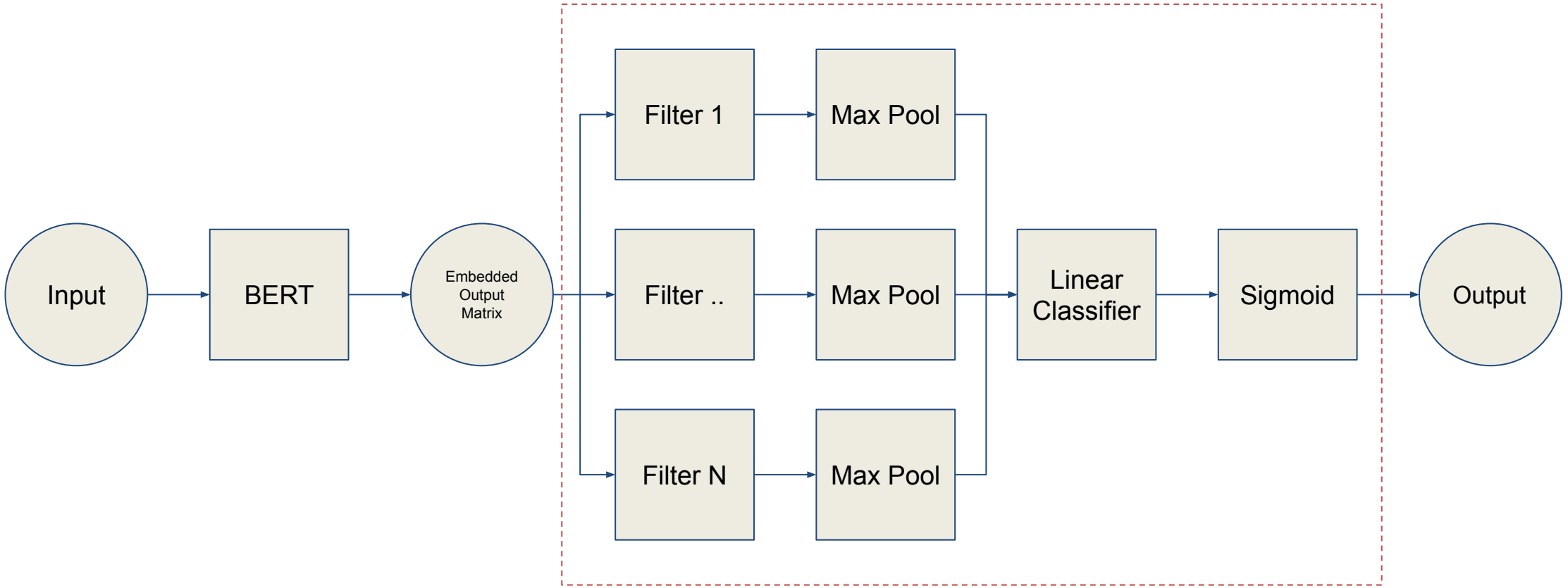
- Dataset is skewed
  - Majority of comments are non toxic
  - Not too many examples of toxic comments
- Model overfitting
  - Manipulating hyperparameters
- Capturing subtones and non explicit data





Main: Architecture

RUTGERS



Kim CNN

Encoder: BERT

RUTGERS

- Bidirectional Encoder Transformation from Transformers
- Looks at context of ALL Words not just the preceding ones
- Masked LM
  - Train with entire input, mask 15% of words
  - Mask 80%, Keep 10%, Incorrect 10%
- Next Sentence Prediction
  - Receives 2 sentences, determines if 2nd is subsequent sentence
  - 50% 2nd is subsequent, 50% is a random sentence

Decoder: Kim CNN

RUTGERS

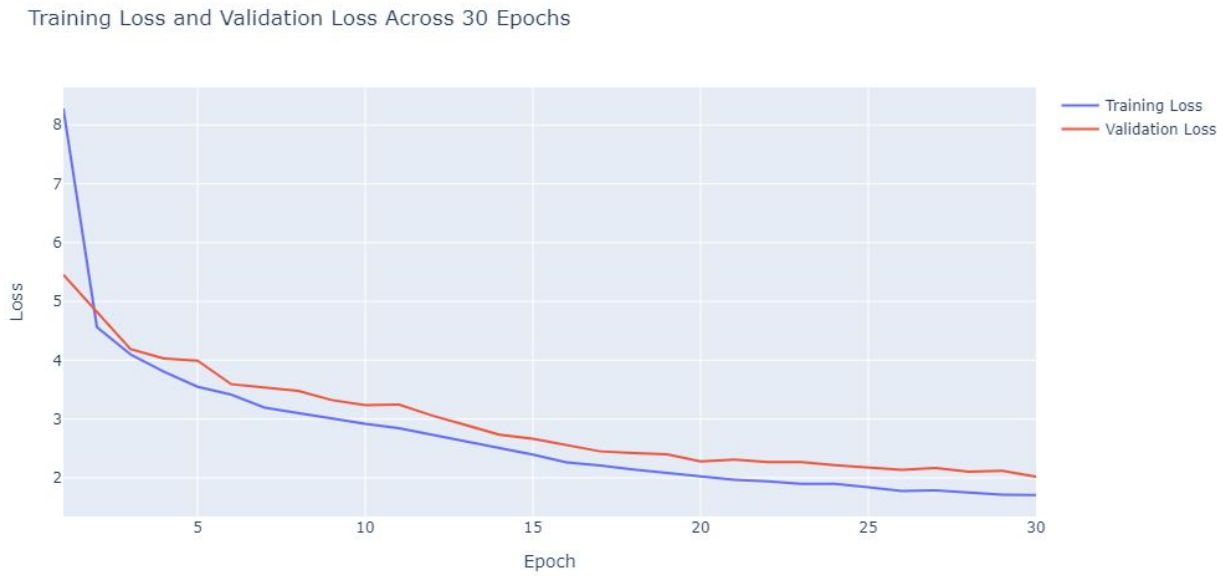
- Embedded Word Vectors have ‘Universal Features”
- Apply multiple filters to the input sequence in parallel
- Max Pool: Extract the highest value feature from each filter
- The max value features are passed into a linear classifier
- Dropout and Sigmoid are applied to produce classifications

Contribution Evaluations:

RUTGERS

- Our Model was very successful in classifying hate speech
  - 96.7% overall accuracy (test set)
  - Used an ROC-AUC Curve to measure individual label accuracy
    - At least ~94% score across all 6 labels in dataset

label	auc
insult	0.973251
obscene	0.970923
identity_hate	0.960268
toxic	0.957884
threat	0.953939
severe_toxic	0.939353



Conclusion:

RUTGERS

- Our results support Kim's findings that encoders are universal feature extractors (The BERT embedding we used was not trained for hate speech classification)

- Severely Toxic was relatively the most difficult hate speech to detect
  - Most likely because it was one of the labels with the fewest data points

- We found the best learning rate to be 9.625e-6
  - Smaller learning rates prevented overfitting with large number of epochs

RUTGERS

Thank You!

Malav Modi - (mhm114)
Shyam Patel - (spp128)
Praveen Sakthivel - (ps931)

Department of Computer Science
Rutgers University

References

RUTGERS

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.

[2] A. G. D'Sa, I. Illina, and D. Fohr. Bert and fasttext embeddings for automatic detection of toxic speech. In 2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA), pages 1–5, 2020.

[3] P. Fortuna, J. Soler, and L. Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 6786–6794, 2020.

[4] Y. Kim. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[5] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. Hate speech detection: Challenges and solutions. PloS one , 14(8):e0221152, 2019.

[6] Rani Horev. BERT Explained: State of the art language model for NLP , 2018. Available at https://towardsdatascience.com/bert-explained-state-of-the-art-language\-model-for-nlp-f8b21a9b6270, Uploaded: November 10th, 2018.

[7] I. Tenney, D. Das, and E. Pavlick. Bert rediscovers the classical nlp pipeline. arXiv preprint arXiv:1905.05950 , 2019.

[8] B. van Aken, J. Risch, R. Krestel, and A. Löser. Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572 , 2018.

[9] Y. Zhang and B. Wallace. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, 2016.