GHOST at SemEval-2021 Task 5: Is explanation all you need?

Kamil Pluciński and **Hanna Klimczak**
Institute of Computer Science
Poznan University of Technology, 60-965 Poznań, Poland
{kamil.plucinski97,hanna.klimczak}@gmail.com

Abstract

This paper discusses different approaches to the Toxic Spans Detection task. The problem posed by the task was to determine which words contribute mostly to recognising a document as toxic. As opposed to binary classification of entire texts, word-level assessment could be of great use during comment moderation, also allowing for a more in-depth comprehension of the model's predictions. As the main goal was to ensure transparency and understanding, this paper focuses on the current state-of-the-art approaches based on the explainable AI concepts and compares them to a supervised learning solution with word-level labels. The work consists of two xAI approaches that automatically provide the explanation for models trained for binary classification of toxic documents: an LSTM model with attention as a model-specific approach and the Shapley values for interpreting BERT predictions as a model-agnostic method. The competing approach considers this problem as supervised token classification, where models like BERT and its modifications were tested. The paper aims to explore, compare and assess the quality of predictions for different methods on the task. The advantages of each approach and further research direction are also discussed.

1 Introduction

The popularity of social media platforms has been continuously increasing over time. As reported by Social (2020), 3.8 billion people have been active users of social media in January 2020. While user inter-connectivity carries a lot of positive effects, there is still a significant threat of cyberbullying and harassment caused by the illusion of anonymity online. Statistics released by Facebook (Richter, 2020) identified that in the first quarter of 2020, there were 2.3 million bullying/harassment posts and 9.6 million hate speech posts detected as a violation of Community Standards.

The importance of keeping the online community safe for users has caused many researchers to focus their work on detecting toxic contents in order to assist moderators in their difficult and mentally exhausting work. A lot of progress has been done in the task of toxic comment classification, but unfortunately, complex models still lack clear explanation and cannot gain much moderators' trust. A step towards increasing the transparency and therefore trust in automatic comment classifiers would be extending current approaches to operate on word-level rather than document-level. The ability to extract and highlight key text fragments that cause the toxic character of a comment could be of great use in post moderation.

The lack of extensive datasets for word-level toxicity detection poses an obstacle to traditional, supervised learning classification, as current state-of-the-art models are very complex and normally require large-scale data. Therefore, with the subject of this task being defined in terms of understanding and transparency for endpoint users, the authors decided to explore the explainable AI methodology. Ongoing research in xAI community examines many different approaches, with the two of them being in the focus of current work: model-specific attention-based (Mohankumar et al., 2020) and model-agnostic (Lundberg and Lee, 2017) explanations. This paper aims to compare the results obtained by these methods on Toxic Spans Detection task (Pavlopoulos et al., 2021) for the English language using supervised models (Devlin et al., 2018), acting as a baseline approach.

This paper is organised as follows. Related works and the backgrounds of discussed methods are described in Section 2. Section 3 presents three approaches to toxic spans detection, each with method-specific details in the corresponding

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 852–859
Bangkok, Thailand (online), August 5–6, 2021.

subsection. Finally, the results of the experiments are included in Section 4, with the discussion and conclusions in Section 5 and 6 respectively.

2 Related work

Pavlopoulos et al. (2017) introduced the application of a GRU-based Recurrent Neural Network for toxicity detection in documents. This approach was compared to the previous state-of-the-art for this task, presented in Wulczyn et al. (2017), which involved the use of Logistic Regression and Multilayer Perceptron and operated on n-grams. Another model compared in their work was a Convolutional Neural Network with pretrained word embeddings. RNN model outperformed other approaches, and it was further extended by the attention module, which improved the quality of classification.

Attention has been commonly used by many researchers as a medium of explanation for the predictions made by the model (Ghaeini et al., 2018). It allows to analyse what part of the input the model focuses on the most while making a prediction. Current research does not provide a unified answer to whether the attention mechanism in models is a good source of explanation. Experiments carried out in Jain and Wallace (2019) point out that attention might not be a reliable method for interpreting the model's predictions. This was due to the fact that the attention distribution did not align well with other feature-importance measures. The other argument was that providing very different attention distribution often does not impact the predictions made by the model significantly. However, Wiegreffe and Pinter (2019) challenges the aforementioned work, claiming that while the answers given by attention should not be uncontrollably trusted and the definition of explanation must be clearly stated, it could still be a useful tool for model understanding and should not be disregarded easily.

An alternative formulation of the Toxic Spans Detection is to treat it as a supervised learning task, which involves training the models to predict the toxicity of each word separately. Previously used, Recurrent Neural Networks have been replaced as the state-of-the-art approach for many tasks by transformer-based models (Vaswani et al., 2017). Transformer architecture, consisting of an encoder and decoder enriched with multi-head attention modules, turned out to outperform previous approaches on a series of NLP tasks. Bidirectional Encoder Representations from Transformers (Devlin et al., 2018) has been proposed as a powerful tool for many language-based problems. Typically, BERT is pre-trained on two unsupervised tasks when it is fed with large-scale text corpora and later adapted to a specific task during the fine-tuning stage, requiring much less data and computing time.

3 Toxic Spans Detection

This section describes the following approaches: analysing the attention of an LSTM model with orthogonalization of hidden states 3.1; using SHAP to provide explanation of BERT predictions for toxic comment classification 3.2; training BERT for classification of toxic tokens 3.3.

3.1 Orthogonal LSTM

As stated in the previous section, attention in RNN models is still a questionable medium of explanation. An interesting approach to LSTM with attention has been proposed in Mohankumar et al. (2020). The authors claim, that attention vectors in LSTM models are too similar to each other and therefore, cannot be used to explain the predictions. They propose an orthogonalization technique to increase the conicity of hidden states, enabling better interpretability. The equations of LSTM units are updated in order to orthogonalize the hidden state at a given time with respect to the previous states. The implementation used in this work has been adapted from the source code[1] provided by the authors of the original paper (Mohankumar et al., 2020).

During the preprocessing stage, the inputs with toxicity higher than 0.5 were considered as toxic examples. Due to the memory limitations of our computing architecture, the entire CivilComments (Borkan et al., 2019) dataset could not be used for training. Therefore, Random Undersampling (Cochran, 1977) of majority class was performed. The resulting set consisted of 350000 examples. The text was tokenized using spaCy tokenizer, special characters were removed as well as newline characters, multiple spaces were compressed to one space and capital letters were replaced by a lowercase equivalent. Finally, tokens were represented using FastText embeddings (Bojanowski et al., 2016) with the size of 300. The data was

[1] https://github.com/akashkm99/Interpretable-Attention at commit 2d8dd37.

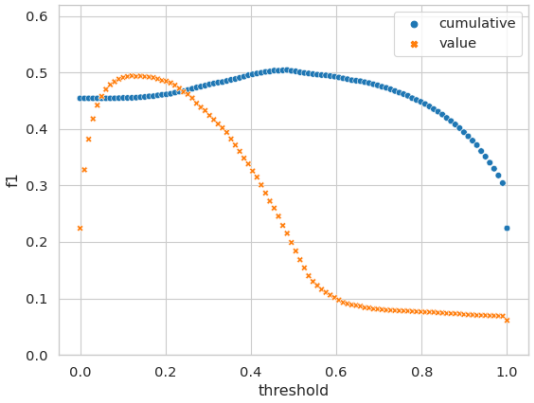

Figure 1: The chart of span-level F1 score with respect to the threshold for OrthoLSTM. Results obtained on the validation set.

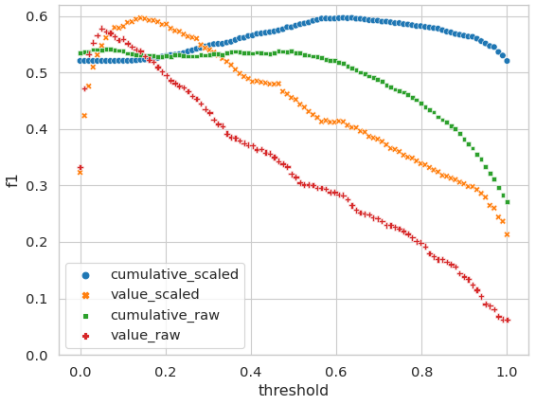

Figure 2: The chart of span-level F1 score with respect to the threshold for SHAP. Results obtained on the validation set.

split into training, validation and testing sets with the 80:10:10 ratio.

The model consisted of 1-layered orthogonal LSTM followed by the dense layer and was trained using Adam optimizer with learning rate of 1e-3, weight decay of 1e-5 and the batch size 32. This model has obtained 0.957 ROC AUC for comment-level classification. The performance of the model proved it to be sufficient for further analysis of the attention. The key parameter to obtain token-level prediction was setting a threshold for toxic token selection based on the attention value. Two approaches were investigated:

- value-based: all tokens with an attention score higher than the certain threshold were selected as toxic
- cumulative: sorting the tokens by the highest attention score, each token was added to the toxic set as long as the cumulative value of attention was not covered (e.g. 70% of the whole model's attention)

The span-level F1 score on the validation set with different threshold values for both approaches is presented in Figure 1. The best results were obtained by 0.12 threshold for the value method and 0.48 threshold for the cumulative method.

3.2 SHAP

Shapley Additive Explanations (Lundberg and Lee, 2017) method has been introduced as a model-agnostic framework that provides interpretation for model predictions. The authors of SHAP recognise that common feature importance measures rely on the same explanation model and propose a new method consisting of Shapley values approximated with kernel methods e.g. LinearLIME.

The authors decided to apply SHAP framework to explain the BERT model, trained for toxic comment classification. An implementation provided by Hanu and Unitary team (2020) in Detoxify[2] repository was used. The model is the `bert-base-uncased` [3] from `transformers` library trained on Wikipedia Comments (Wulczyn et al., 2017) using Adam optimizer with learning rate 3e-5 and weight decay 3e-6. The model obtains 0.909 ROC AUC in the binary classification task on CivilComments dataset (Borkan et al., 2019).

As in the previous method, the problem of threshold selection has been a significant bottleneck of this approach as well. The difference, as opposed to LSTM with attention, is that SHAP scores do not sum up to 1. This allowed to treat the values in two more ways: operating on unchanged values or rescaling them to sum up to 1 and performing operations in the previous fashion. The results are presented in Figure 2. Overall, the scaled approaches outperformed methods using raw SHAP scores.

3.3 Token classification

As opposed to explanation-based solutions presented in the previous section, the authors also

[2] https://github.com/unitaryai/detoxify at commit `18fd29e`.

[3] https://huggingface.co/bert-base-uncased

tested a classic, supervised learning approach to the task. This solution is based on pre-trained BERT model (Devlin et al., 2018), which is fine-tuned to predict toxic spans.

The trial part of the dataset was used for validation, where the training and testing parts were used according to their purpose. Due to the size of the training set, a model as complex as BERT is prone to overfitting. Therefore, the training time of the model consisted of 3 epochs and the learning rate started from 4.7e-5 and was divided by 10 after each epoch. The authors also proposed data augmentation to deal with this problem which was described in details in Section 3.3.3. As optimizer AdamW was used and batch size was set to 8.

3.3.1 Dealing with long inputs

The dataset for the task contained many comments with a noticeably large number of tokens. However, the BERT model is limited to 512 tokens for input data. The problem was solved by reformatting the input to fit the given size. A sample was built by adding whole sentences from the input as long as the size limit was not exceeded. A set of possible breakpoints consisted of end of sentence positions to ensure that each sentence is not split in half. Chosen breakpoints and sentence IDs were remembered in order to calculate the span-level F1 score. The length of the preprocessed samples did not have to be based only on BERT limitations and can be adjusted for better training complexity. The shorter splits were tested as a way to speed up the learning process. The results were presented in Table 1. The training time was reduced by 40%, but the decrease in quality of prediction appeared due to the much narrower context fed into the model.

		Time [m]	span-level F1
Tokens	128	43	0.660
	512	72	0.672

Table 1: The learning time in minutes and span-level F1 score on validation set according to the length of input in tokens.

3.3.2 BERT extensions comparison

Furthermore, given promising results obtained by BERT, the authors have decided to compare it to other BERT-based models. The models selected for testing were ELECTRA (Clark et al., 2020), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), SqueezeBERT (Iandola et al., 2020) and for each of them the implementation from the `transformers`[4] library was taken. The models were tested in the same manner as BERT. The hyperparameters learning rate, batch size etc. were also the same. The results are presented in Table 2. One can notice that models with a higher number of parameters tend to work better. However, learning curves analysis indicates that they are also more prone to overfitting.

Model	**span-level F1**
XLNet	0.678
RoBERTa	0.676
BERT	0.672
SqueezeBERT	0.657
ELECTRA	0.646

Table 2: The comparison of BERT and BERT-based models on the validation dataset.

3.3.3 Data augmentation

In order to deal with overfitting, the authors decided to apply a simple augmentation method. The augmentation is performed by random swaps of words appearing in text no further than 3 words from each other. The number of swaps depends on the length of the input, for each input it is calculated as follows: $alpha * number\_of\_tokens$. The results obtained while using augmentation presented in Table 3 are ambiguous and do not give a clear answer whether it helps or not. However, after a thorough analysis of train vs. loss curves, the authors noticed that the model does not overfit quite as much and the metric have smaller fluctuations. Therefore, the authors decided to use this technique during further training whenever it helps on the validation dataset.

alpha	**span-level F1**
0	0.672
0.2	0.669
0.5	0.675

Table 3: The span-level F1 score on validation set for BERT model with augmentation with a given alpha.

3.3.4 Filling empty spaces in between spans

While analysing the labels provided for the training set, the authors noticed that at times, whole text

[4] `https://huggingface.co/transformers` at version `4.0.1`

fragments were labelled as toxic, including spaces in between words. Further analysis showed that 9.54% spans in the training set are spaces. Due to the fact that tokenization resulted in omitting those spaces, an experiment had been performed in order to deal with this problem. If the output contained toxic spans separated by one or two characters, the character was also considered toxic. This technique slightly improved the performance as noted in Table 4, therefore the authors decided to include it in the final solution.

Chars	span-level F1
0	0.6720
1	0.6735
2	0.6730

Table 4: The results of marking the given number of characters in between toxic spans as toxic by BERT on validation set.

3.3.5 Ensemble

Models previously described in Section 3.3.2 were concatenated into an ensemble. The aggregation of predictions was performed with hard voting - a span was considered toxic only when 3 out of 5 models returned a toxic prediction.

4 Results

The results obtained by different methods on the test set are presented in Table 5. The highest performing model in this paper turned out to be an ensemble of different BERT-based models scoring **13th out of 91** solutions. The source code for all approaches is publicly available on GitHub[5].

[5] https://github.com/hancia/ToxicSpansDetection

Method	span-level f1
OrthoLSTM	0.4970
SHAP	**0.5987**
BERT	0.6513
XLNet	0.6624
BERT + aug 0.5 + fill 1	0.6780
Ensemble	**0.6859**

Table 5: The comparison of selected models' performance on the test dataset. Ensemble consists of all models stated in the section 3.3.2 - BERT was used with augmentation (*alpha*=0.5) and filling chars (*char*=1), other models were only used with filling chars (*char*=1).

5 Discussion

As presented in the previous section, the supervised solution to token classification outperformed both xAI approaches. The BERT model was trained specifically to solve this type of problem and therefore achieved a score slightly better than LSTM or SHAP. High labelling costs need to be taken into account when comparing those solutions, as training a robust BERT model would require much more data, that would not be necessary for xAI approaches, which are generally unsupervised.

Surprisingly, a model-agnostic approach turned out to perform much better than the attention-based solution, even though the explained models achieved very similar results on the toxic comment detection task. In Figure 3 one can see that the number of false negatives was significantly higher while relying on attention than it was while using Shapley values. This could be because the model might pay a significant portion of attention to a very toxic word, which is enough to recognise the comment as toxic. While analysing the model's focus can improve the understanding of how the model works, it might not necessarily be enough to

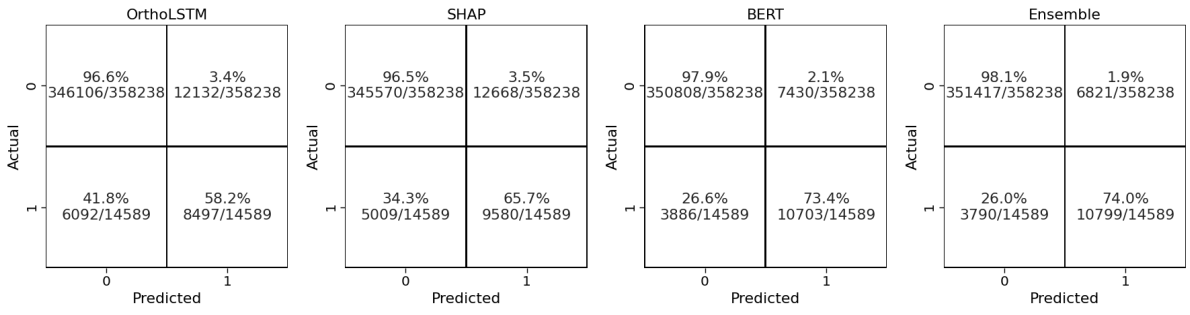

Figure 3: Confusion matrices for each method. 1 refers to toxic class, non-toxic samples are marked by 0.

translate into clear decisions for each of the input components. Furthermore, the need for threshold tuning somehow forced explainable approaches to mark a certain number of tokens as toxic, which could be reflected in a slightly higher number of false positives. Visualisations of example predictions can be seen in Table 6.

While xAI approaches might not be fitted to solve the problem of predicting the toxicity of each input word, they might still be useful for improving the transparency and understanding of predictions made by comment-level models. As recognised in error analysis, the number of false negatives was significantly higher for LSTM and SHAP. But in terms of explanation, it might be enough for the user to obtain the few most toxic words per comment, rather than marking all of them, no matter how low the toxicity score is. This would not only provide the explanation on what the model considered while making a prediction but would also be a clear and transparent answer for the user in many real-time use cases. Further examination might need to be done in order to assess the performance of those methods for a task specified in the aforementioned way.

6 Conclusions

This work discussed different approaches to the Toxic Spans Detection task. Supervised toxic token classification and xAI methods were examined to compare the results and assess whether explaining high-performing models can lead to a similar quality of prediction as models dedicated to a more detailed task. The supervised approach using the BERT model achieved the best result in this task, but the xAI methods have proven to be an interesting alternative that could reduce data preparation costs and improve transparency and understanding of the model's predictions. While not currently outperforming BERT, explainable methods can be sufficient for many tasks where binary decision models are used. A system for toxic spans detection was prepared, achieving a 0.6859 span-level F1 score and placing 13th out of 91 in the overall ranking.

Acknowledgments

The authors are grateful to Mateusz Lango for his invaluable guidance. The research by Kamil Pluciński was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

William G. Cochran. 1977. *Sampling Techniques, 3rd Edition.* John Wiley.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, Brussels, Belgium. Association for Computational Linguistics.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Forrest N. Iandola, Albert E. Shaw, Ravi Krishna, and Kurt W. Keutzer. 2020. Squeezebert: What can computer vision teach nlp about efficient neural networks? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 124–135. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *CoRR*, abs/1902.10186.

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.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4206–4216, Online. Association for Computational Linguistics.

John Pavlopoulos, Léo Laugier, Jeffrey Sorensen, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection (to appear). In *Proceedings of the 15th International Workshop on Semantic Evaluation*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

Felix Richter. 2020. Toxic content runs rampant on facebook.

We Are Social. 2020. Digital in 2020: Global digital overview.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

A Examples

OrthoLSTM	SHAP	Ensemble
That's right. They are not normal. And I am starting from the premise that they are ABNORMAL. Proceed wth the typical racist, bigot, sexist rubbish. Thanks !	That's right. They are not normal. And I am starting from the premise that they are ABNORMAL. Proceed wth the typical racist, bigot, sex ist rubbish. Thanks!	That's right. They are not normal. And I am starting from the premise that they are ABNORMAL. Proceed wth the typical racist, bigot, sexist rubbish. Thanks!
A rock has more mental horsepower than the indubitably ignorant. And more integrity as well. They're largely very dishonest people who repeat the same lies over and over. Just like their hero .	A rock has more mental horsepower than the indubitably ignorant. And more integrity as well. They're largely very dishonest people who repeat the same lies over and over. Just like their hero.	A rock has more mental horsepower than the indubitably ignorant. And more integrity as well. They're largely very dishonest people who repeat the same lies over and over. Just like their hero.
Ha ha, HILLARY LOST. All you hillary fools need to be rounded up and put in prison along with your leader. You are all anti–American scum. Perhaps we should dump you in Mexico along with all the illegal trash you let in....if you love 'em so much, go live with 'em. But you are not welcome here, you hate filled Marxist morons.	Ha ha, HILLARY LOST. All you hillary fools need to be rounded up and put in prison along with your leader. You are all anti–American scum. Perhaps we should dump you in Mexico along with all the illegal trash you let in....if you love 'em so much, go live with 'em. But you are not welcome here, you hate filled Marxist morons .	Ha ha, HILLARY LOST. All you hillary fools need to be rounded up and put in prison along with your leader. You are all anti–American sc um. Perhaps we should dump you in Mexico along with all the illegal trash you let in....if you love 'em so much, go live with 'em. But you are not welcome here, you hate filled Marxist morons .
Rabidly anti-Canadian troll .	Rabidly anti -Canadian troll .	Rabid ly anti-Canadian troll .
No, my poor benighted correspondent, your reductionism reveals an archaic mindset. Intersex, questioning, nongendered, and a vast variety of others have rights too!	No, my poor benighted correspondent, your reductionism reveals an archaic mindset. Intersex, questioning, nongendered, and a vast variety of others have rights too!	No, my poor benighted correspondent, your reductionism reveals an archaic mindset. Intersex, questioning, nongendered, and a vast variety of others have rights too!
justin is an embarrassment to Canada. he needs a muzzle. he needs a brain .	justin is an embarrassment to Canada. he needs a muzzle. he needs a brain .	justin is an embarrassment to Canada. he needs a muzzle. he needs a brain.
You are a simpleton, who is full of racial hatred. That is too bad. You make America a worse place.	You are a simpleton, who is full of racial hatred . That is too bad. You make America a worse place.	You are a simple ton, who is full of racial hatred. That is too bad. You make America a worse place.
When you consider the source - he writes like the Trump we've all come to know - "I could stand in the middle of 5th Avenue and shoot somebody and I wouldn't lose voters", a racist, misgynistic, liar who only brings hate to the table .	When you consider the source - he writes like the Trump we've all come to know - "I could stand in the middle of 5th Avenue and shoot somebody and I wouldn't lose voters", a racist, misgynistic, liar who only brings hate to the table.	When you consider the source - he writes like the Trump we've all come to know - "I could stand in the middle of 5th Avenue and shoot somebody and I wouldn't lose voters", a racist, mis gy nistic, liar who only brings hate to the table.
Total rubbish ! The birther bit was started by Crooked Hillary and perpetuated by Zero his own damn self.	Total rubbish ! The birther bit was started by Crooked Hillary and perpetuated by Zero his own damn self.	Total rubbish ! The birther bit was started by Crooked Hillary and perpetuated by Zero his own damn self.
Damn, you beat me to it	Damn, you beat me to it	Damn, you beat me to it
I don't think they eat them, just kill them, chop them up and sell off the parts .	I don't think they eat them, just kill them, chop them up and sell off the parts.	I don't think they eat them, just kill them, chop them up and sell off the parts.
F*cking nasty...	F* ck ing nasty...	F*cking nasty...

Table 6: Example predictions done by models. No background colour refers to true negative, green - true positive, red - false negative and blue - false positive. Examples were selected where the predictions between models were different.