Automated Content Moderation: NLP Against Toxic Comments

Team 1: Jiashu Chen, Qianqian Liu, Irene Yang, Chesie Yu

Table of contents

Table of contents

Motivation

Motivation

Toxic Content on Social Media

- **Harmful & far-reaching** impact
  - **Perpetuates & amplifies** as social media platforms **broadcast** their content to a wider audience
- Social media companies need to **act responsibly**
  - Section 230 of the Internet
  - Provide a forum of **free speech & safe environment**

Significance

Automated Content Moderation (NLP/DL)

- Need to investigate **creation & dissemination** of toxic content
  - Understand the **patterns** of abusive language
- **Combat** cyberbullying and online hate speech
  - Detection & Prevention
  - Create a **safer and more inclusive** online environment

Research Question

How to leverage NLP techniques to detect and classify six subtypes of toxic comments:

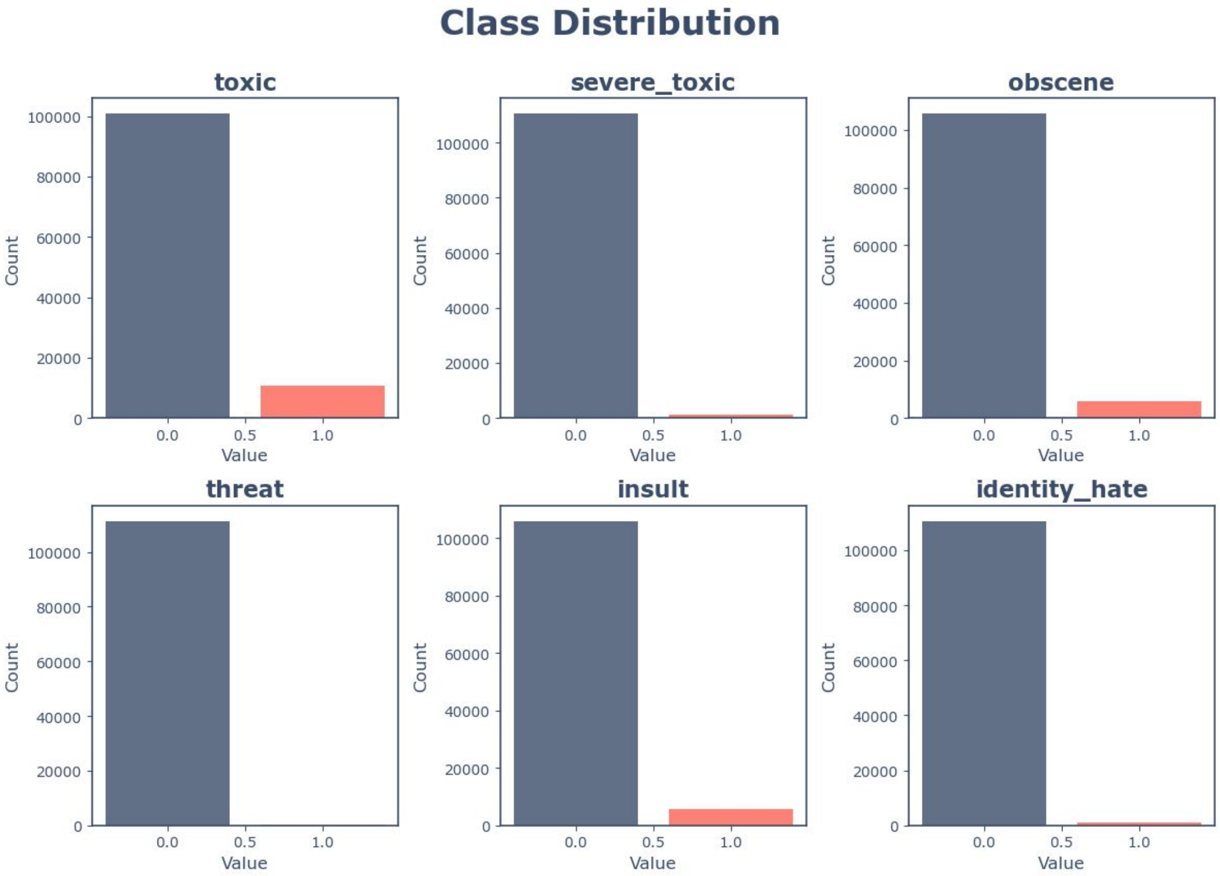
toxic, severe toxic, obscenity, threat, insult, and identity-hate.

Related Work

Year	Researchers	Study	Data	Approach
2016	Waseem & Hovy	Hateful symbols or hateful people? predictive features for hate speech detection on Twitter	Tweets	**Hand labeling criteria with n-grams**
2017	Vigna et al.	Hate me, hate me not: Hate speech detection on facebook	Facebook comments	**SVM, LSTM**
2018	Schmitt et al.	Joint aspect and polarity classification for aspect-based sentiment analysis	SemEval 2017	**CNN, LSTM**
2019	Kraus et al.	Sentiment analysis based on rhetorical structure theory	Movie Database (IMDb), Food Reviews (Amazon)	**Tree-LSTM, Discourse-LSTM**

Data

Type	Count
toxic	15294
severe_toxic	1595
obscene	8449
threat	478
insult	7877
identity_hate	1405



Data

id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York	0	0	0	0	0	0
000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)	0	0	0	0	0	0
000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk pa	0	0	0	0	0	0
0001b41b1c6bb37e	" More I can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of accidents"" -I think th There appears to be a backlog on articles for review so I guess there may be a delay until a reviewer turns up. It's listed in the relevant form eg Wikipedia:Gooc	0	0	0	0	0	0
0001d958c54c6e35	You, sir, are my hero. Any chance you remember what page that's on?	0	0	0	0	0	0
00025465d4725e87	" Congratulations from me as well, use the tools well. · talk "	0	0	0	0	0	0
0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
00031b1e95af7921	Your vandalism to the Matt Shirvington article has been reverted. Please don't do it again, or you will be banned.	0	0	0	0	0	0

Training Set — Development Set — Testing Set

Approach

Metrics	Definition	Importance
Accuracy	• The ratio of correct predictions	• Accuracy is an important indicator, yet accuracy alone is not informative for unbalanced datasets
Confusion Matrix	• FP: Predicting negative class as positive • FN: Predicting positive class as negative	• Focus on FP, FN part of the confusion matrix
***Precision**	• Precision is the ratio of the correct positive predictions to all positive predictions.	• Low precision means that users' regular comments / posts being identified as toxic comments. It will negatively influence user experiences.
***Recall**	• Recall is the ratio of the correct positive predictions to all observations in the positive class.	• Low recall means that toxic comments are not captured. It indicates low effectiveness of content moderation model.
***F1 Score**	• F1 score = 2*(Precision * Recall) / (Precision + Recall)	• The F1 score combines precision and recall, which is a good indicator of overall our model performance.

Text Preprocessing

Explanation\nWhy the edits made under my username Hardcore Metallica Fan were reverted?....I'm retired now.9.205.38.27

Remove space, tab

Explanation Why the edits made under my username Hardcore Metallica Fan were reverted?...I'm retired now.9.205.38.27

Lowercase sentence

explanation why the edits made under my username hardcore metallica fan were reverted?...i'm retired now.89.205.38.27

Remove non-alphabets

explanation why the edits made under my username hardcore metallica fan were reverted...i am retired now

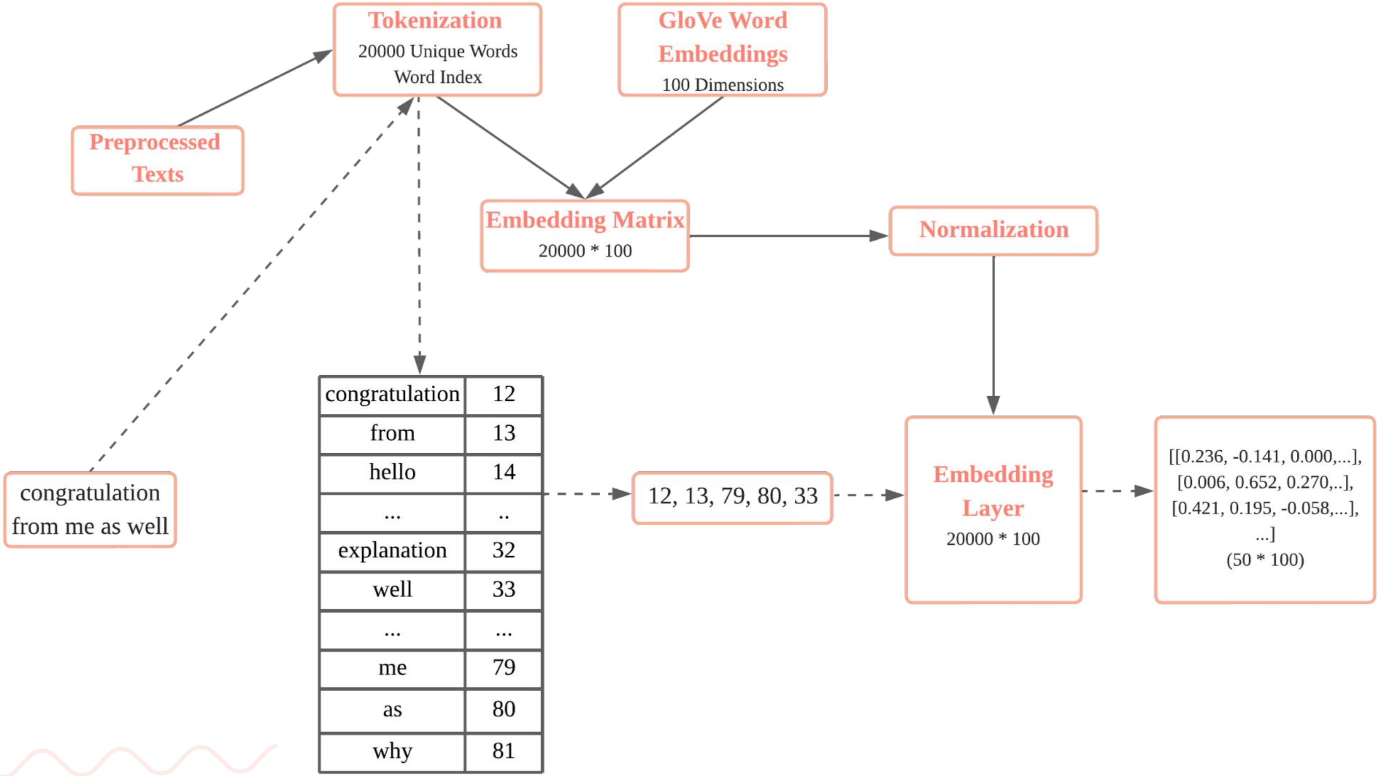
POS Tagging + Lemmatization

'explanation', 'why', 'the', 'edits', 'make', 'under', 'my', 'username', 'hardcore', 'metallica', 'fan', 'be', 'revert'...'i', 'be', 'retire', 'now'

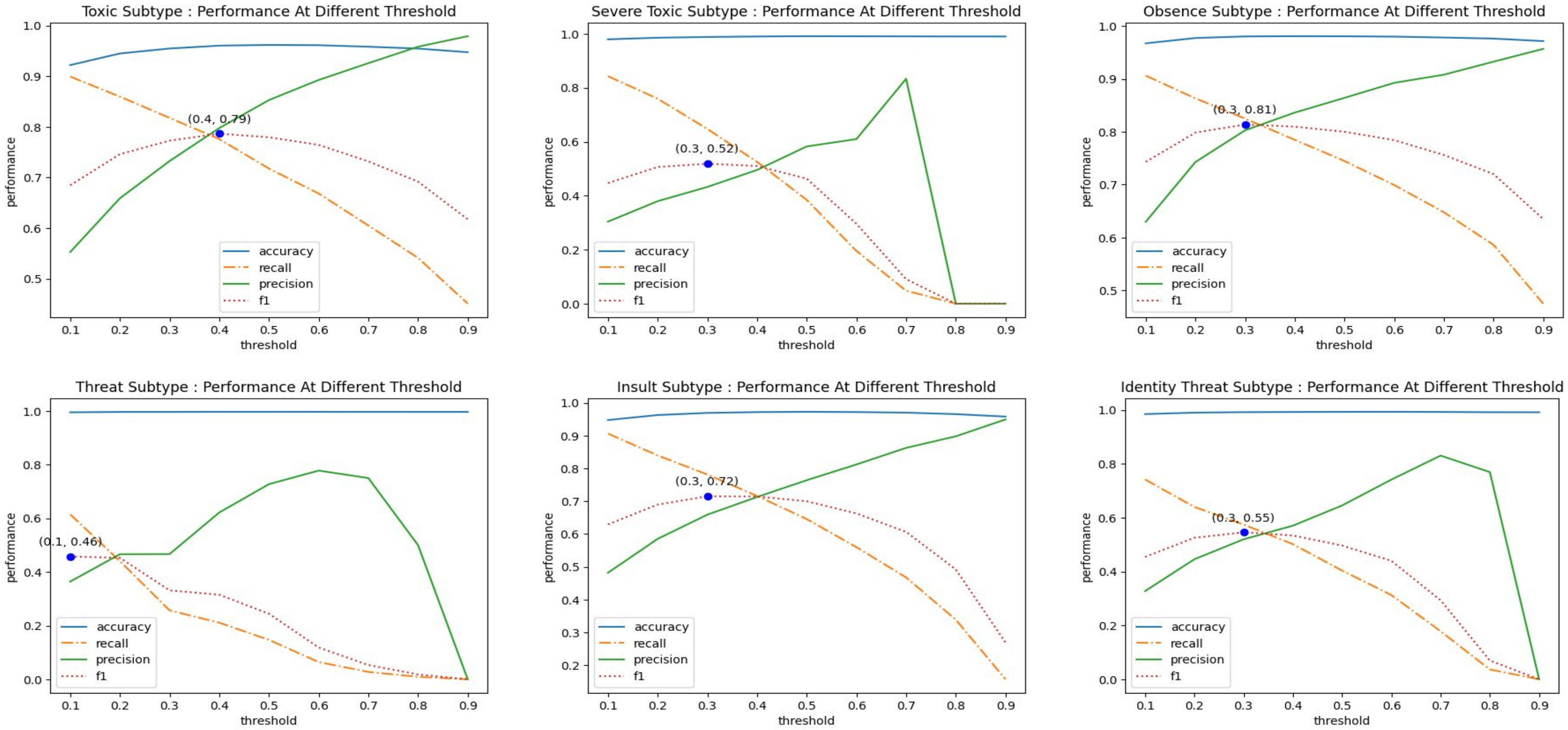
Remove stopwords

'explanation', 'edits', 'make', 'username', 'hardcore', 'metallica', 'fan', 'revert'...'retire'

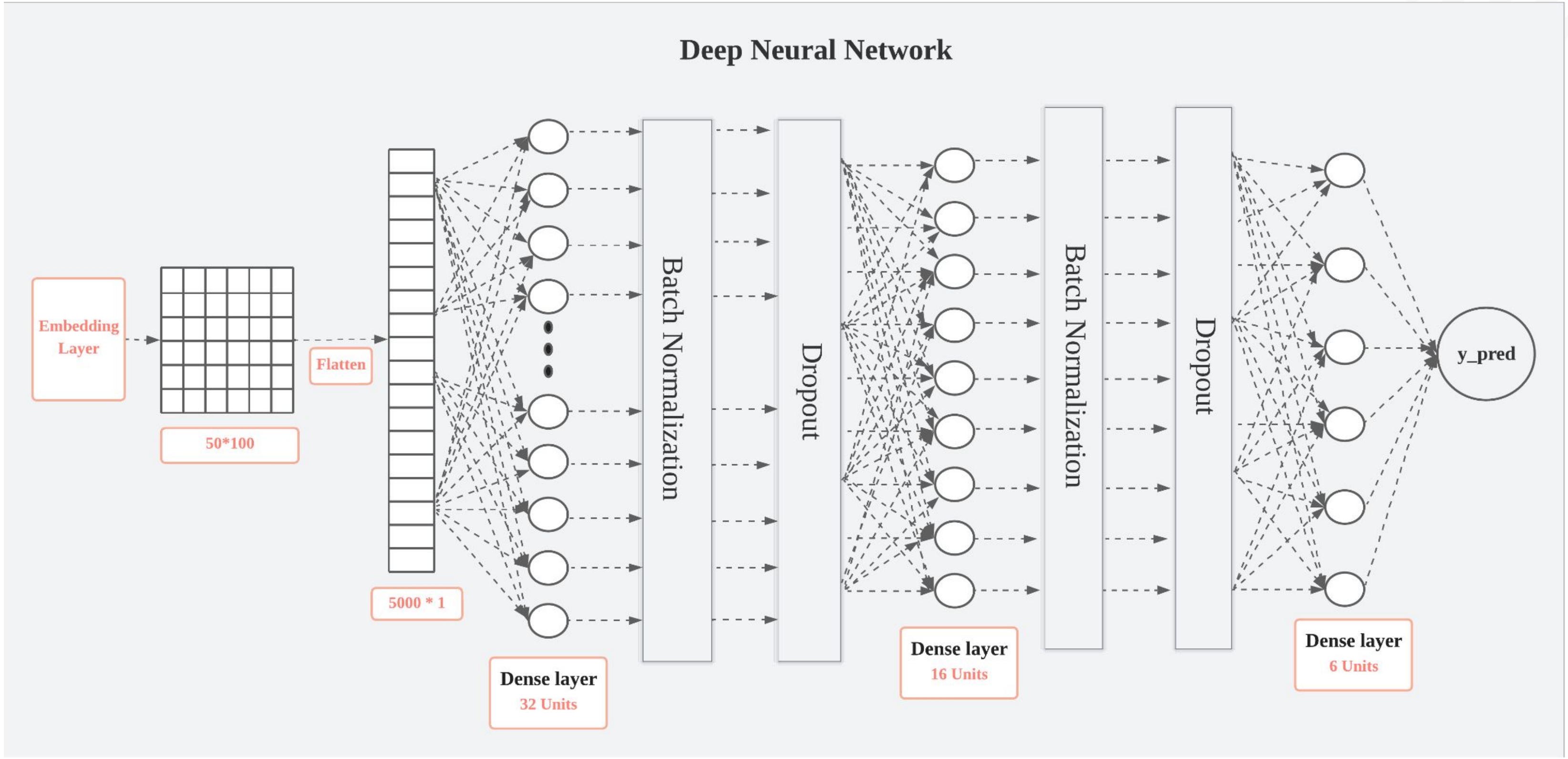
Feeding Text Data Into Neural Network



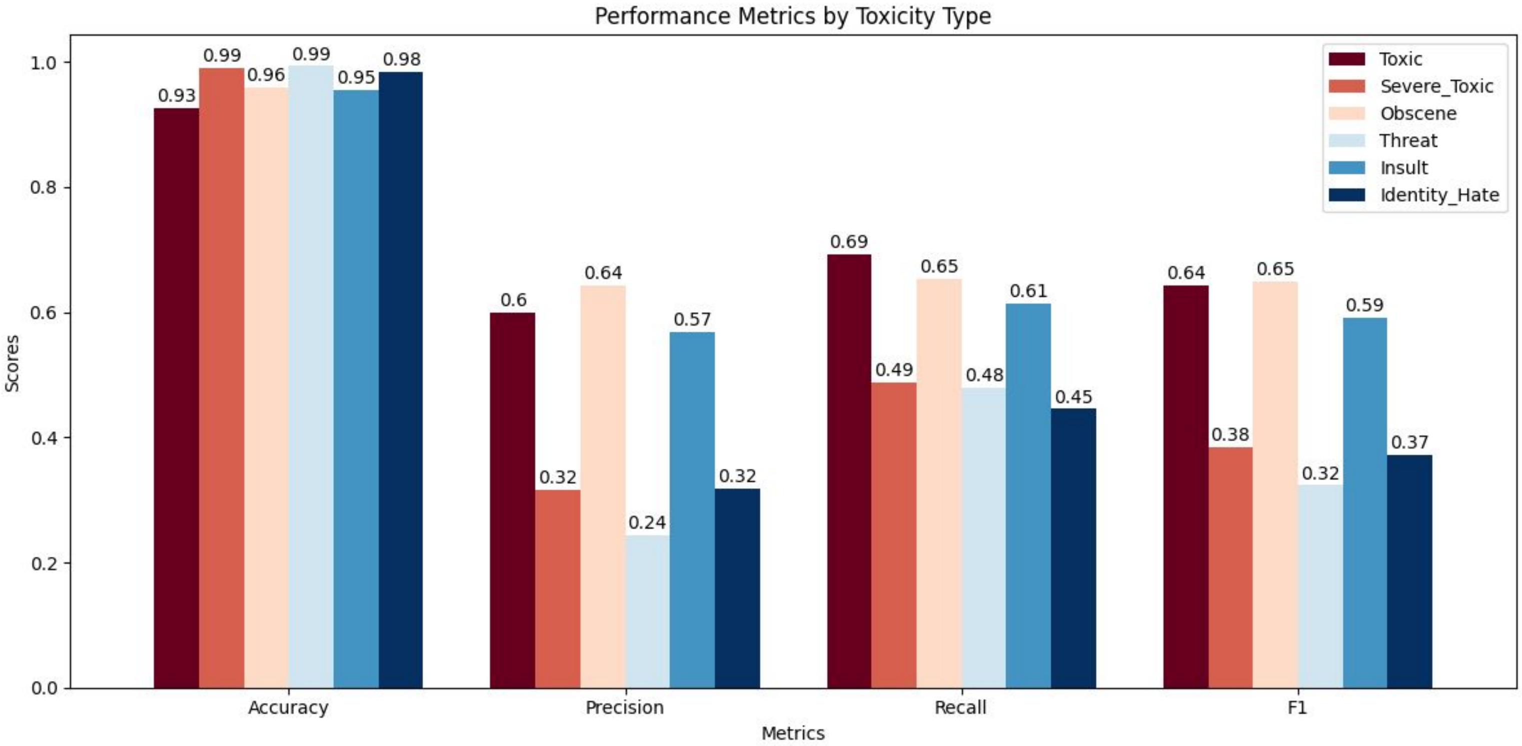
Classification Threshold



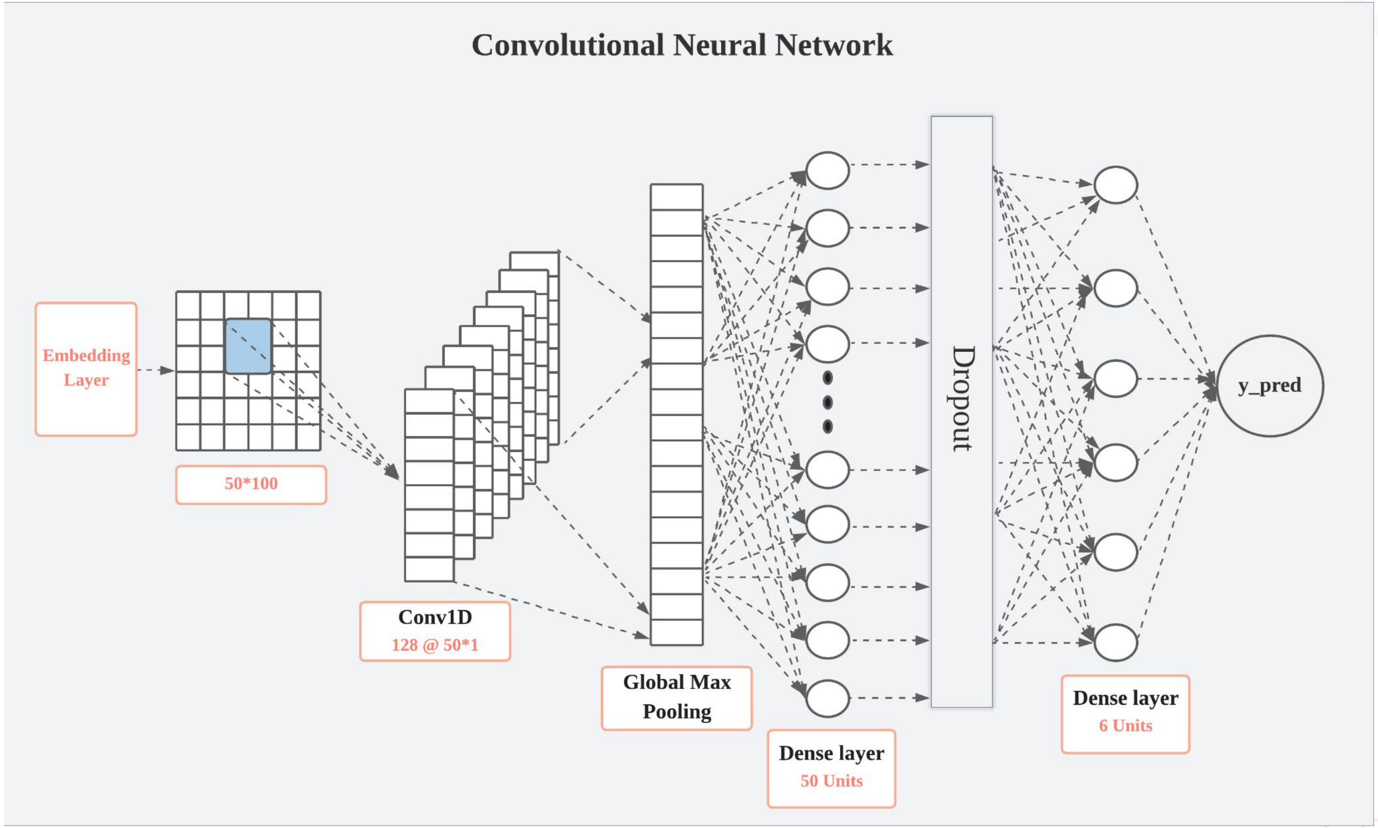
Experiment 1 - Deep Neural Network



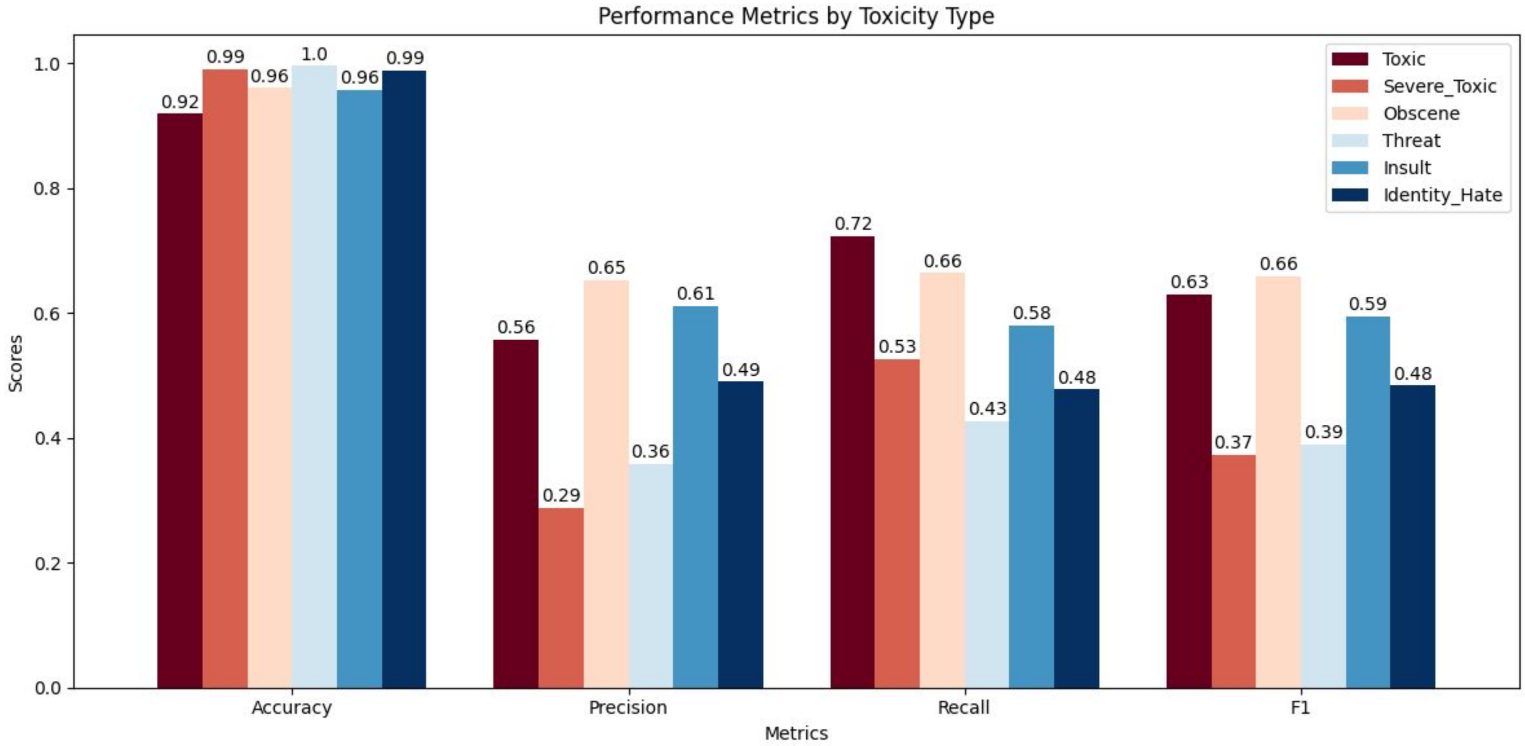
Experiment 1 - Deep Neural Network



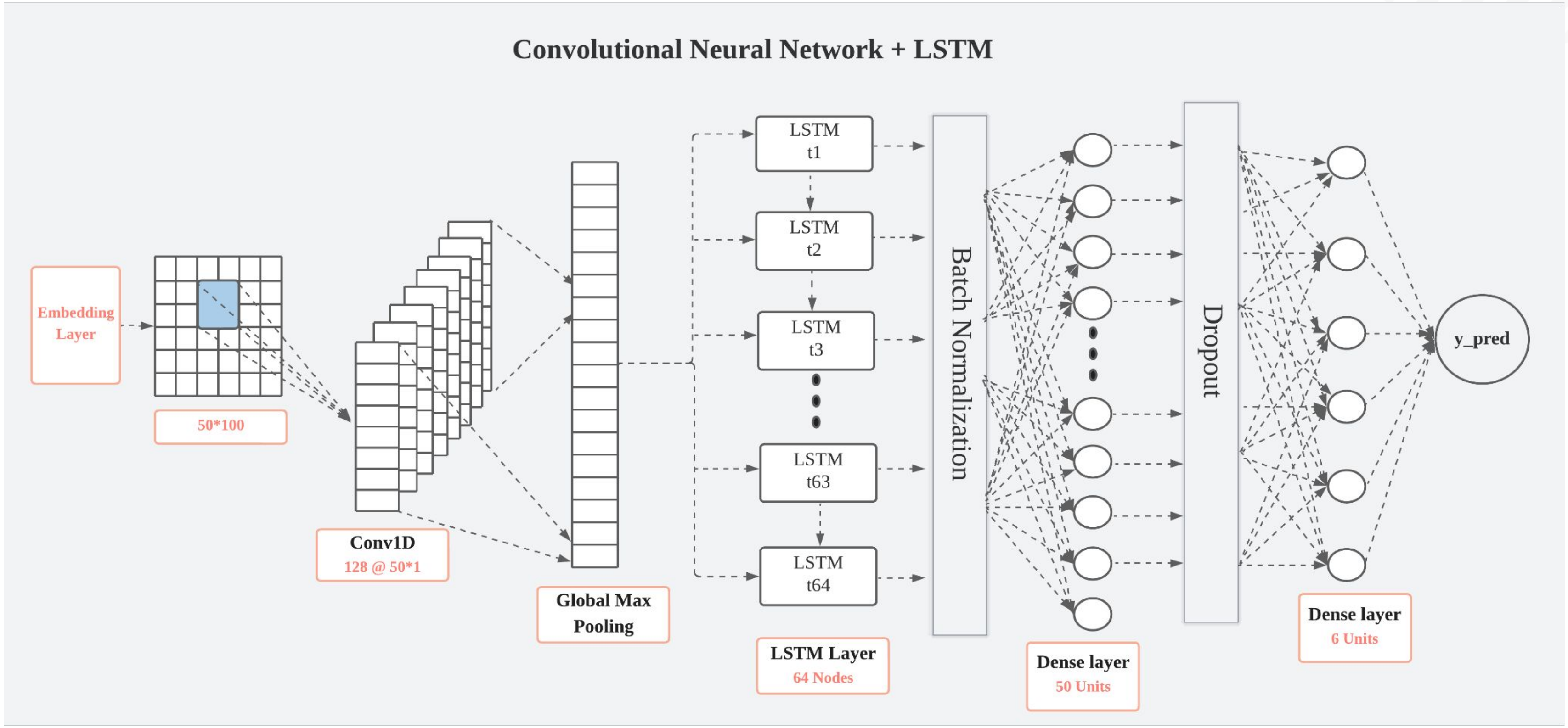
Experiment 2 - Convolutional Neural Network



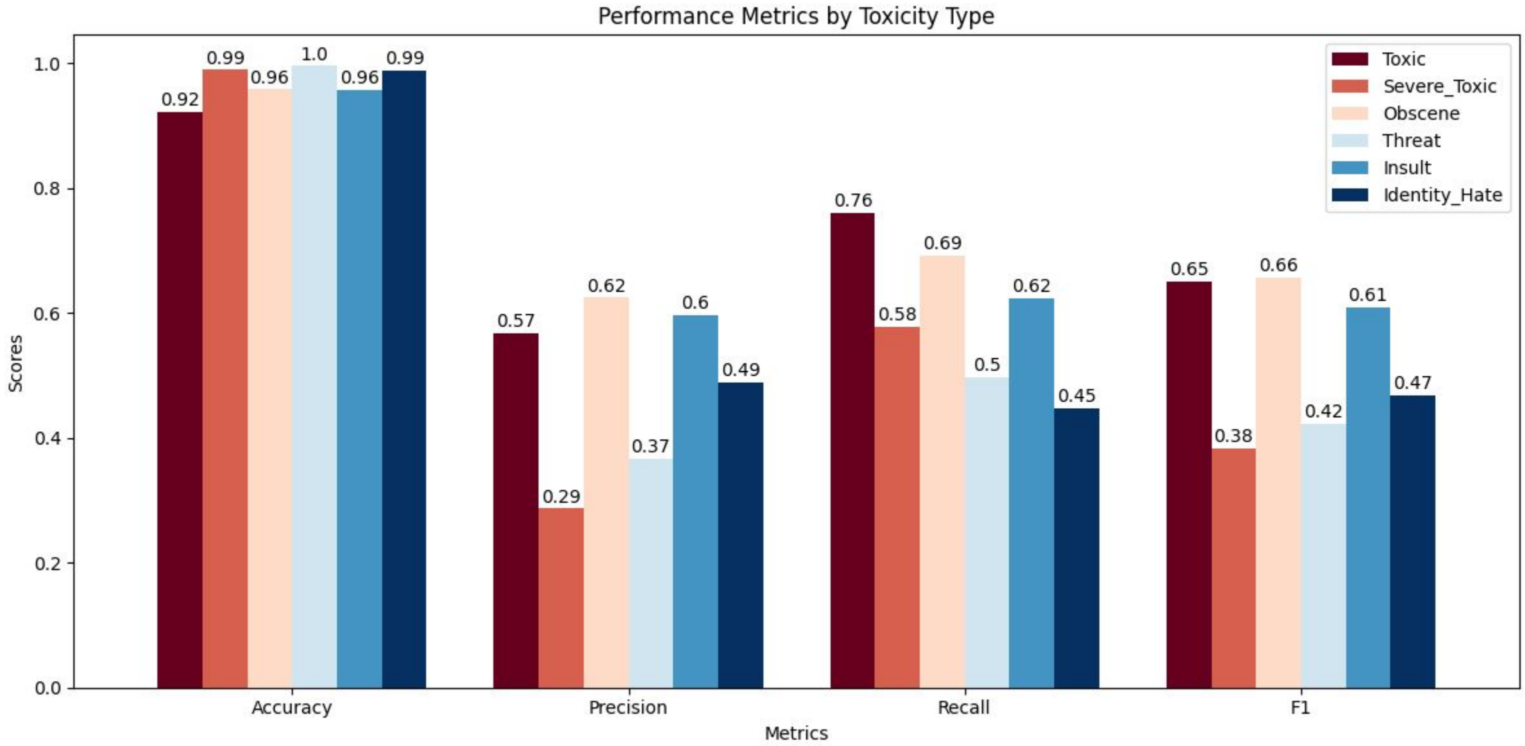
Experiment 2 - Convolutional Neural Network



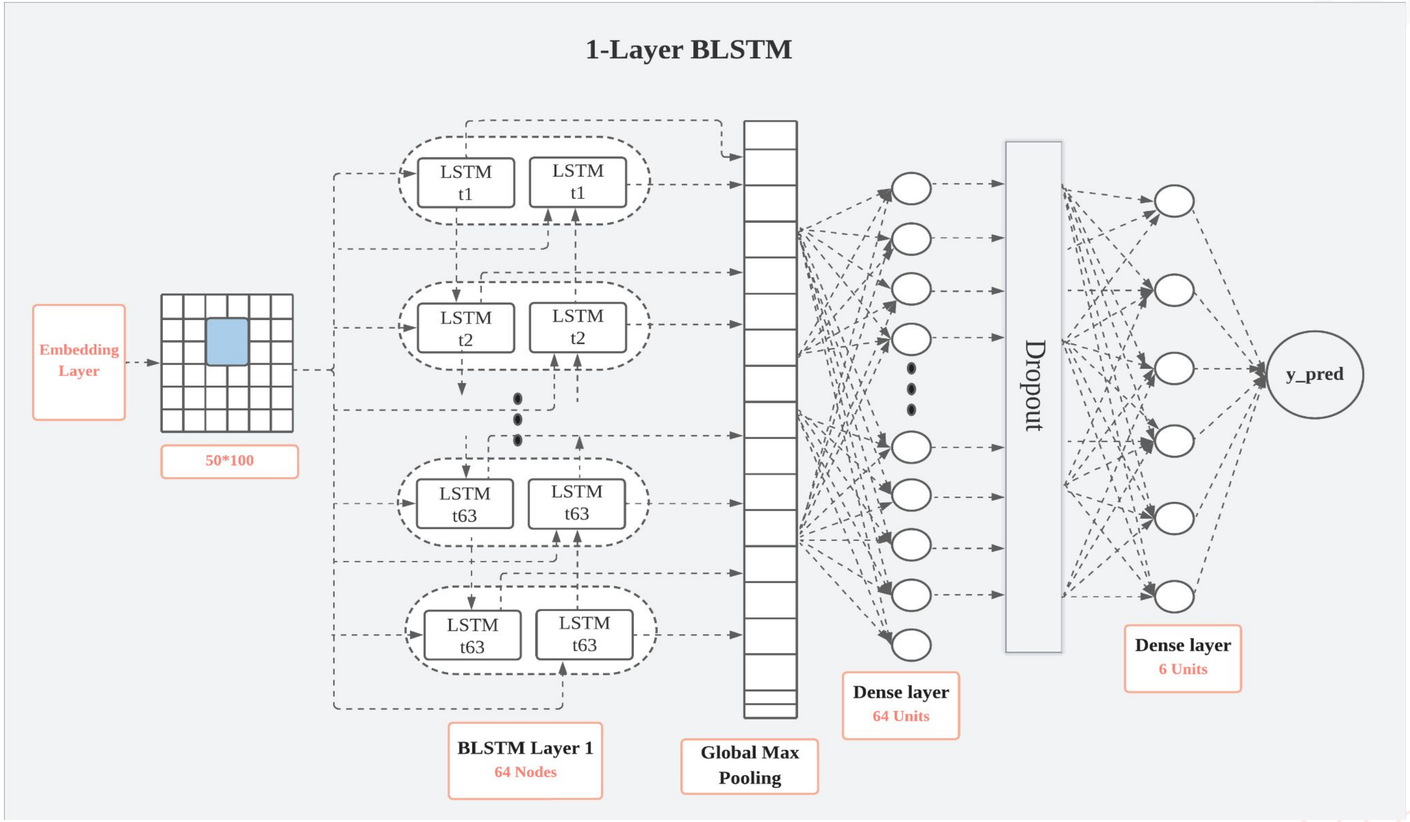
Experiment 3 - CNN + LSTM



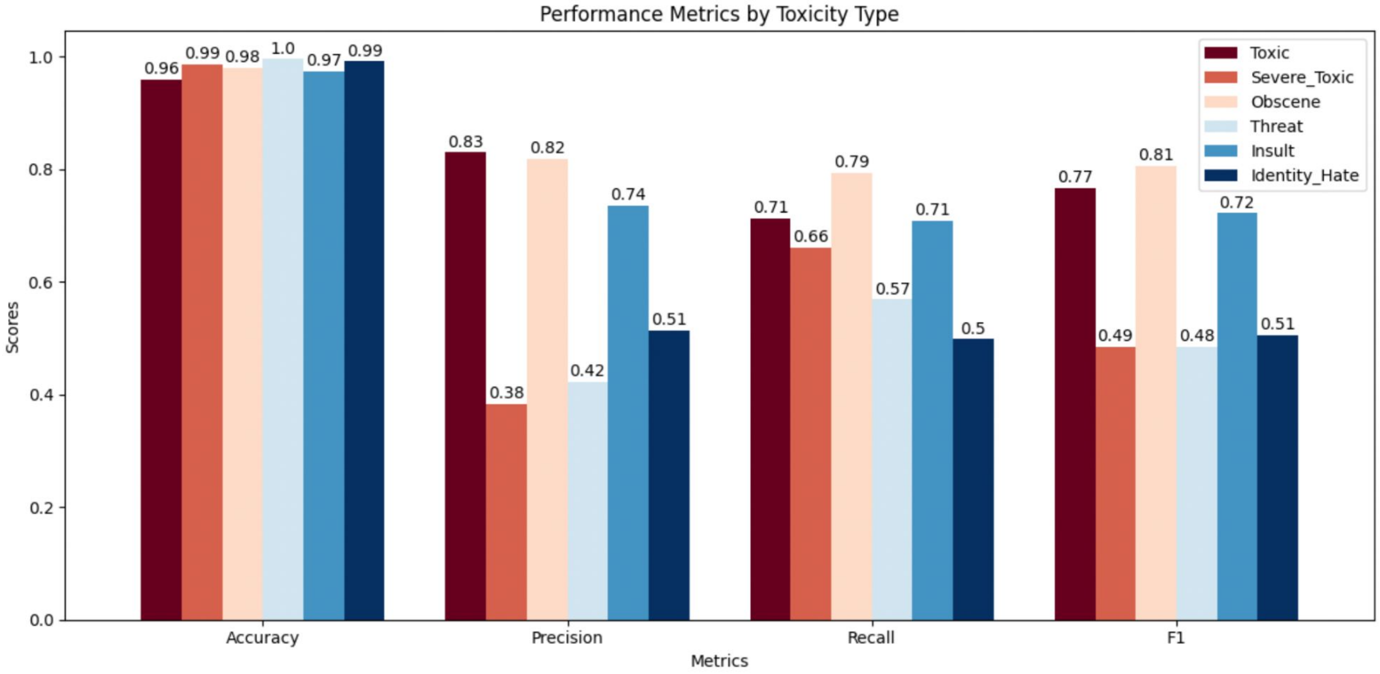
Experiment 3 - CNN + LSTM



Experiment 4 - RNN: BLSTM

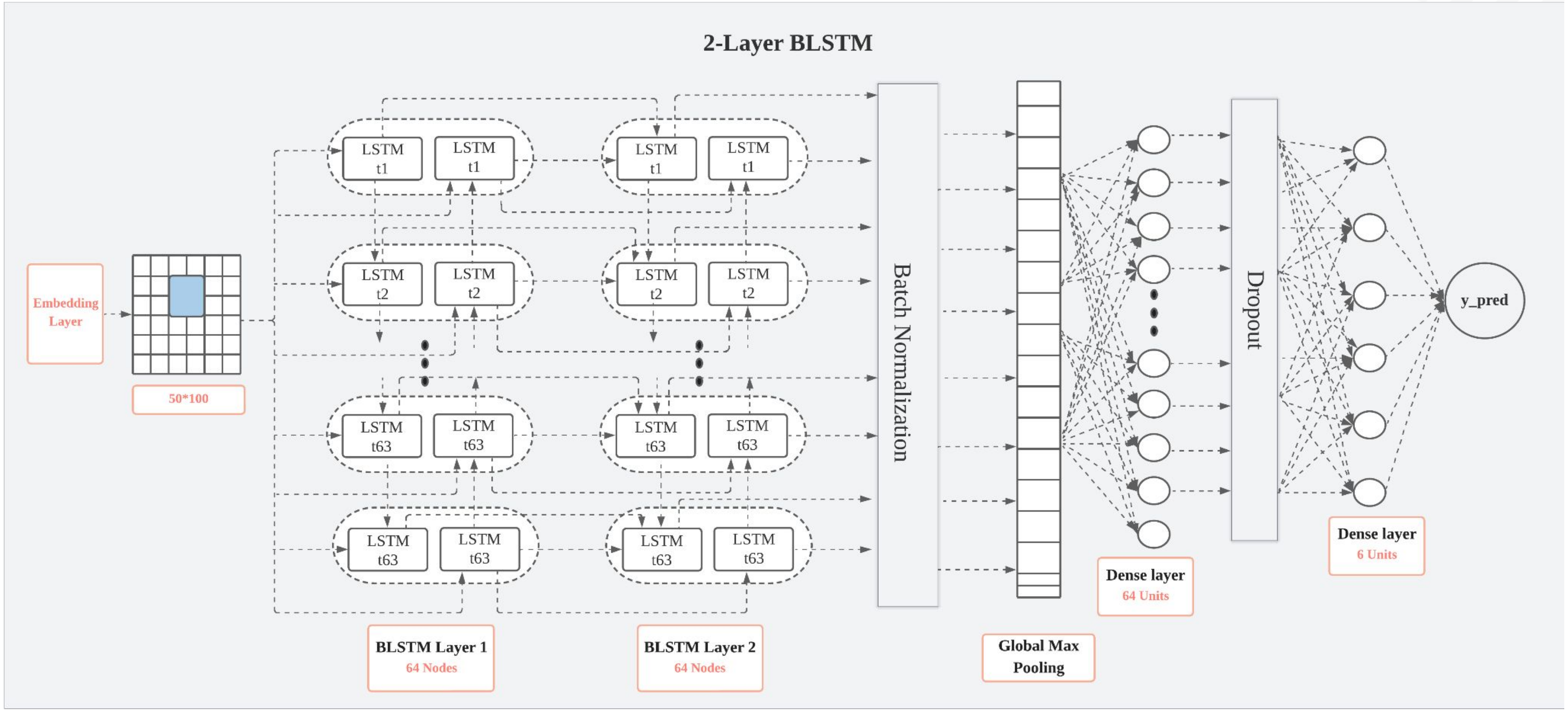


Experiment 4 - RNN: BLSTM

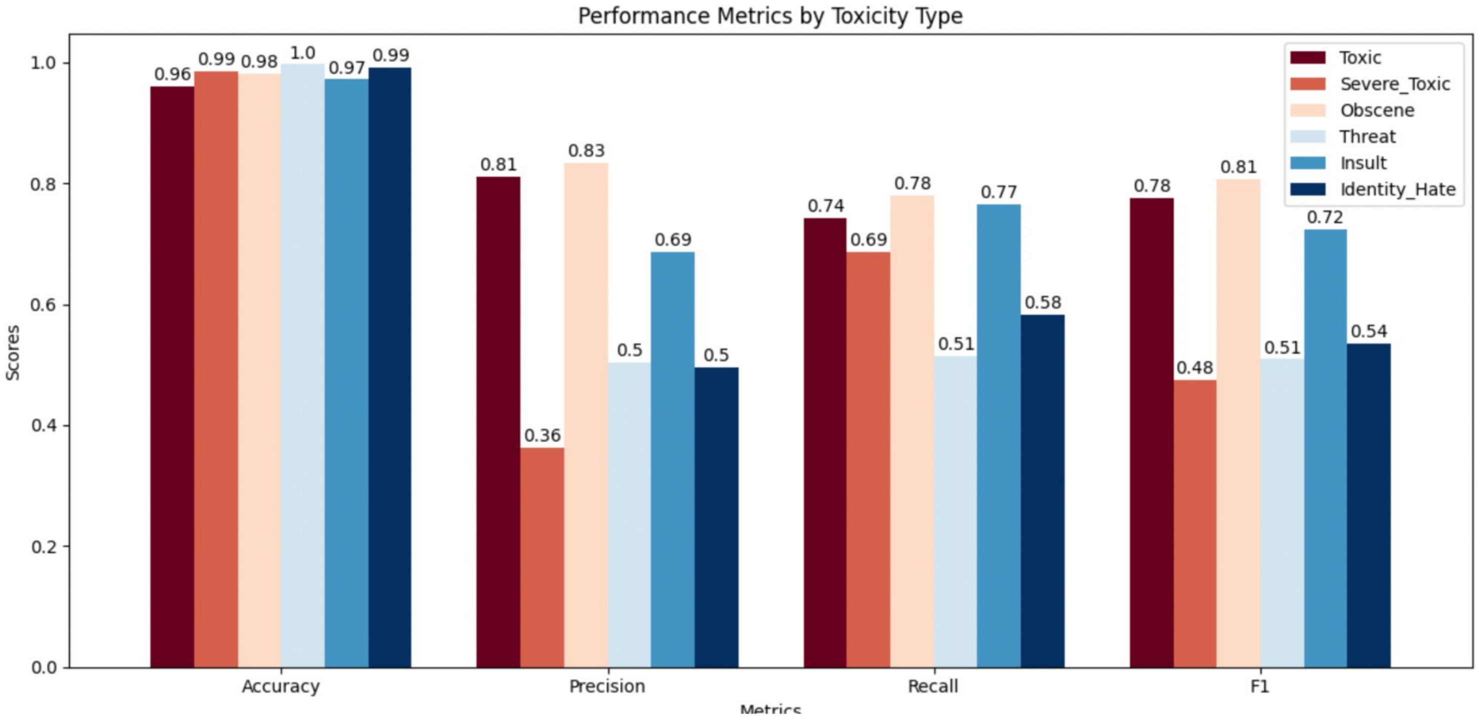


LSTM works really well!
Still have space for progress ?

Experiment 5 - RNN: 2BLSTM

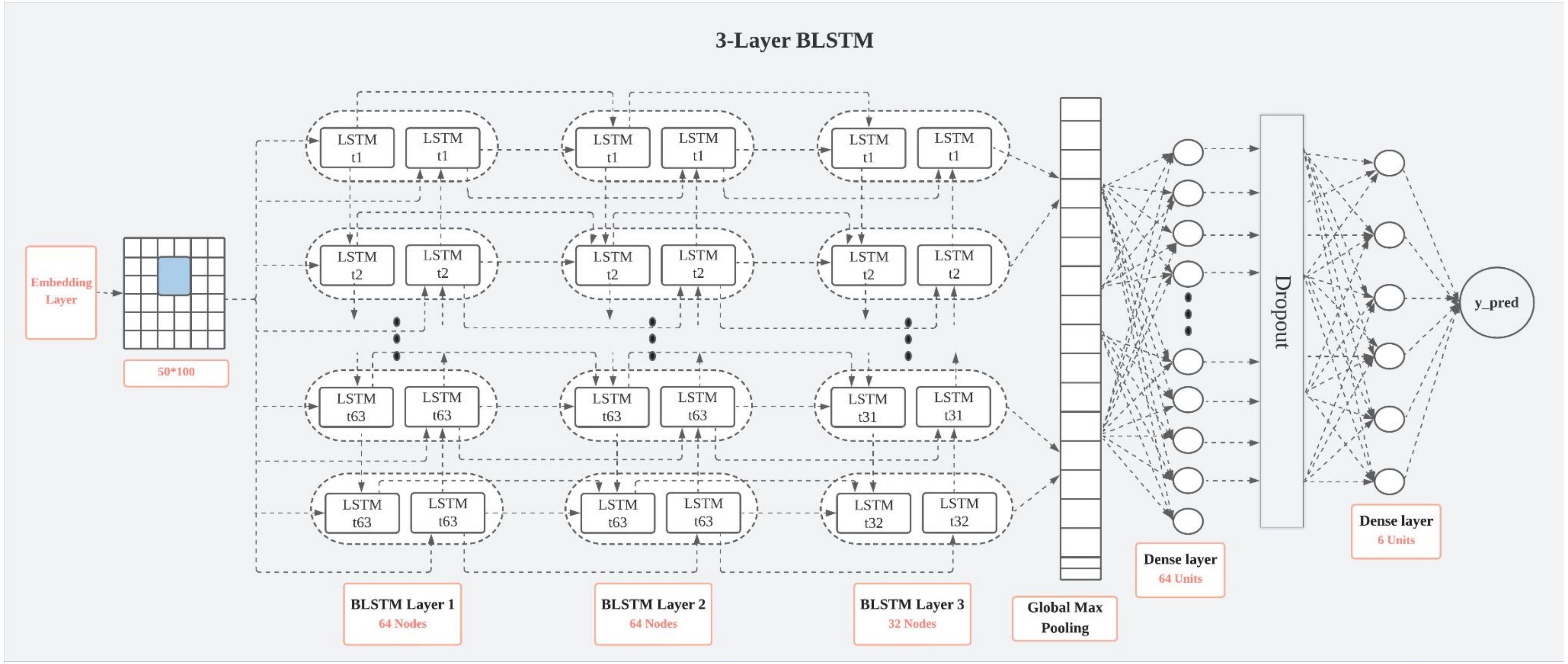


Experiment 5 - RNN: 2BLSTM

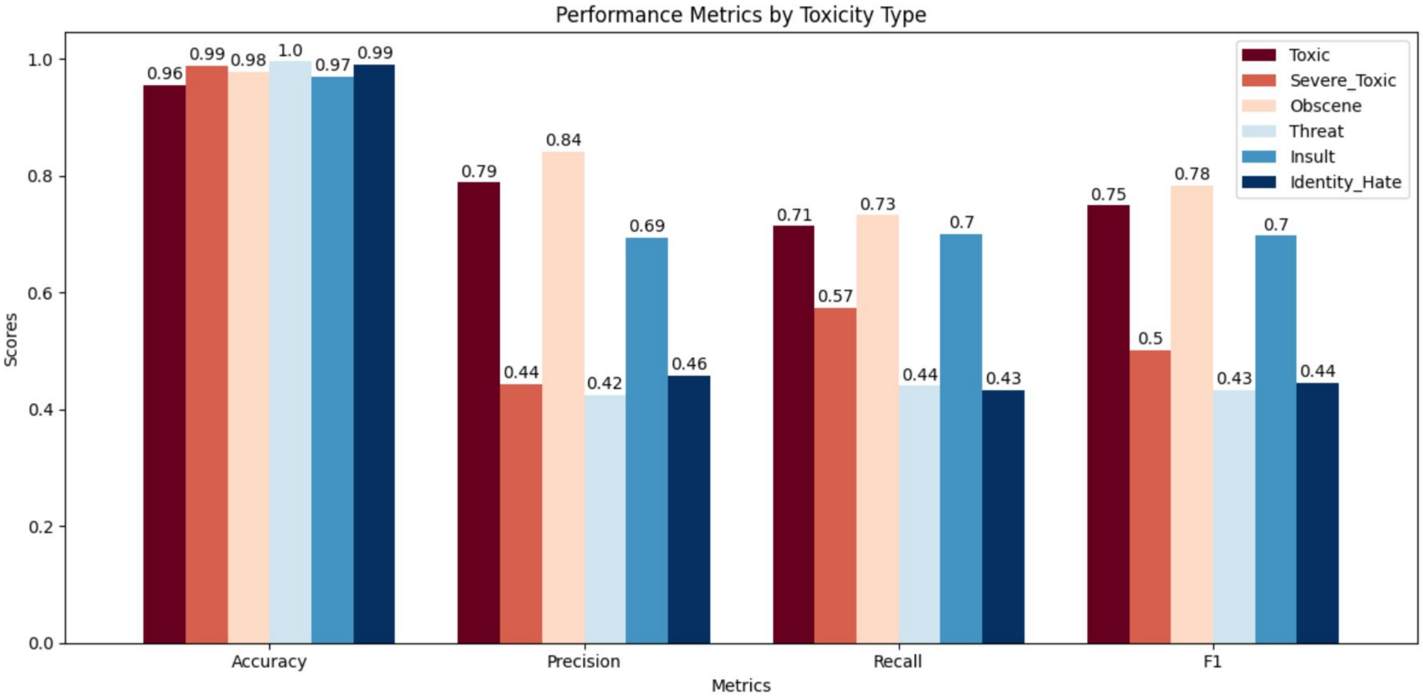


An even better LSTM model!

Experiment 6 - RNN: 3BLSTM



Experiment 6 - RNN: 3BLSTM



A little bit overfitting?

Results: F1 Scores on Test Data

Subtypes	DNN	CNN	CNN + LSTM	RNN: BLSTM	RNN: 2BLSTM	RNN: 3BLSTM
Toxic	0.64	0.63	0.65	0.68	**0.68**	0.66
Severe_Toxic	0.38	0.37	0.38	0.43	**0.43**	0.38
Obscene	0.65	0.66	0.66	**0.69**	0.68	0.68
Threat	0.32	0.39	0.42	0.41	**0.50**	0.37
Insult	0.59	0.59	0.61	0.63	**0.64**	0.62
Identity_Hate	0.37	0.48	0.47	0.52	**0.57**	0.49

2 BLSTM is our best model

Results: Comparison with previous research

Models	Strong Hate	Strong Hate	No Hate
SVM	0.256	0.519	0.757
LSTM	0.097	0.221	0.747

Vigna et al.

Models	Aspect + Sentiment
Pipeline LSTM + fasttext	0.342
End-to-end LSTM + fasttext	0.384
Pipeline CNN + fasttext	0.342
End-to-end CNN + fasttext	0.465

Schmitt et al.

Our models perform well on extremely imbalanced classes!

Findings & Insights

- Our most robust model is the **2-Layer BLSTM** in combination with **batch normalization**.
- **RNN** demonstrated the **best performance**, DNN performed the least well, and CNN gave average results.
- The sequential structure of LSTM allows the model to handle text data effectively by **retaining information over time**. This attribute makes LSTM particularly adept at **managing sequence modeling tasks** such as sentiment analysis.
- The main bottleneck of DNN and CNN models is that sometimes it does not perform well on **underrepresented classes**.
- Adding batch normalization can help improve model performance.
- LSTM models can be prone to overfitting.

Limitations & Future Work

Addressing Class Imbalance:

- Use of Appropriate Evaluation Metrics & Resampling Technique
- For our project, we used F1 scores to evaluate the model performance.
- In the future, we can explore methods of oversampling and undersampling.

RNN LSTM Model Performance:

- Good at detecting minority classes in unbalanced datasets.
- We have not found any research that supports our guess.
- Moving forward, we would like to find out whether the good performance of the RNN LSTM model is due to its strength in learning textual data or whether it indeed has some advantages in detecting minority classes.

Conclusion

- **Toxic Comments Detection:**
  - Most studies are focused on developing a robust model to detect toxic comments in general. The contribution of our study is that we examine subtypes of toxic comments.
- **Minority Classes:**
  - In our experiments, the DNN model has extremely low performance in detecting the threat class and the identity hate class.
- **Classification Thresholds:**
  - We can improve the model performance by adjusting the classification thresholds according to different subtypes.
- **Ethical Consideration:**
  - While moderating toxic content, the models should be cautious not to restrict legitimate discussions, diverse viewpoints, and freedom of expression.

References

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. *Proceedings of the NAACL Student Research Workshop*. https://doi.org/10.18653/v1/n16-2013

Vigna, F. D., Cimino, A., Dell'Orletta, F., & Petrocchi, M. (2017, January). *Hate me, hate me not: Hate speech detection on facebook - researchgate*. researchgate. https://www.researchgate.net/publication/316971988_Hate_me_hate_me_not_Hate_speech_detection_on_Facebook

Schmitt, M., Steinheber, S., Schreiber, K., & Roth, B. (2018). Joint Aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d18-1139

Kraus, M., & Feuerriegel, S. (2019). Sentiment analysis based on rhetorical structure theory:learning deep neural networks from discourse trees. *Expert Systems with Applications*, *118*, 65–79. https://doi.org/10.1016/j.eswa.2018.10.002

THANK YOU!
ANY QUESTIONS?