EasyChair Preprint

№ 2773

Explainable AI approach towards Toxic Comment Classification

Aditya Mahajan, Divyank Shah and Gibraan Jafar

February 27, 2020

Explainable AI approach towards Toxic Comment Classification

Aditya Mahajan
BE-Computer, VIIT, Pune

Divyank Shah
BE-Computer, VIIT, Pune

Gibraan Jafar
BE-Computer, VIIT, Pune

Bhushan Garware, Persistent Systems Limited, Pune
Vaishali Mishra, Assistant Professor, Computer Engineering, VIIT, Pune

Abstract - In this paper, we have documented results and discussed our approach towards solving a serious problem that people face online (i.e. on internet) which is encountering *"toxic"*, *"abusive"*, *"inappropriate"* and *"offensive"* content in the form of textual comments on social media. The purpose of taking up this project topic is to stop *"cyber bullying"* and having a safe online environment. The methodology followed includes data collection from online resources, data preprocessing, converting textual data to vectors (TF-IDF, Word Embeddings), building machine learning and deep learning models, comparing the models using standard metrics as well as interpretability techniques, and thus selecting the best model. After training and evaluating various models, we have come up with a conclusion that standard model evaluation metrics (such as accuracy, precision, recall, f1-score) can often be deceiving as almost every single model we trained gave very good accuracy scores on the test set. After using a model-interpretability technique like LIME and checking out some of the explanations that the models generated on a common set of comments we created manually, we noticed that some of the models were considering incorrect words for a sentence to be predicted as toxic. So even with a high accuracy or any evaluation score, we can't deploy such models in real world scenarios. The combination that gave us the best result in our study is Gated Recurrent Unit (GRU) + pre-trained word embeddings with a high accuracy score along with intuitive LIME explanations. This paper provides a comparative study of various machine learning and deep learning models in Toxic Comment Classification. Through this project and study we also want to emphasize that model interpretability techniques (like LIME, etc.) are pivotal while selecting the best model for any ML/DL project and solutions as well as establishing the trust of the end user on the deployed model.

Keywords - toxic comment classification, machine learning, deep learning, explainable AI, LIME, interpretability

I. INTRODUCTION

Toxicity is a major problem that people face when they post and share photos, videos, opinions, etc. online publicly on the internet. The issue of online bullies, trolls [5] and spammers who post and comment inappropriate and offensive content on social media is becoming increasingly prevalent. This is mainly due to the fact that people can hide their true identities on social media websites by having fake names and profile pictures. There have been several cases where, because of this online toxicity, people had to deactivate their social media accounts or had to disable comments on their photos or posts. Also there have been extreme cases which have even resulted in loss of an innocent life. A few notable ones are as follows:-

1. Manchester United Manager Ole Gunnar Solskjaer deactivated his Twitter Account after being subjected to online abuse due to poor results. Similarly, Chelsea full back Marcos Alonso was also subjected to online abuse after delivering poor performances in matches. Because of this constant negativity, he had to disable comments on his instagram posts.

2. Megan Meier (1992–2006), age 13, from Dardenne Prairie, Missouri, died of suicide by hanging three weeks before her fourteenth birthday. After prompting an investigation, it came to light that cyberbullying and online harassment through the social networking website Myspace was the main reason behind the suicide.

   Several such cyberbullying cases can be found on this wikipedia article -

   https://en.wikipedia.org/wiki/List_of_suicides_that_have_been_attributed_to_bullying.

Victims of cyberbullying [1] and online harassment tend to experience low confidence, low self-esteem, increased suicidal ideation, and a wide range of negative emotions such as depression, frustration, anger and even fear.

Content Moderation [12] is a process of flagging user-submitted content (in our case textual comments) according to certain rules or protocols to determine whether a submission is acceptable or not. We live in a time period where content posted online has the potential to cause damage and influence the minds of young children. Also, in general, people prefer having positive and socially acceptable content on their social media timeline/wall, profiles, posts, etc instead of negative, abusive or toxic content. Content moderation is essential to have a safe and comfortable online experience. It helps in protecting the users from being the victims of inappropriate and abusive content. Moreover, it also helps in protecting brand's reputation and image even if users post undesirable content.

Large amounts of comments and tweets are posted online on a day-to-day basis, so to track and flag inappropriate and toxic comments manually with the help of moderators becomes inefficient. In such scenarios, automated moderation can be used to flag/block the occurrences of toxic and inappropriate comments.

Through this study we wanted to check the performance of various machine learning and deep learning models to determine whether a comment is toxic or not. Which preprocessing techniques and models would work the best in combination with one another. And whether interpretability techniques like LIME can help us in selecting the best models and whether it can also help the end users *"trust"* the model.

Ⅱ. Model Interpretability Through LIME

Machine Learning and deep learning models work great when it comes to prediction and classification problems. But these models do not explain the reason behind a certain prediction. Hence it becomes difficult for end-users to *"trust"* the predictions made by these models.

To explain the reason and cause of prediction, concept of model interpretability and explainable AI was introduced. Interpretability is the extent to which a human can understand the cause of a decision made by a black box model. [8]

LIME [10] is a model agnostic interpretability technique used to explain individual local predictions. Model agnostic means that it can be applied for any machine learning/deep learning model to generate explanations. The general working principle for LIME is that it localises the problem and explains the problem at that locality instead of generating explanation for the whole model.

LIME Algorithm [6]:

1. To find an explanation for a single data point and a given classifier, sample the locality around the selected single data point uniformly and at random.
2. Generate a dataset of perturbed data points with its corresponding prediction from the model we want to be explained.
3. Use the specified feature selection methodology to select the number of features that is required for explanation.
4. Calculate the sample weights using a kernel function and a distance function. (this captures how close or how far the sampled points are from the original point)
5. Fit an interpretable model on the perturbed dataset using the sample weights to weigh the objective function.
6. Provide local explanations using the newly trained interpretable model.

III. Approach / Methodology

a. Data Collection

We started our study with data collection for which 3 online sources were used -

- Jigsaw Toxic Comment Classification Challenge (hosted on Kaggle) [11]
- Hate Speech Dataset (hosted on Github) [4]
- Sexual Abusive YouTube Comments (hosted on Zenodo) [2]

The Jigsaw Toxic Comment Classification Challenge data contained comments from Wikipedia and is labelled into six categories (toxic, severe_toxic, obscene, threat, insult, identity_hate).

The Hate Speech Dataset hosted on Github contained tweets labelled into 3 categories (hate speech, offensive language, neither).

The third dataset contained sexually abusive youtube comments posted on viral youtube videos.

We converted all the labels in these datasets into a binary classification :

‘0’ label - indicating a non- toxic comment

‘1’ label - indicating a toxic comment.

b. Data Cleaning and Merging

Once the dataset labels were converted into proper format, we cleaned the comments in the dataset by removing all the unnecessary symbols, URLs, mentions, etc. using regular expressions. Stopword removal and word lemmatization was done using spaCy library.

Finally after cleaning all the comments, the three datasets were merged into a single big dataset.

c. Converting textual data into vectors

To represent each comment as a series of features we used 2 approaches :

- Term Frequency-Inverse Document Frequency (TF-IDF)

  TF-IDF score indicates how relevant a word is to a comment in a collection of comments. [13] We had to switch from TF-IDF to word embeddings for reasons described in subsequent sections.
- Pretrained Word Embeddings

  Word embedding is a feature learning technique in which words in the vocabulary are mapped to vectors of real numbers. [14] We used two famous pretrained embeddings available online -
  - GloVe (Global vectors for word representation) [9]
  - Fasttext [3]

d. Model Training, Evaluation, Interpretation and Selection process

In this step various models (listed below) were trained along with their corresponding preprocessing steps. Hyper-parameter tuning was done to get maximum accuracy. Then the models were then compared using two approaches -

- Standard model evaluation metrics such as accuracy and f1-score,
- LIME explanations.

The reason for including LIME explanations in model comparison and selection process is because, as explained earlier, machine learning and deep learning models are black box models so we cannot rely on a single metric like accuracy or f1-score to determine how trustworthy a model is. It is equally important to understand the reason behind a particular prediction being made.

TABLE I

Preprocessing techniques used and Models trained for Toxic Comment Classification

Sr No.	Model Name	Preprocessing Technique used
1.	Logistic Regression	TF-IDF
2.	Multinomial Naive Bayes Classifier	TF-IDF
3.	Random Forest Classifier	TF-IDF
4.	Xg-boost Classifier	TF-IDF

5.	CNN	Glove Embeddings (100d)
6.	GRU	Glove Embeddings (100d)
7.	GRU	Fasttext Embeddings (300d)

IV. Results and Discussion

TABLE II

Accuracy and f1-scores observed for the trained models

Sr No.	Model Name	Preprocessing Technique used	Accuracy Score	F1-Score obtained for toxic prediction
1.	Logistic Regression	TF-IDF	0.96	0.90
2.	Multinomial Naive Bayes Classifier	TF-IDF	0.94	0.82
3.	Random Forest Classifier	TF-IDF	0.93	0.80
4.	Xg-boost Classifier	TF-IDF	0.96	0.88
5.	CNN	Glove Embeddings (100d)	0.95	0.87
6.	GRU	Glove Embeddings (100d)	0.96	0.90
7.	GRU	Fasttext Embeddings (300d)	0.96	0.90

As we can see the accuracy and f1-score (for toxic prediction) for all the above models are significantly good (> 0.90 accuracy) and very close to each other. Selection and rejection of the models based on the standard evaluation metrics like accuracy and f1-score is difficult. Thus we decided to try some sentences manually and compare their LIME explanations.

Some of the comments we tried out manually were -

1. "Well done dude!" (not-toxic comment)
2. "Dude, you might have done it wrong." (not-toxic comment)
3. "You are so stupid. Can't even take a proper photo." (toxic comment)
4. "That girl is a s t." (toxic comment)
5. "These motherf s will pay. Time to f g kill them." (toxic comment)
6. "You are truly a horrible person." (toxic comment)
7. "You are a piece of s t" (toxic comment)
8. "Its p n, let us f p in peace. He already blew his shot." (toxic comment)
9. "This is nothing but a p n video." (toxic comment)
10. "Your opinion is bulls t and should be ignored." (toxic comment)
11. "Do you know you come across as a giant pr k?" (toxic comment)

12. “Your life will be terminated” (toxic comment)

(Note:- Extremely toxic and inappropriate words have been censored out here in this paper, but were passed explicitly to the models for getting results.)

Lime Explanations for Logistic Regression -

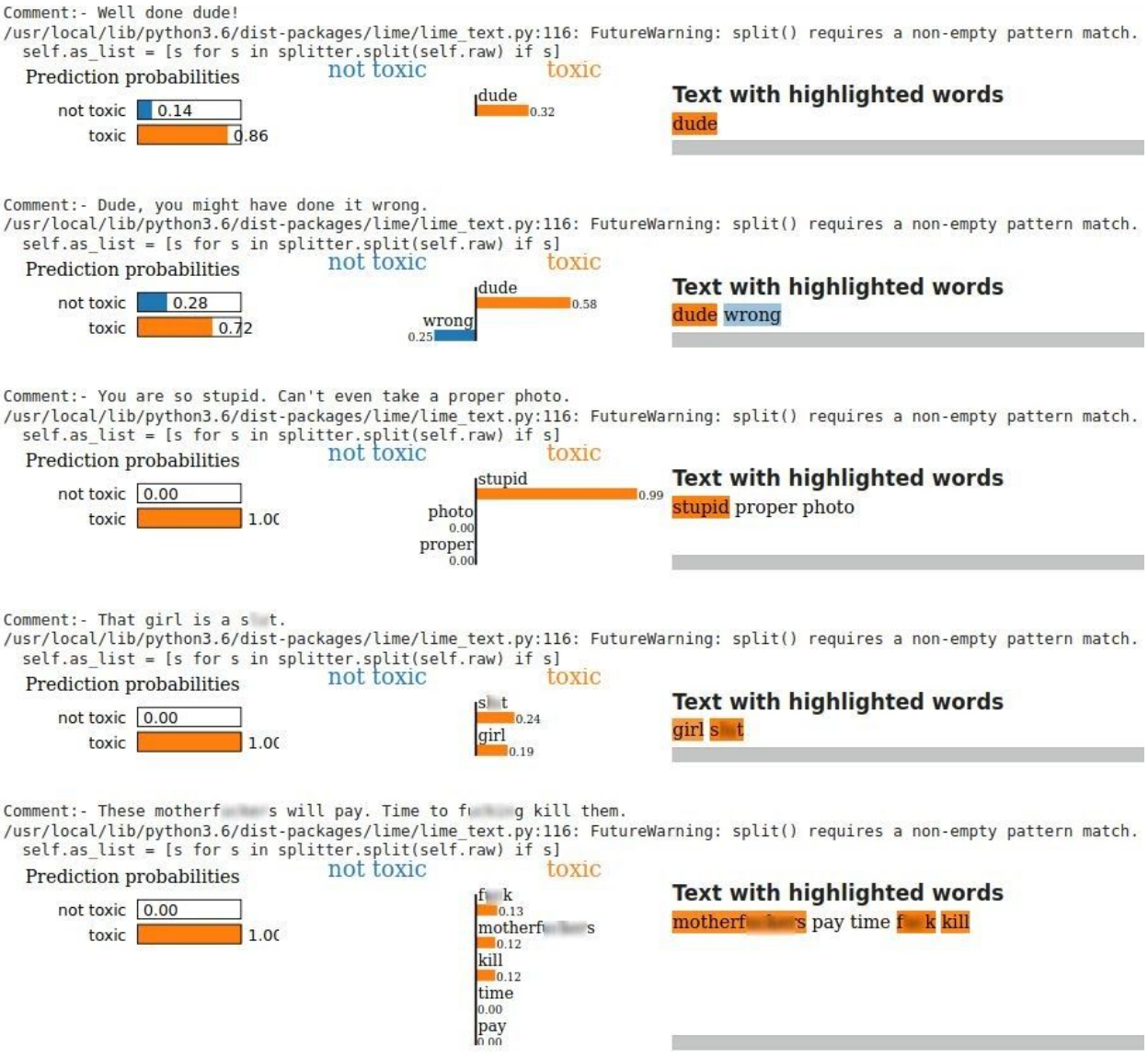


Figure 1: Explanations of Logistic Regression

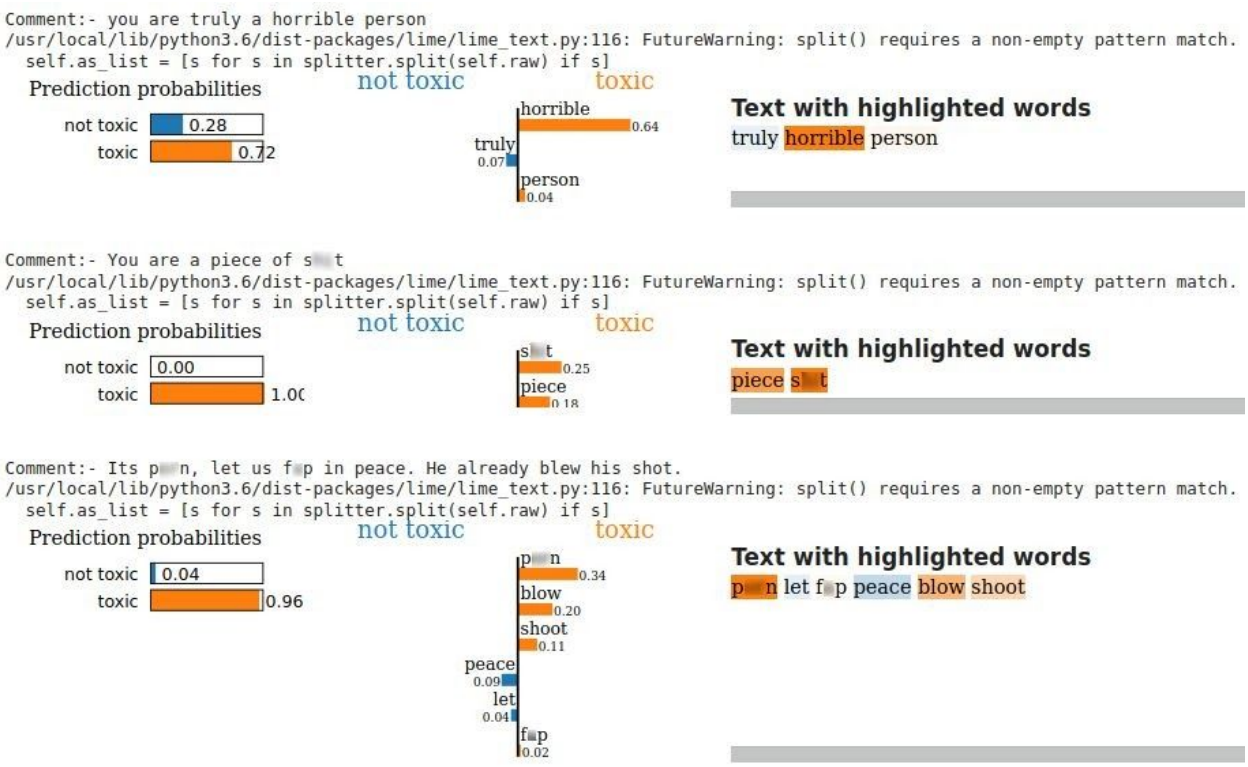


Figure 2: Explanations of Logistic Regression

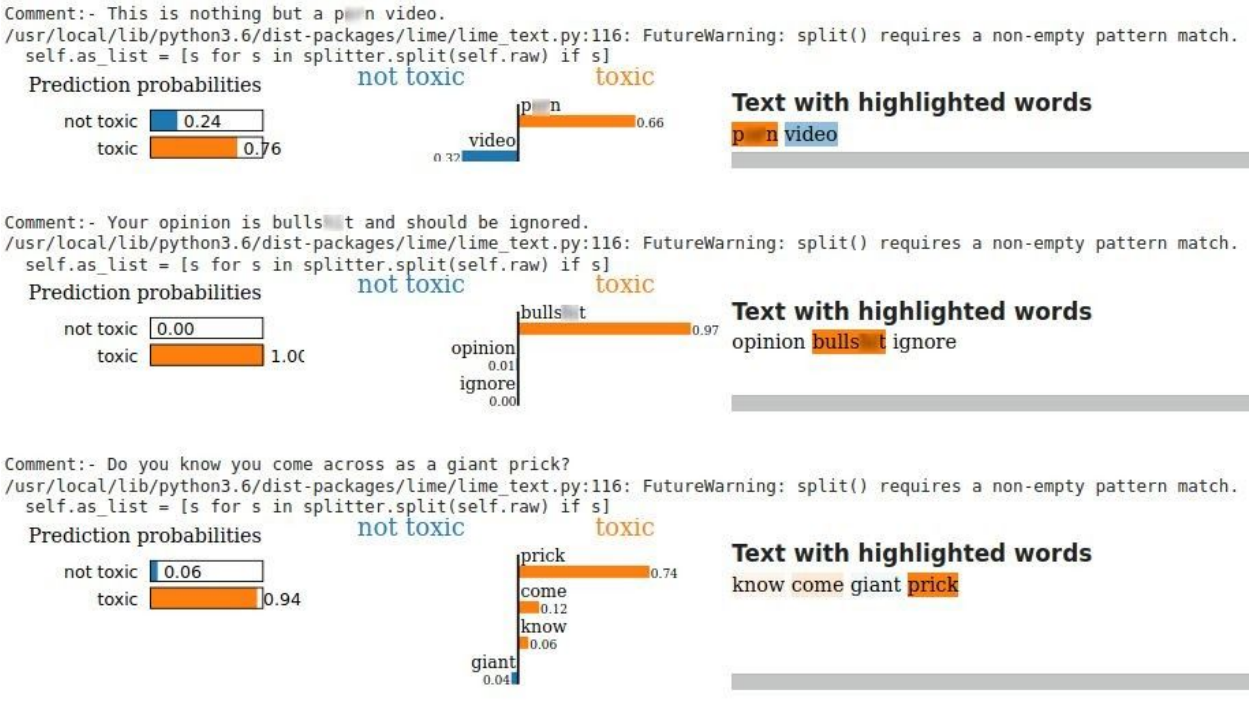


Figure 3: Explanations of Logistic Regression

When we tried out these sentences we observed that Logistic regression was giving incorrect predictions for sentences like *"Well done dude!"*, "*Dude, you might have done it wrong*.". When we observed the explanations behind these incorrect predictions using LIME, we found out that words like *"dude"* were shifting the prediction to the toxic side. As these words are very commonly used in comments, we had to reject the Logistic Regression model even though it gave us an accuracy score of 0.96 and a f1-score of 0.90.

Lime Explanations for Multinomial Naive Bayes -

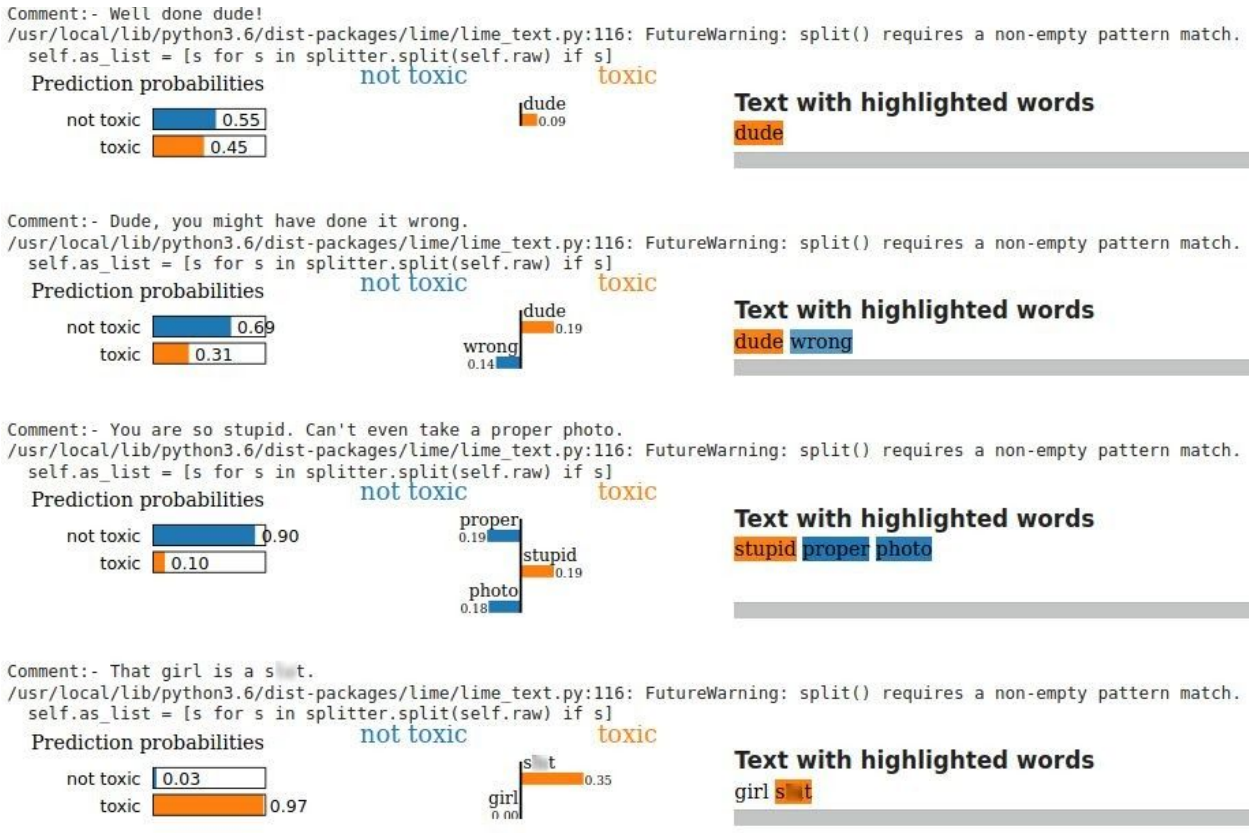


Figure 4: Explanations of Multinomial Naive Bayes

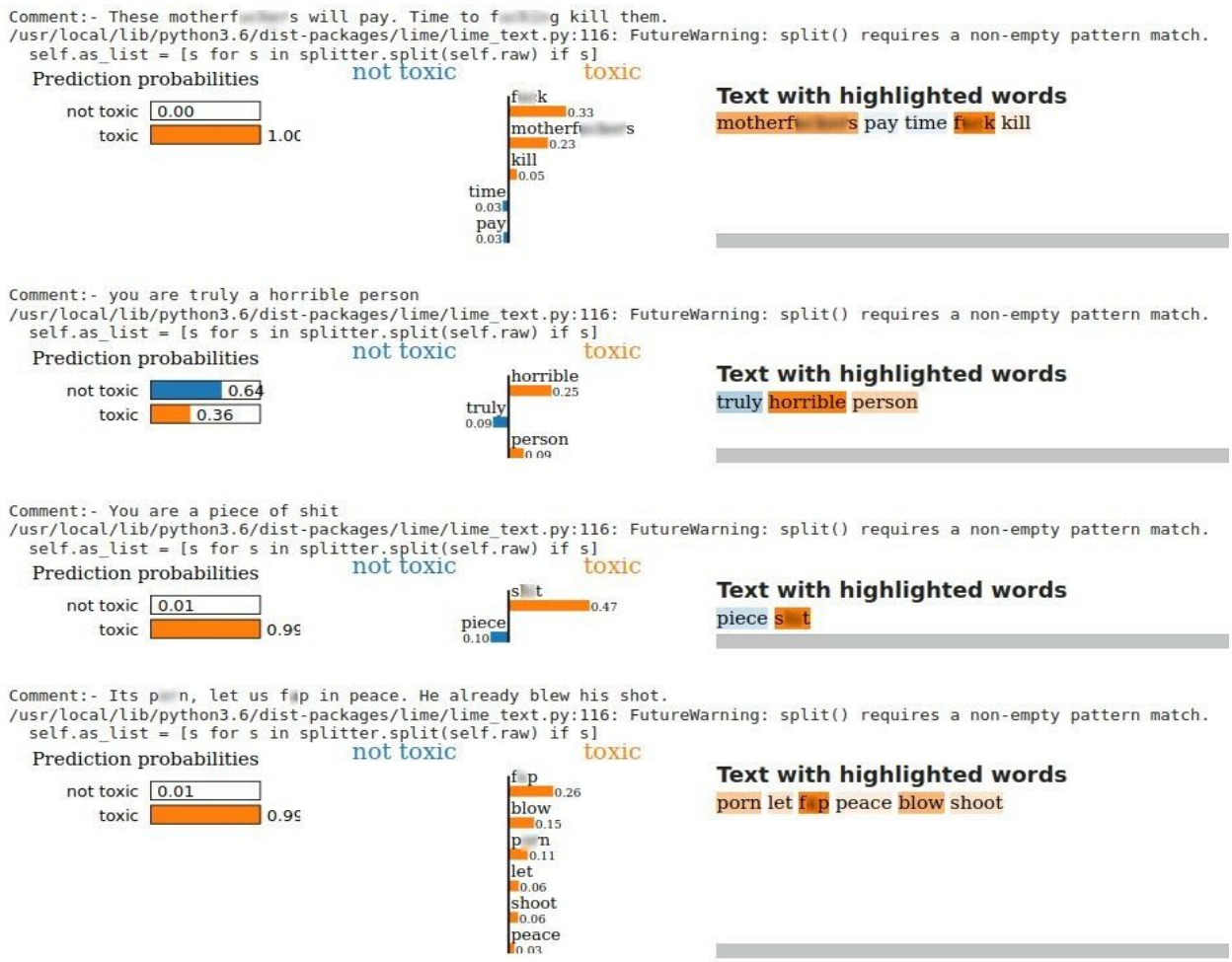


Figure 5: Explanations of Multinomial Naive Bayes

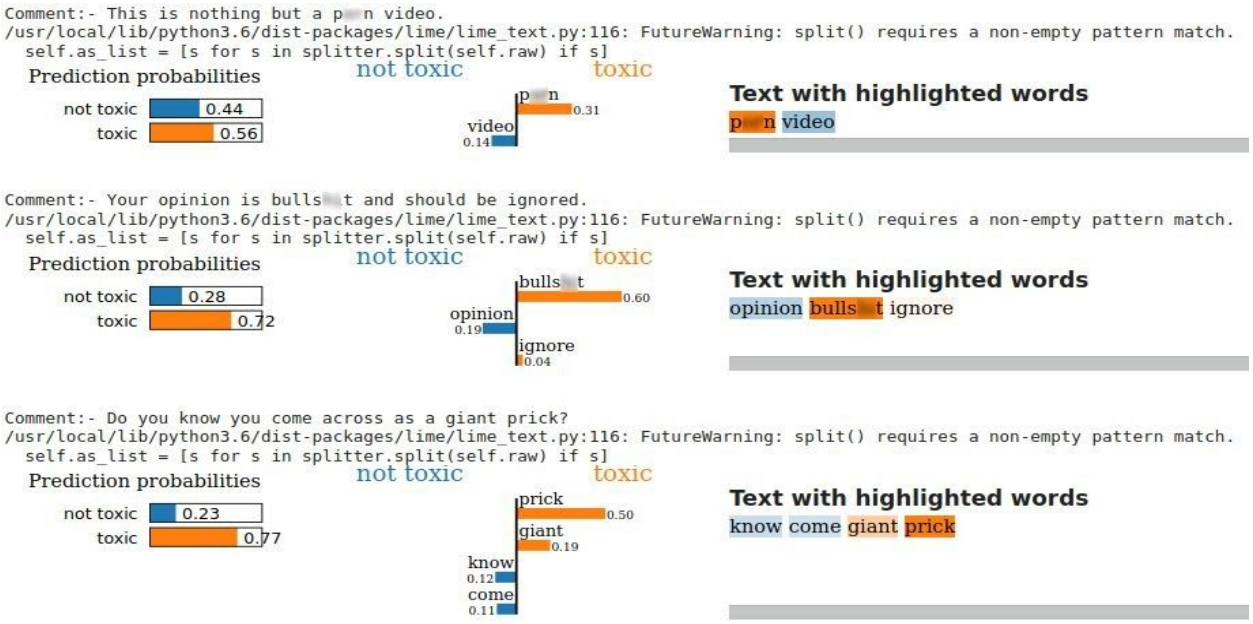


Figure 6: Explanations of Multinomial Naive Bayes

When we tried out the same sentences for Multinomial Naive Bayes it was giving incorrect predictions for sentences like *"You are so stupid. Can't even take a proper photo.","You are truly a horrible person."*. In our study for Naive Bayes f1-score obtained was also very less in comparison to other models. Hence we had to reject Multinomial Naive Bayes.

Lime Explanations for Random Forest -

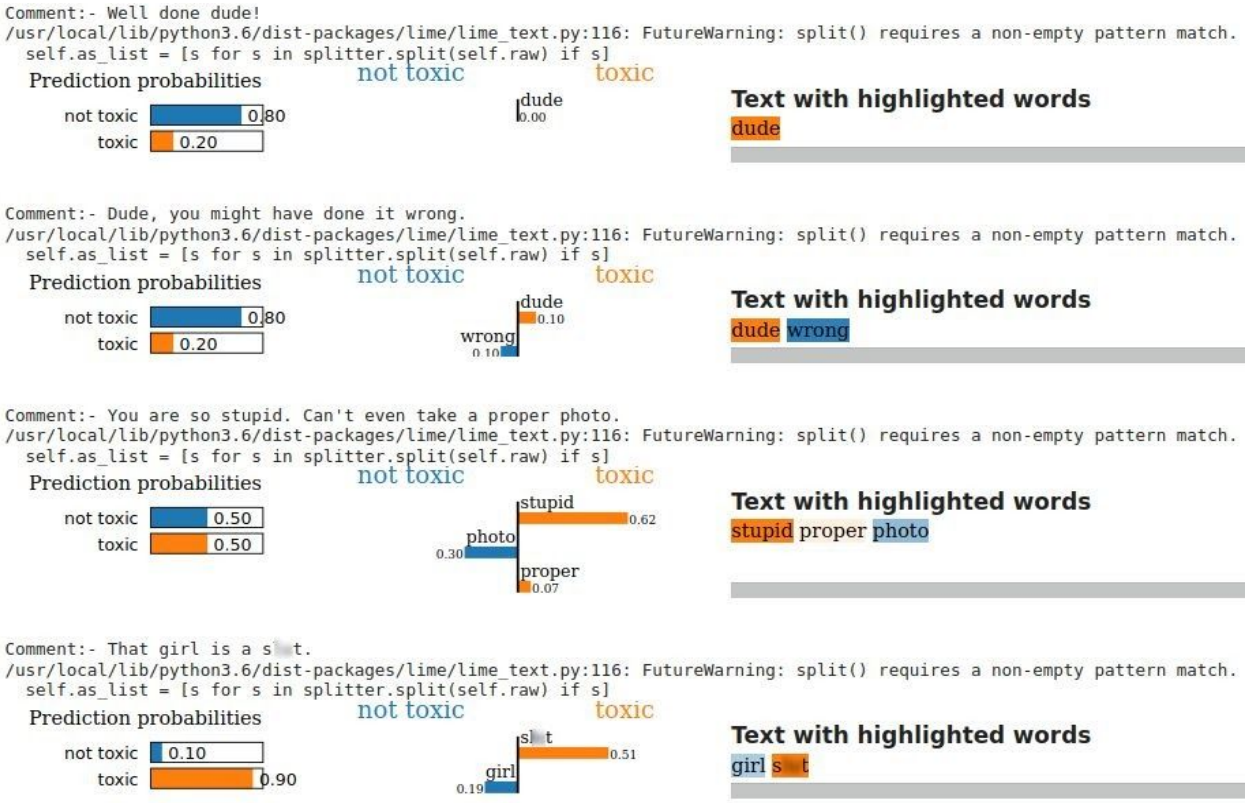


Figure 7: Explanations of Random Forest

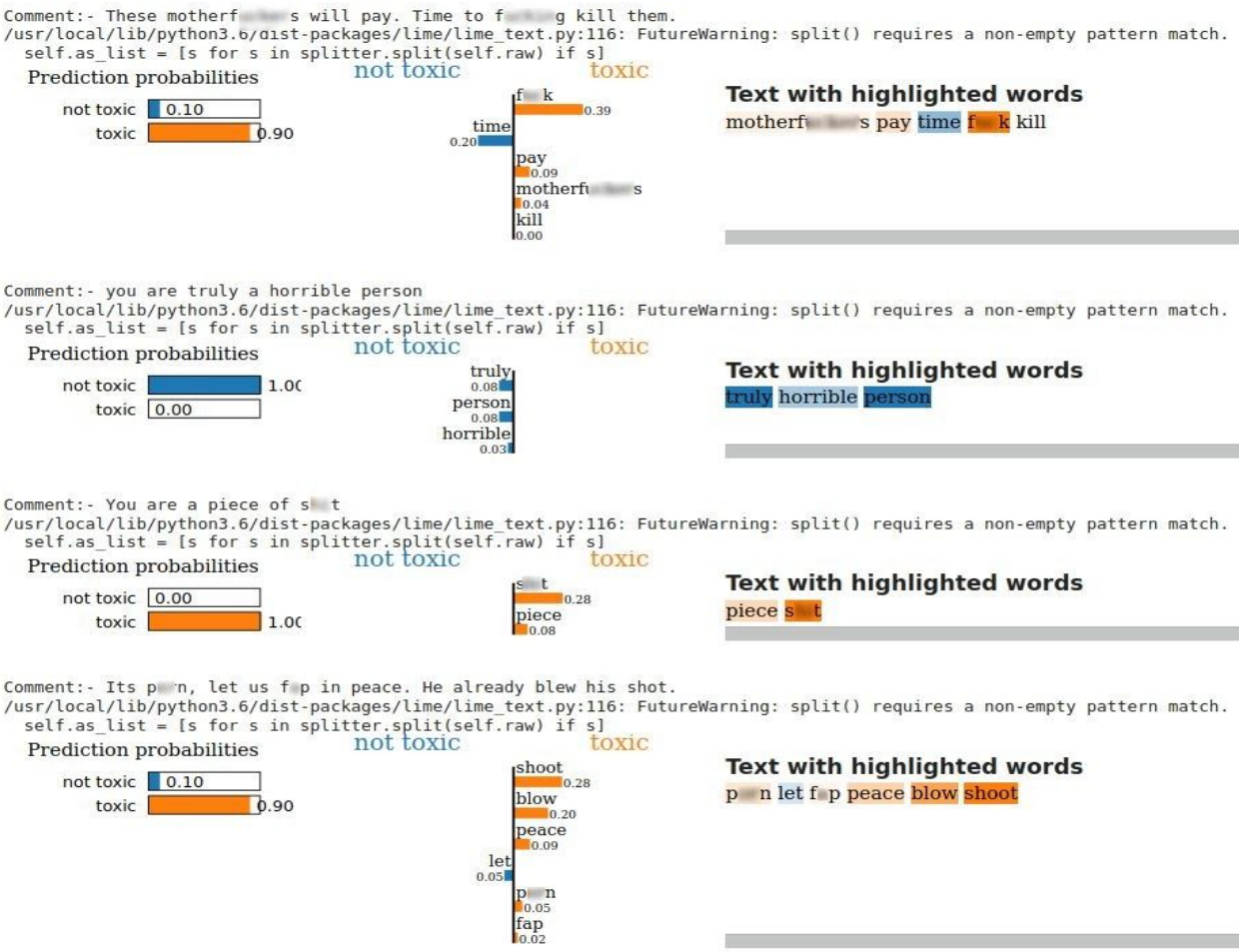


Figure 8: Explanations of Random Forest

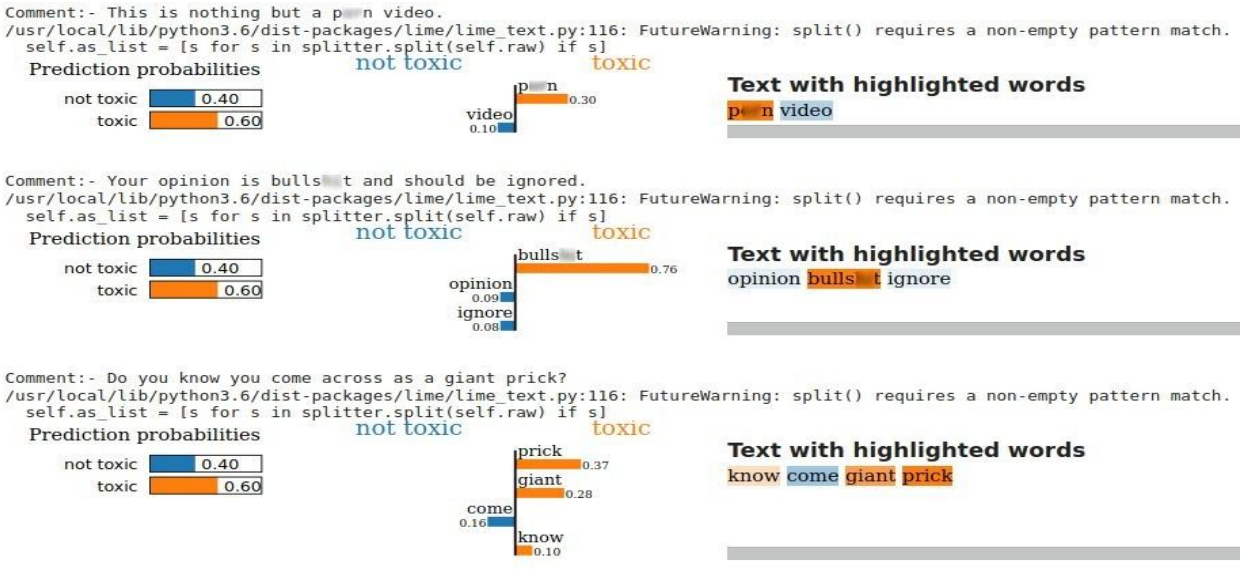


Figure 9: Explanations of Random Forest

In-spite of being an ensemble technique , Random Forest [RF] also suffers from similar weaknesses as previous models . For sentences such as "*You are so stupid. Can't even take a proper photo.*","*You are truly a horrible person*.","*This is nothing but a p[redacted]n video.*" the model fails to classify them correctly . In the sentence "*You are truly a horrible person*" we observed that words like *"horrible"* are shifting the prediction to the non-toxic side. This tells us that the model cannot be trusted which is why we rejected it.

Lime Explanations for XgBoost -

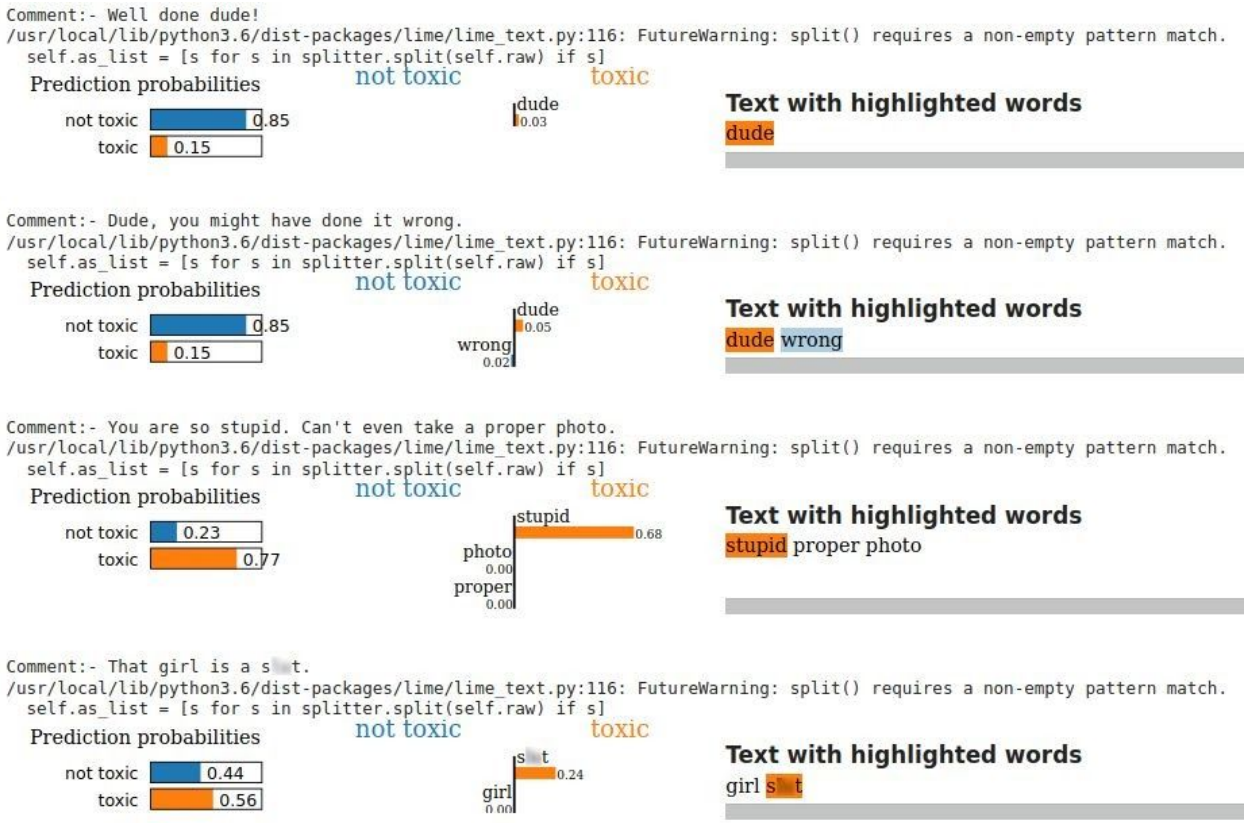


Figure 10 : Explanations of Xg-boost

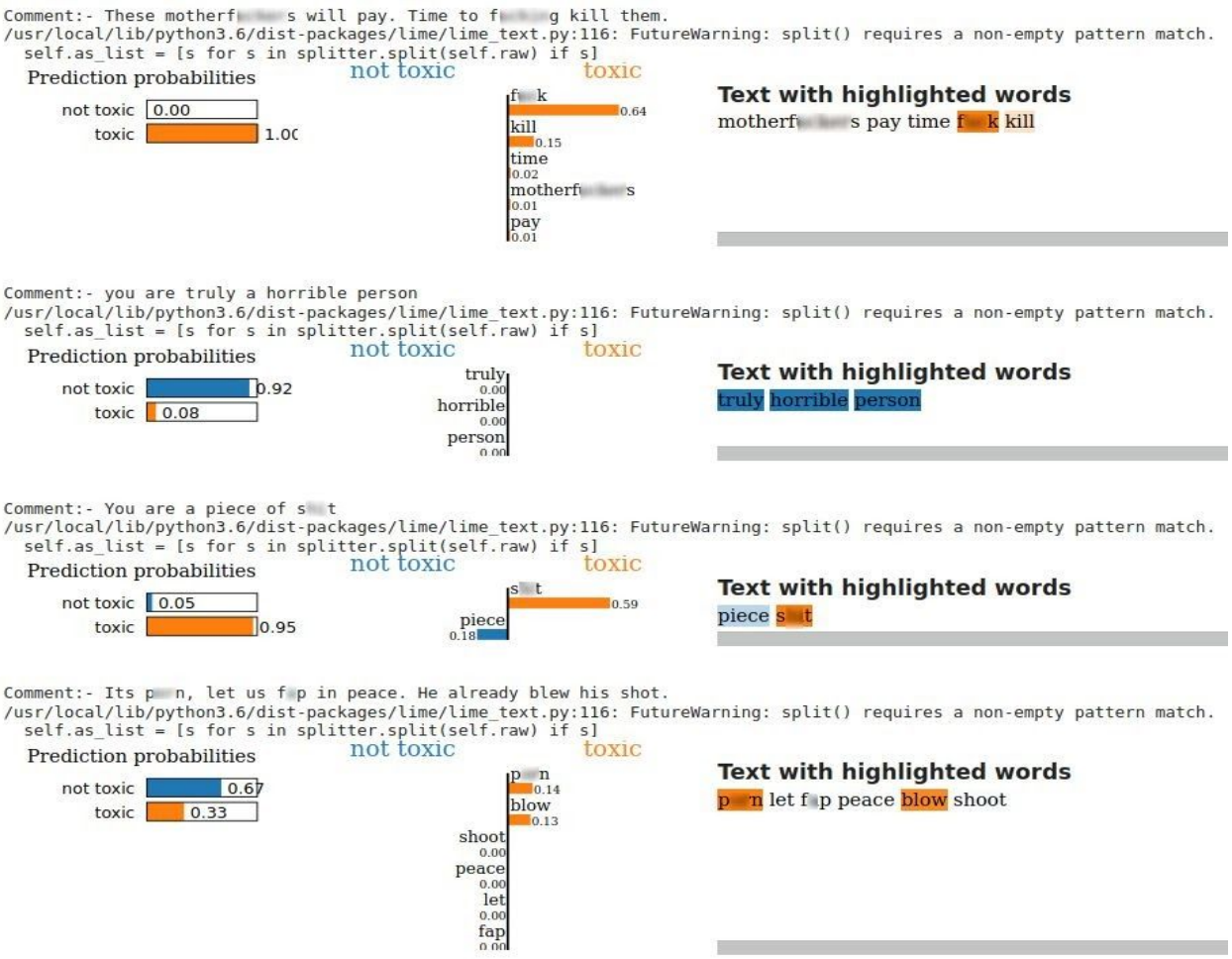


Figure 11 : Explanations of Xg-boost

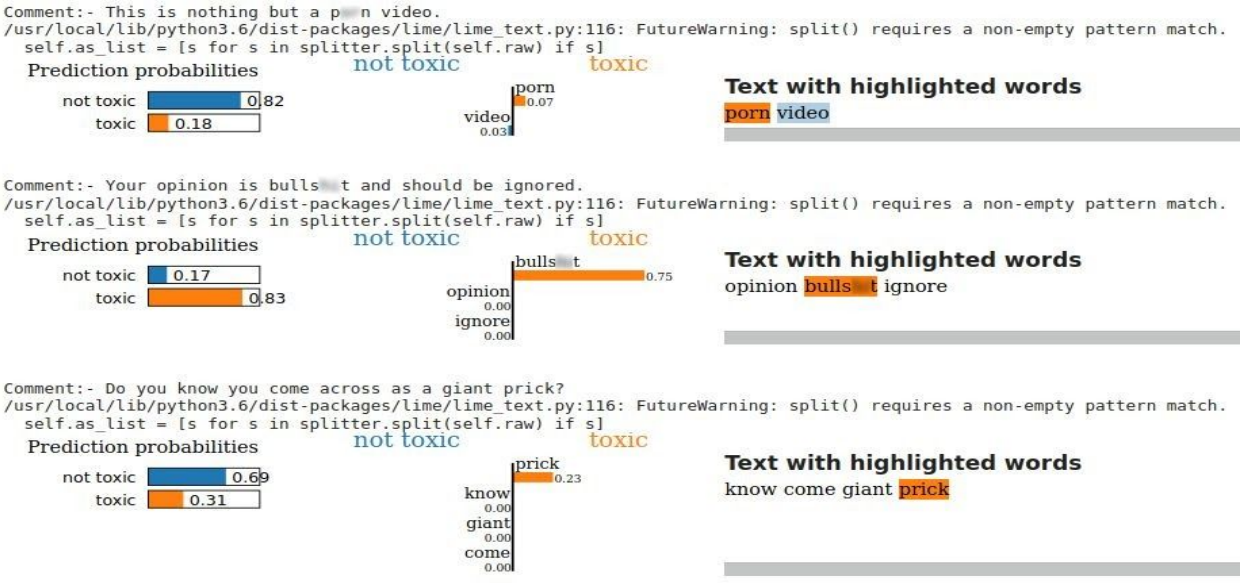


Figure 12 : Explanations of Xg-boost

Xg-boost gave us the best accuracy as well as f1-score as compared to other traditional machine learning models with TF-IDF but when we checked out the explanations we noticed that some inappropriate sentences having words like *"p[redacted]n"* were being classified as non-toxic. Also in the sentences "*You are truly a horrible person.*","*Do you know you come across as a giant pr[redacted]k?*", "*Its p[redacted]n, let us f[redacted]p in peace. He already blew his shot.*" words like "*horrible*", "*f[redacted]p*" were shifting the prediction to non-toxic.

The probable reason behind these traditional machine learning models giving inaccurate results can be that TF-IDF doesn't consider the position of words in text as well as the semantic information. Hence we shifted to word embeddings which are commonly used with deep learning models.

Lime Explanations for CNN + Glove Embeddings -

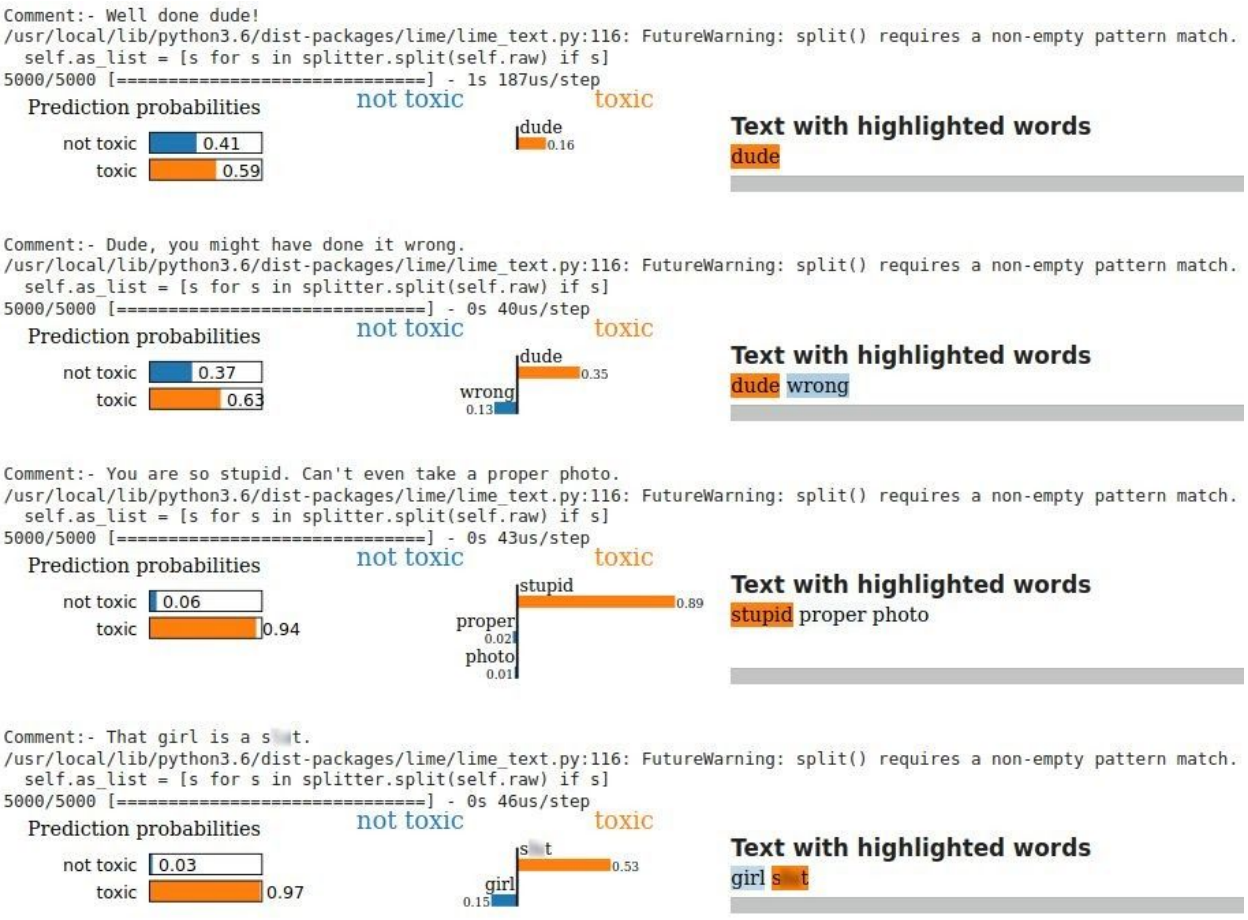


Figure 13 : Explanations of CNN + Glove

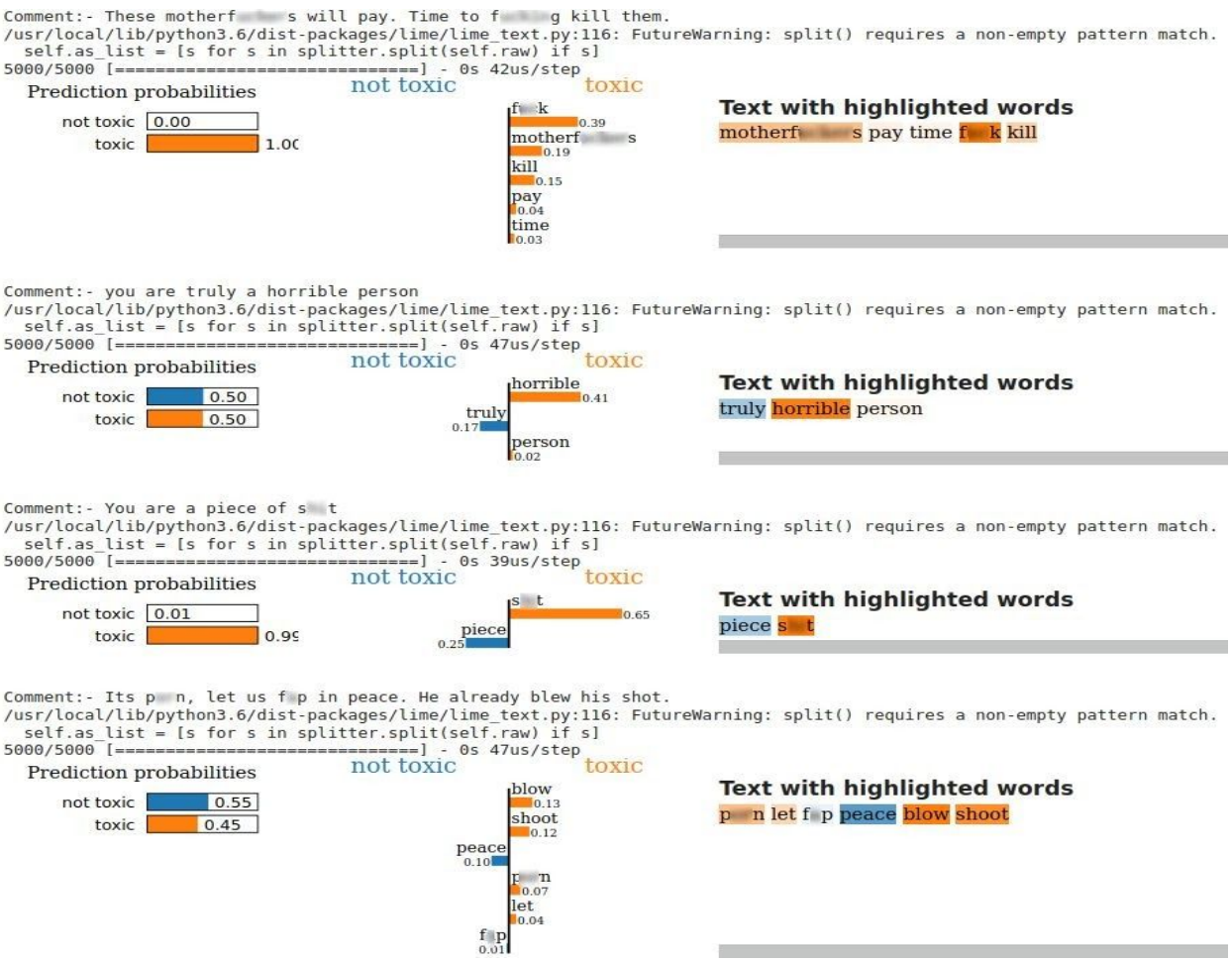


Figure 14 : Explanations of CNN + Glove

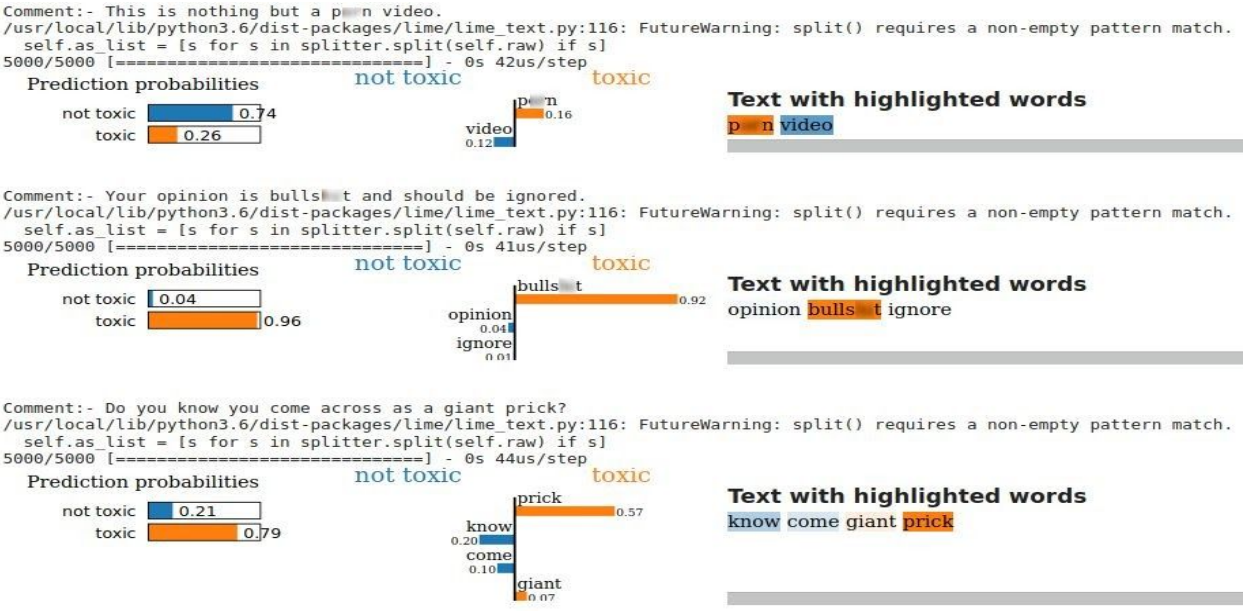


Figure 15 : Explanations of CNN + Glove

CNN also makes some mistakes similar to previous models even after being used with Glove Embeddings. It incorrectly classifies comments in the case of "*Well done dude*", "*Dude, you might have done it wrong*", *"This is nothing but a p[redacted]n video."* and *"Its p[redacted]n, let us f[redacted]p in peace. He already blew his shot."* The LIME explanation tells us that the words like *"f[redacted]p"* and *"dude"* are inappropriately interpreted.

Lime Explanations for GRU + Glove Embeddings -

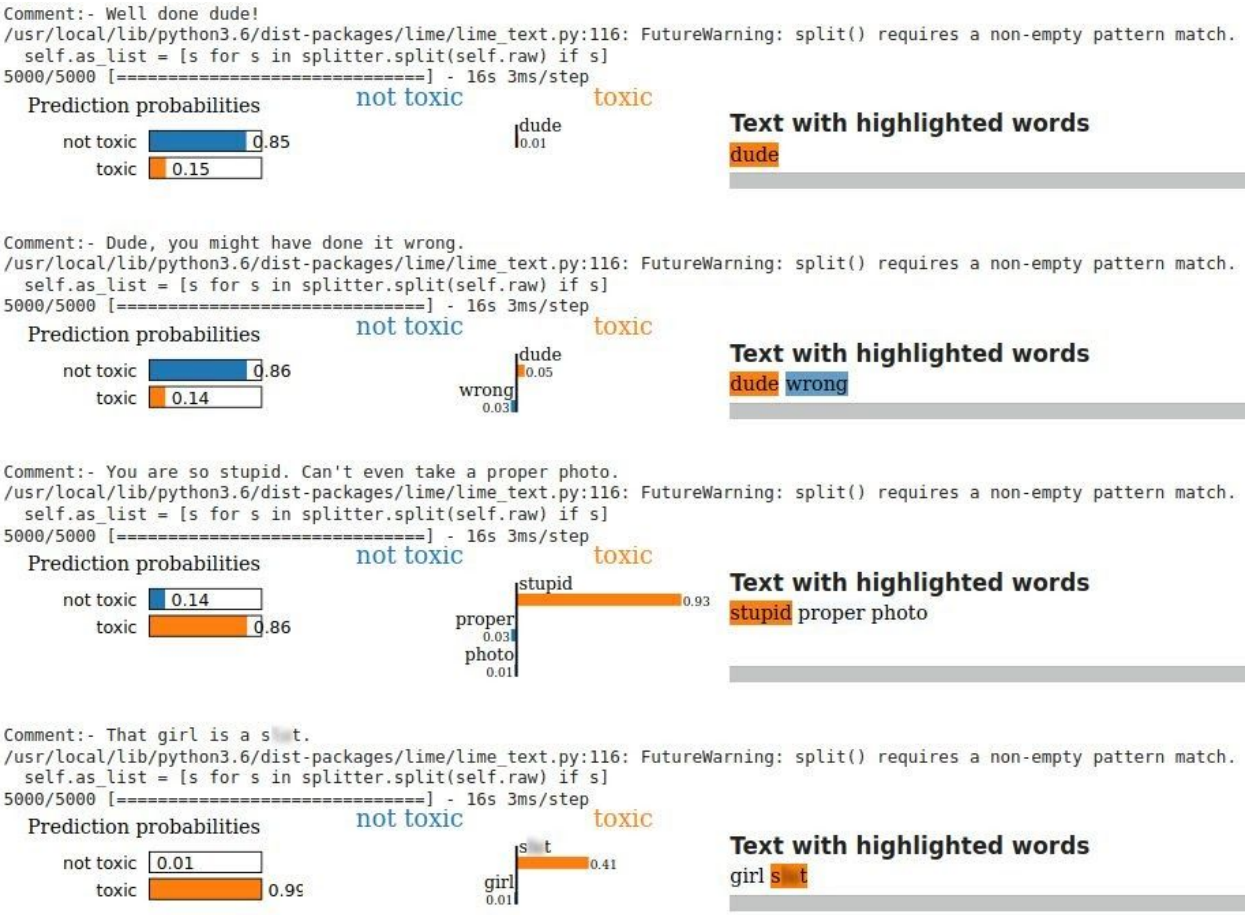


Figure 16 : Explanations of GRU + Glove

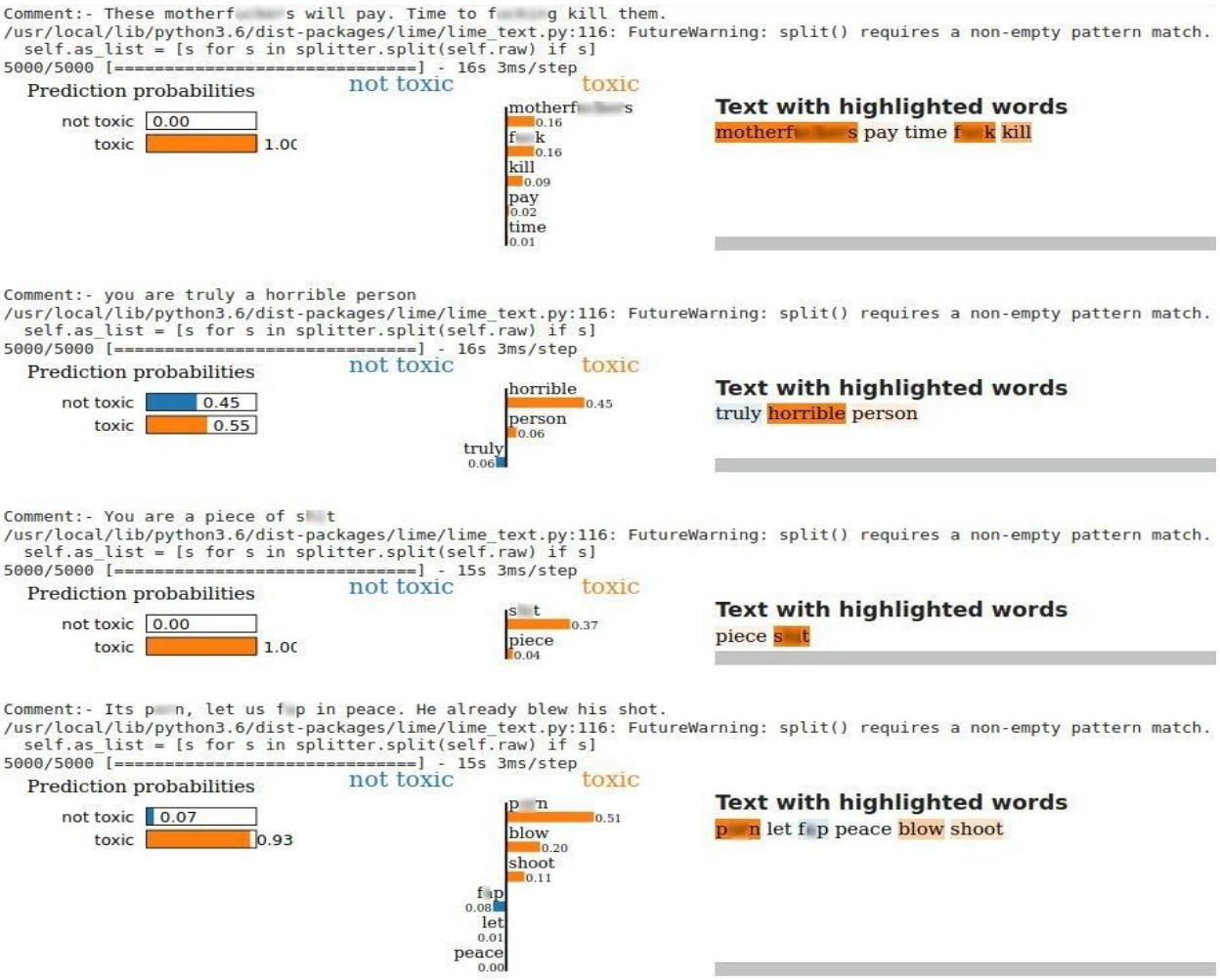


Figure 17 : Explanations of GRU + Glove

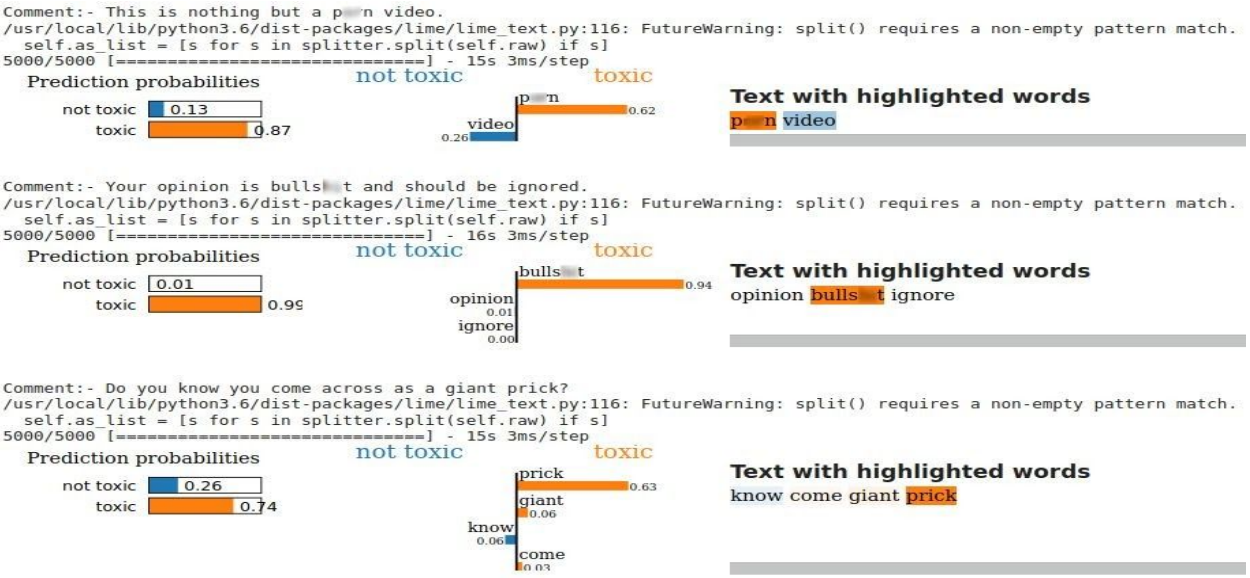


Figure 18 : Explanations of GRU + Glove

GRU (Gated Recurrent Unit) with GloVe Embeddings offers significant improvements over previous models and preprocessing techniques as shown by figures above. All sentences that we tested were correctly classified and the explanations generated were also appropriate.

Lime Explanations for GRU + Fasttext Embeddings-

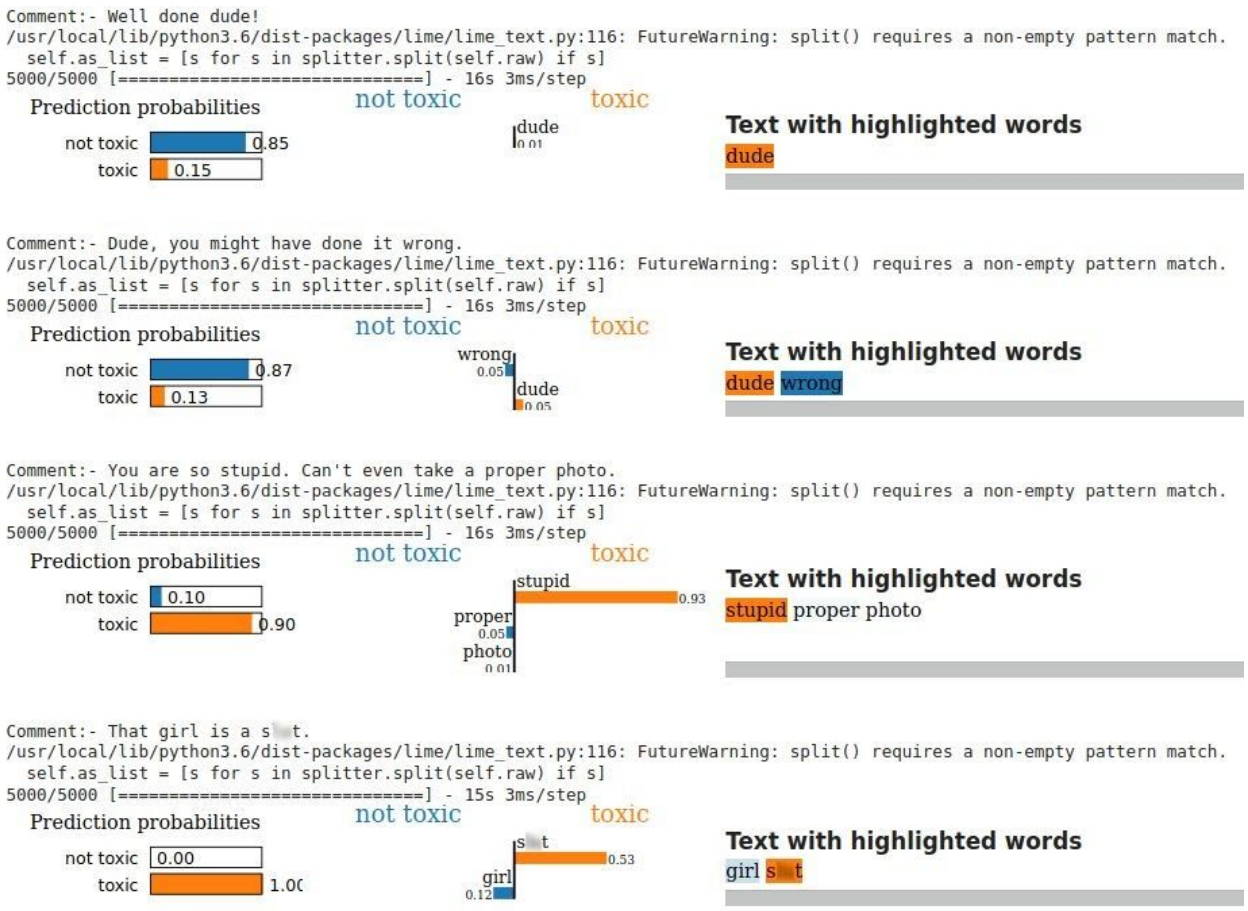


Figure 19 : Explanations of GRU + Fasttext

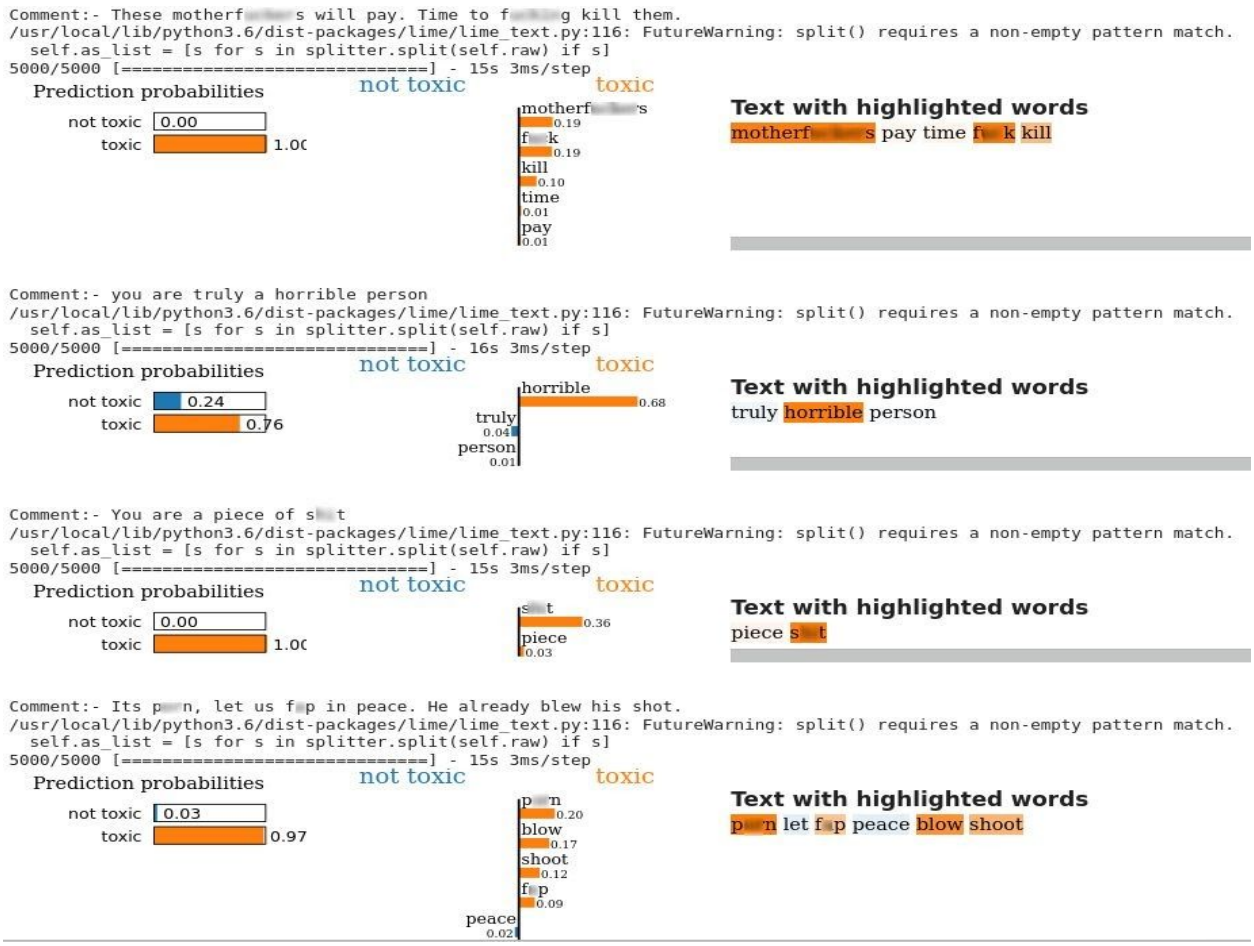


Figure 20 : Explanations of GRU + Fasttext

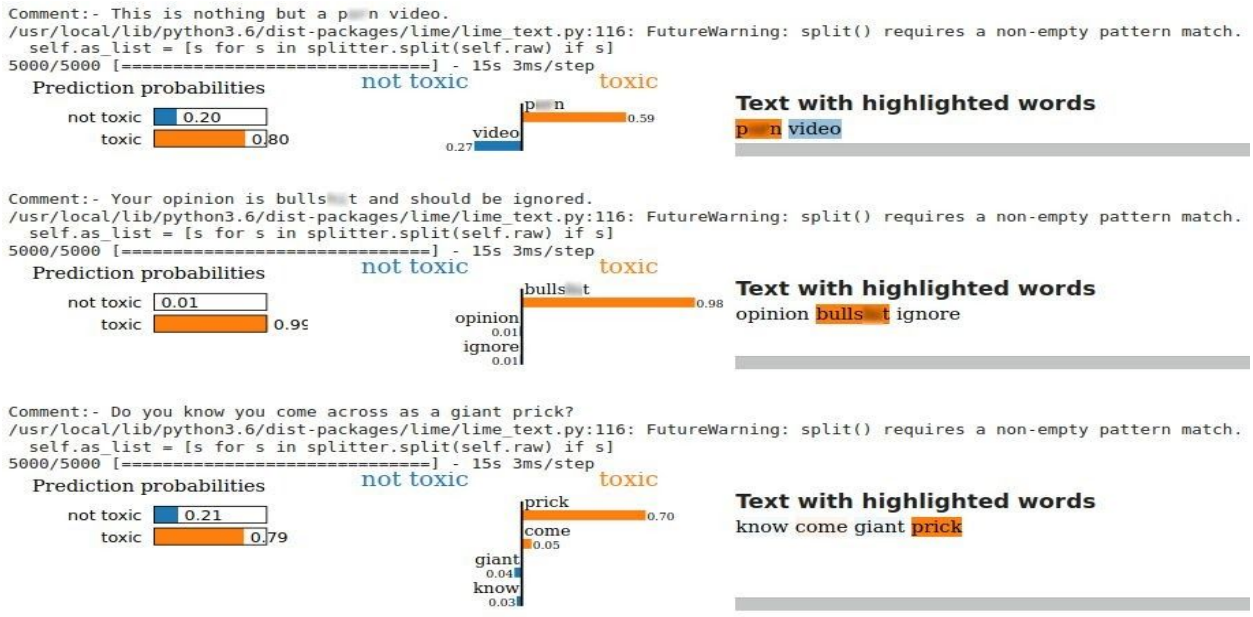


Figure 21 : Explanations of GRU + Fasttext

GRU with Fasttext embeddings (300d) also provides the best performance as shown by the scores mentioned in Table II. It also correctly classifies all the sentences that we tried . Moreover on interpreting the results with LIME , the explanations are also found to be most intuitive.

V. Issues and Limitations

From the above section, we were able to clearly conclude that GRU + pre-trained word embedding combination works the best in our current scenario with best accuracy score and intuitive LIME explanations. However, there are certain issues and limitations with regards to toxic comment classification, which we would like to highlight here in this section -

1. In certain cases, comments need to be of a substantial length (excluding stopwords) to get correctly classified. Individual words passed to the model without any context may get misclassified in a few cases.

   We observed that there are still few words like *"mouth"*, *"mother"*, *"black"* which appear frequently in toxic comments tend to shift the prediction towards the toxic side.

   But when we provide enough context (excluding stopwords), then the comments containing those words seem to get classified correctly.

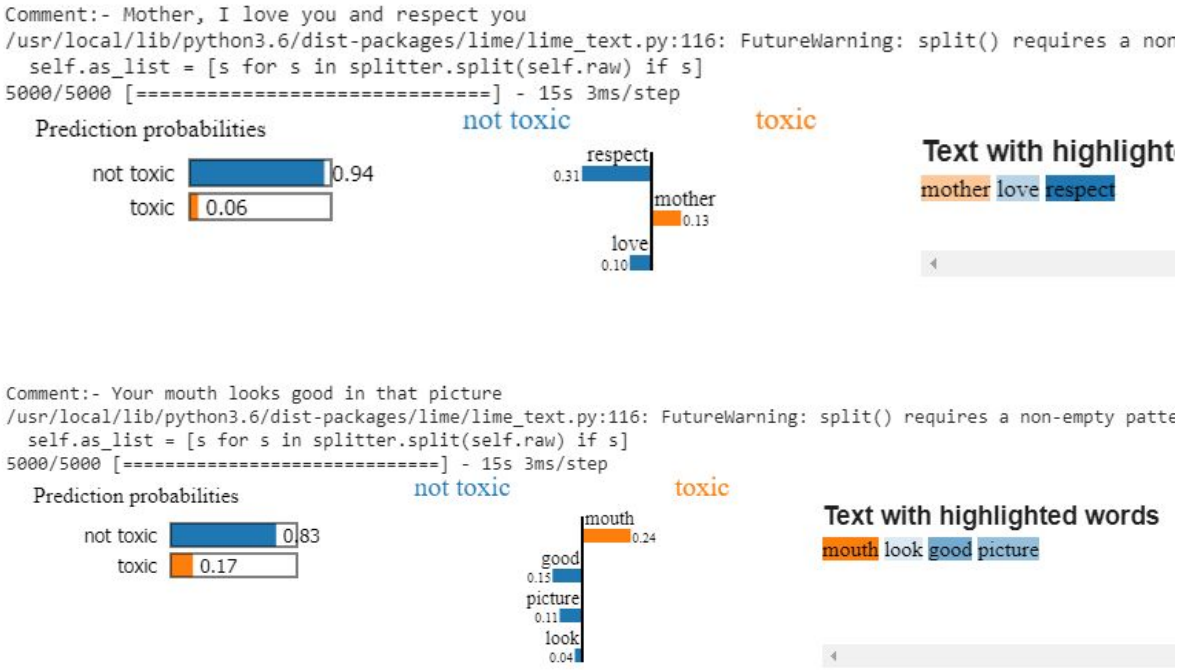


Figure 22 : Explanations of *"Mother, I love you and respect you."* and *"Your mouth looks good in the picture."* on GRU + Fasttext

2. Some threats and inappropriate comments written using "non-toxic" words and phrases may get misclassified.

Example -

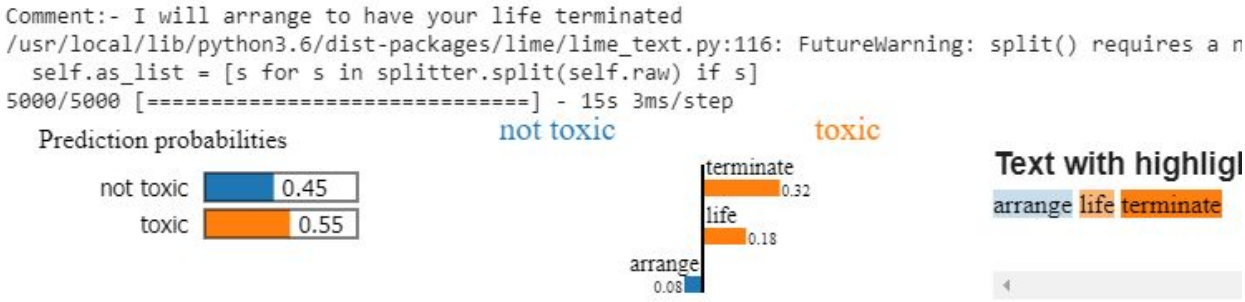


Figure 23 : *"I arrange to have your life terminated"* on GRU + Fasttext getting classified correctly

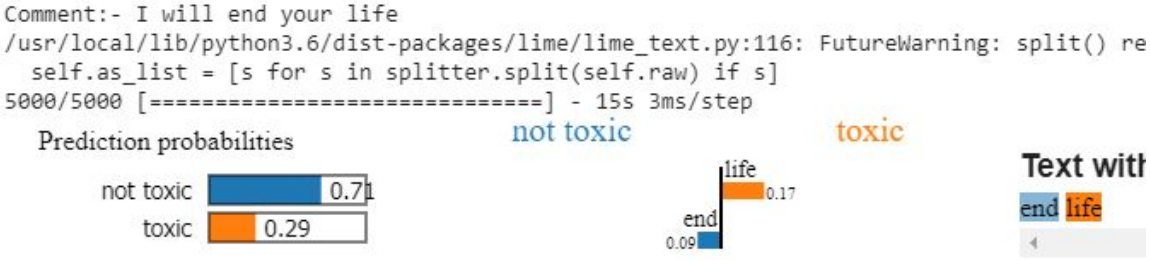


Figure 24 : *"I will end your life"* on GRU + Fasttext getting misclassified

3. The dataset and model work best for detecting and predicting toxic/abusive, inappropriate comments and hate speech found most commonly on the online comment sections but may not be able to properly deal with severe issues like caste/gender/religion/race/nationality/sexual-orientation discrimination, identity hate, personal attacks, etc even when we use our best model.

   Example -.

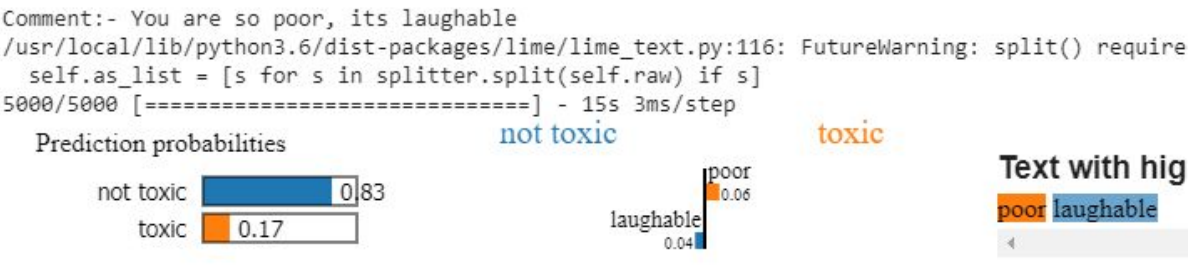


Figure 25 : Personal attacks like *"You are so poor, it's laughable"* on GRU + Fasttext.

VI. Conclusion

In this study conducted, we observed that every model we trained gave us significantly good scores on the test data. However, after looking at the predictions and explanations of those predictions we observed that GRU along with pre-trained Word Embeddings gave us most intuitive LIME explanations for the predictions being made. The model works best for dealing with commonly found toxicity and hate speech in online comment sections and threads.

So now speaking in general, an ML/DL model which gives excellent scores may lead to one believing that they have found a good enough model for carrying out a certain job/task. Model interpretability techniques like LIME can actually help in explaining why a model is making certain predictions and can help in selecting best model and preprocessing techniques. It helps in analyzing the problem, model and techniques much more in-depth. Moreover model interpretability techniques can also help in giving reasons to the end users for a certain prediction/classification made by the model, thus helping in building *trust*. This highlights the importance of model interpretability step in machine learning/deep learning projects and solutions.

VII. References

[1] Cyberbullying. (2005, August 21). Retrieved from https://en.wikipedia.org/wiki/Cyberbullying

[2] Dataset Abusive YouTube Comments. (2018, November 7).

Retrieved from https://zenodo.org/record/1479531#.XjxGNmgzbIU

[3] fastText. (n.d.). Retrieved from https://fasttext.cc/

[4] Hate-speech-and-offensive-language dataset. (n.d.).

Retrieved from https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data

[5] Internet troll. (2001, October 1). Retrieved from https://en.wikipedia.org/wiki/Internet_troll

[6] Interpretability part 3: opening the black box with LIME and SHAP. (n.d.).

Retrieved from https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html

[7] Local Interpretable Model-Agnostic Explanations (lime) — lime 0.1 documentation. (n.d.).

Retrieved from https://lime-ml.readthedocs.io/en/latest/

[8] Molnar, C. (2019). *Interpretable Machine Learning*. Morrisville, NC: Lulu.com.

[9] Pennington, J. (n.d.). GloVe: Global Vectors for Word Representation.

Retrieved from https://nlp.stanford.edu/projects/glove/

[10] Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?":

Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. doi:10.18653/v1/n16-3020

[11] Toxic Comment Classification Challenge | Kaggle. (n.d.).

Retrieved from https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[12] What Is Content Moderation and Why Companies Need It. (2019, September 3).

Retrieved from https://bridged.co/blog/what-is-content-moderation-why-companies-need-it/

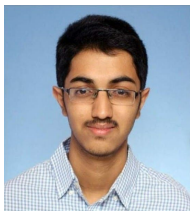
[13] What is TF-IDF? (2019, December 26). Retrieved from https://monkeylearn.com/blog/what-is-tf-idf/

[14] Word embedding. (2014, August 14). Retrieved from https://en.wikipedia.org/wiki/Word_embedding

[15] XGBoost Documentation. (n.d.). Retrieved from https://xgboost.readthedocs.io/en/latest/

VIII. Author Profiles

Aditya Mahajan, currently a BE Computer Engineering Student at Vishwakarma Institute of Information Technology (VIIT), Pune, Maharashtra, India. Very good academic track record and previously has done projects in the fields of Software Development and Web Development. Keen on exploring different domains and research areas which come under Computer Engineering.



Divyank Shah, currently a BE Computer Engineering Student at Vishwakarma Institute of Information Technology (VIIT), Pune, Maharashtra, India. Previously has worked on various Software Development and Android projects. Keen on exploring different domains and research areas related to Machine Learning and AI.

Gibraan Jafar, currently a BE Computer Engineering Student at Vishwakarma Institute of Information Technology (VIIT), Pune, Maharashtra, India. Founded CyberCell in VIIT, established ISACA, VIIT chapter. Primarily interested in Cyber security. Has worked on various Web and Android development projects. Exploring applications of ML and AI in the domain of information security.

IX. Acknowledgements

We would like to thank our mentors for their guidance and constant support throughout this study.

Bhushan Garware, Ph.D works as Data Scientist at Persistent Systems Ltd., Pune, Maharashtra, India. He holds a Bachelor Degree in Engineering with Electronics and Telecommunication specialization from University of Mumbai, Master of Technology in VLSI & Embedded systems from COEP Pune and Ph.D with Gold medal in Fingerprint Biometrics from University of Pune.

Ms. Vaishali Mishra, has received M.E. in Computer Science & Engg. From Bharati Vidyapeeth University, Pune. She is pursuing a PhD from KLEF Vijayawada in Cyber Security. She is working as Asst. Prof. at Vishwakarma Institute of Information Technology, Pune in Computer Engineering Department from 2006. She has been teaching Microprocessor, Information and Cyber Security, Computer and Cyber Forensic . She is CEH Certified by EC-Council USA.