Oscar the GROUCH: Graphical Representations and Observations for Understanding Classification of Hate

Benny Longwill, Courtney Mansfield, Preeti Mohan, Simola Nayak, David Nielsen, and Bill Presant
University of Washington
{longwill, coman8, preetmhn, simnayak, ndavidl, wpresant}@uw.edu

Abstract

We perform an analysis of pre-trained BERT language models to identify what type of syntactic and semantic information is attended to by BERT attention heads. We utilize BERT for transfer learning in a multi-label hate speech classification task, using dataset modification and augmentation for contrastive analysis. Multiple models are trained for the task, each using a different dataset: an unmodified monolingual dataset, a modified monolingual dataset, and a multilingual dataset. We feed inputs to these classification models and use visualization techniques to contrast BERT attention head behavior across the different models. We find that punctuation and stopword removal causes attention heads to focus more intensely on relevant semantic hate-speech tokens. We also identify particular multilingual BERT attention heads which attend to relevant semantic hate-speech tokens regardless of input language, suggesting that BERT might encode linguistic semantic properties common across all languages.

1 Introduction

In the present day, moderation on most major social media platforms consists of a mixture of automated and human efforts, typically employing neural models for the first stage of detecting hateful and abusive content. It is crucial that such models exhibit a nuanced notion of what constitutes hate and abuse on platforms, and understanding how such mechanisms work is crucial to maintaining user trust and ensuring healthy discourse and safe online communities. Usually, when models fail to detect such content or send out false alarms, the responsibility shifts to human moderators, who end up seeing the hateful content themselves.

Our paper reports the results of dataset ablation methods in order to investigate the activation of attention heads and examines the role of punctuation and stopwords in BERT's determination of hatefulness or abuse. We investigate its effects on precision, accuracy, and attention, and posit if BERT encodes or attends to semantic information related to pejoratives.

2 Related Work

2.1 BERT and Transfer Learning

Transfer learning uses general representations from a pre-trained model to train another task-specific model. BERT language models have been particularly successful, utilizing a pre-training objective of missing-word prediction, such that each token attends to all other tokens in the sequence (Devlin et al., 2018). If a task-specific model performs well, then BERT's representations presumably encode information relevant to that task.

BERT's Transformer architecture consists of multiple layers of attention heads. In multilayer models, lower layers encode "local" features, while higher layers encode "global" features (Belinkov and Glass, 2019). Particular attention heads attend to specific linguistic concepts; for example, a few attention heads encode "dobj" syntactic dependency relations, although no single head individually encodes *all* syntax (Clark et al., 2019).

Sequence alignments can be used to visualize attention weights between input and output tokens at each attention head at each layer (Ruder et al., 2019). Open source software tools such as BertViz exist for this type of Transformer attention visualization (Vig, 2019).

2.2 Hate Speech Classification

Hate speech classification has been approached as a transfer learning problem, with BERT as the transfer model, yielding good performance on annotated Twitter datasets (Mozafari et al., 2019). Recent

work has attempted hate speech classification using multilingual datasets and models (Ousidhoum et al., 2019). Previous authors using multilingual BERT models have demonstrated that supplementing low-resource language datasets with machine-translated data can yield state-of-the-art results (Sohn and Lee, 2019). Perfectly annotating a hate speech dataset, however, is difficult: annotators will disagree on what constitutes hate speech based on their inherent racial biases (Sap et al., 2019), and non-expert annotators are likely to mark non-hate speech as hate speech. (Waseem, 2016).

2.3 Linguistic features

Hate speech classification models can be enhanced by utilizing linguistic features beyond simple word-grams, such as punctuation and capitalization.

Nobata et al. (2016) use number of punctuation, repeated punctuation, number of tokens with non-alpha characters, and similar properties to enhance their classification system. Punctuation may be a particularly useful feature for hate speech, as it has been referred to as the prosody of written language (Chafe, 1988), where prosody is an important means for a speaker to convey emotion. Furthermore, punctuation has been directly associated with conveying emotion in online communication (Hancock et al., 2007). While classifiers take punctuation into account, the degree of its usefulness is not clear. The above studies also focus on English; the effectiveness of punctuation in the classification of other languages is not well studied.

Chan and Fyshe (2018) find that capitalization on social media is a marker of sentiment, and is often associated with negative sentiment. Unsurprisingly, capitalization has also proven to be useful in hate speech detection (Nobata et al., 2016).

3 Methods

3.1 Datasets

We used three Twitter-derived hate speech datasets for our analysis, each from a different language. The datasets contain tweet text accompanied by a class label of "Hate", "Abuse", or "None"[1].

The English dataset from Founta et al. (2018) contains 100,000 tweet IDs. Tweets were filtered if they were marked as unoriginal (e.g. retweets), non-text, or non-English. The Arabic dataset ("L-HSAB") from Mulki et al. (2019) contains 5,846 tweets. Tweets were collected using queries usually targeted for hate speech (e.g., "refugees") and from popular users known to frequently produce hate speech. Tweets were normalized by removing Twitter-inherited symbols and non-Arabic characters. The Indonesian dataset from Ibrohim and Budi (2019) contains 13,169 tweets. Annotators came from various demographic, political, spiritual, educational, and social backgrounds.

We preprocessed each dataset by removing usernames and URLs. Retweets and hashtags were left intact. We randomly split training and test data 90-10. The number of training and test samples can be found in Table 1. The distribution of the labels reflects their distribution in the original dataset. Label counts for test data are shown in Table 2.

Table 1: Number of training and test tweets for each language. The Multilingual dataset is a concatenation of all languages.

Dataset	Train	Test
Arabic	5262	585
English	54947	6106
Indonesian	8860	985
Multilingual	69070	7675

Table 2: Distribution of labels in the test set for each language. The Multilingual dataset is a concatentation of all languages.

Dataset	Normal	Abusive	Hate	Spam
Arabic	334	206	45	0
English	1676	3305	779	346
Indonesian	576	182	227	0
Multilingual	2918	3374	815	568

3.2 Experiment Setup

We constructed a set of 16 experiment configurations. Given the English dataset, we would train 4 models on: (1) the unmodified dataset, (2) the dataset with punctuation removed, (3) the datset with stopwords removed, and (4) the datset with both punctuation and stopwords removed. We repeated this for three other languages: Arabic, Indonesian, and "multilingual" (a concatenation of the Arabic, English, and Indonesian datasets) to train 16 hate speech classification models in total.

As a starting point, we adapated McCormick and Ryan's (2019) tutorial code for training and fine-

[1]The English dataset also included labels for "Spam". The Indonesian dataset contained subtags for finer distinctions but these were not utilized.

tuning a probing task, using a BERT transfer model and the Hugging Face transformers library. We introduced additional dependencies on AllenNLP libraries.

We loaded pretrained BERT models: English for English dataset, Multilingual for all others. Text inputs were tokenized and padded to ensure that all examples have the same number of features. Padding was determined by looking at the length of each input in each dataset and taking the 95th percentile among them. For each experiment, we trained a BertForSequenceClassification multi-class hate speech classifier. We used BertViz to produce visualizations of each model's attention heads, given sample hate and non-hate text inputs.

3.3 Hypotheses

Linguistic features such as casing and punctuation have been shown to improve model performance in previous work. Thus, we hypothesize that a model trained with all linguistic features will achieve higher performance and that its attention heads will attend more to the relevant linguistic features.

If multilingual BERT models have a universal notion of "hatefulness", we hypothesize that the same attention heads attend to both Arabic hate speech and Indonesian hate speech. If multilingual BERT models do not encode universal semantic concepts, then we would expect to see one unique node attending to "Arabic hatefulness", and another attending to "Indonesian hatefulness".

4 Results

4.1 Model Performance

We report the results of our BERT-based models using accuracy and F1 score. Monolingual English trained models used BERT-base uncased; all other models used BERT-base multilingual uncased. Model results are indicated in Table 4.

Model performance against the baseline is mixed. Although the English BERT model is considerably improved over a naive bayes baseline, an RNN baseline with GLoVE embeddings performs best. The Arabic model shows a higher recall compared to a naive bayes baseline, but lower precision. We expect additional hyperparameter tuning would result in performance gains for our BERT models.

Across all BERT models, we see that performance on normal and abusive categories is higher than that of hate speech. This is reflected in the multilingual model, where the F1 score for hate speech is 64% compared to 88% on normal speech and 86% on abusive speech.

4.2 Dataset Modification

The results of the ablation of casing, punctuation, and stop words in each language are presented in Table 3. The English model, which relied on the monolingual BERT model, did not degrade when casing, punctuation, or stop words were removed.

The Arabic and Indonesian model performed best when all features were taken into account [2]. Removing stop words from these models resulted in a 3% drop in F1 for both languages. Punctuation had a smaller effect, resulting in a 2% drop in accuracy for Indonesian.

Attention visualizations revealed a difference in BERT attention head weights for transfer models trained on modified data versus transfer models trained on unmodified data. Going forward, these will be referred to as `english_1` (unmodified data), `english_2` (no stopwords), `english_3` (no punctuation), and `english_5` (no stopwords or punctuation).

4.2.1 General Model Differences

Figure 1 presents the difference between the BERTViz visualization graphics of two entire models' attention on example abusive input. The differences are more pronounced in downstream layers (between 6-11). Congruent with prior work (Belinkov and Glass, 2019), our layers 1-5 focus more on the local features (hence why there are very few differences, given that the input sentence is the same).

We see relatively strong differences between the heads in layers 10 and 11. It can be inferred that removing punctuation has a higher impact on these layers than removing stop words (because there are stronger differences visually present), although removing either shows significant differences.

4.2.2 Layers 10 and 11

In layer 10 (Figure 2), `english_3` pays comparatively less attention to punctuation but more to "picture", especially in the third head (L10H3). `english_2`, `english_3`, and `english_5` focus on profanity more than `english_1`. `english_5` attends less to "filthy" and "picture",

[2] Arabic does not make a distinction between upper and lower case characters.

Table 3: Results for ablation of linguistic features.

	Arabic			English			Indonesian		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
all features	**0.80**	**0.81**	**0.79**	0.81	0.82	0.82	**0.89**	**0.89**	**0.89**
w/o casing	-	-	-	0.81	0.82	0.82	0.88	0.88	0.88
w/o punctuation	0.79	0.80	**0.79**	0.81	0.82	0.82	0.87	0.87	0.87
w/o stop words	0.75	0.78	0.76	0.81	0.82	0.82	0.86	0.86	0.86

and very little towards [CLS]. `english_5` attends strongly to both "bitch" and [SEP].

In Layer 11 (Figure 3), the processed datasets attend less to [SEP]. Compared to `english_1`, `english_2` pays less attention to "the". `english_3` pays even less attention to the apostrophe, and its visualization seems more centered around the profanity "bitch". `english_5` pays the most attention to "bitch", and the least attention to the apostrophe.

Layer 10 focuses on the words "filthy" and "picture" more than layer 11, though not for `english_5`, which has a hyper-fixation on the word "bitch".

We note distinct changes in L11H0 related to punctuation-removal (Figure 5). `english_1` and `english_2` (trained on datasets with punctuation) attend to the apostrophe. `english_3` and `english_5` (trained on datasets without punctuation) attend more to "bitch". `english_3` (trained on a dataset without stopwords), attends slightly to "filthy", while the other models do not.

4.3 Multi vs Monolingual Trained Models

Figure 4 presents several diff visualizations of the model views of `all_1` and `indonesian_1` across three different labels (i.e., hate, abusive, normal). Looking at the pattern of results displayed in Figure 4, it appears that, across all labels, there are few differences in layers 0 through 5 with significantly more differences demonstrated in layer 6. These results are congruent with our hypothesis by demonstrating that the initial layers are more generalized across different models; whereas, later layers specialize more deeply on the learned dataset. That is to say that the lack of differences in layers 0 through 5 suggest that `all_1`

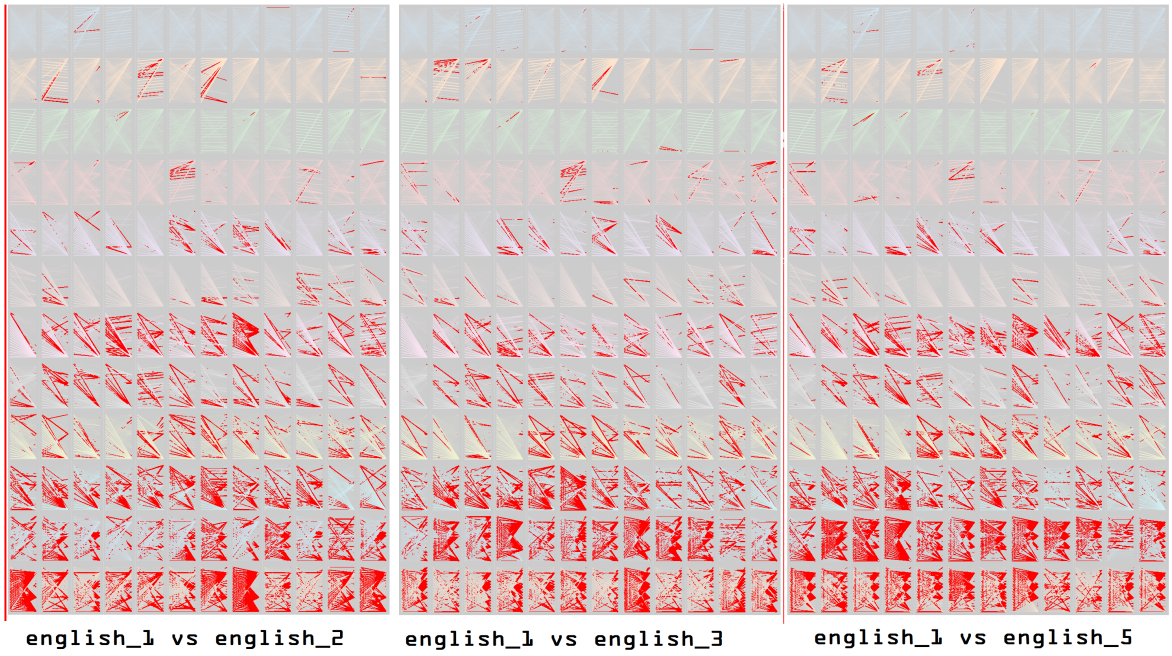

Figure 1: The differences between BERT trained on an unmodified dataset versus, from left to right, a dataset with no stopwords, a dataset with no punctuation, and a dataset with neither stopwords or punctuation. The red lines on the diagram highlight the differences in attention allocation between the compared models.

and `indonesian_1` likely have a shared focus on general sentence representation; whereas, the multitude of differences denoted in red for layer 6, suggest that this layer is particularly specialized to encoding unique information from each of the languages from which these models were trained.

4.4 Profanity Analysis

Per the analysis above, we see that there are specific nodes, especially in layers 10 and 11, that attend to profanity (e.g., "bitch"). The following analyses are done on `english_5`, the model trained on a dataset without punctuation or stop words.

4.4.1 Varying Levels of Profanity

There are varying "levels" of profanity in tweets. Below are three examples, ordered from least to most offensive:

> USER USER your an idiot TROLL and the furthest thing from what you think you are 007 is your frigging IQ
>
> RT USER Horse face hoe stop playing before I show the world yo lil ugly ass
>
> It's always the filthy bitch that comes in the picture..

We saw that there were heads that strongly focus on profanity. Other heads we examined, for example L11H0 above, that focused intensely on

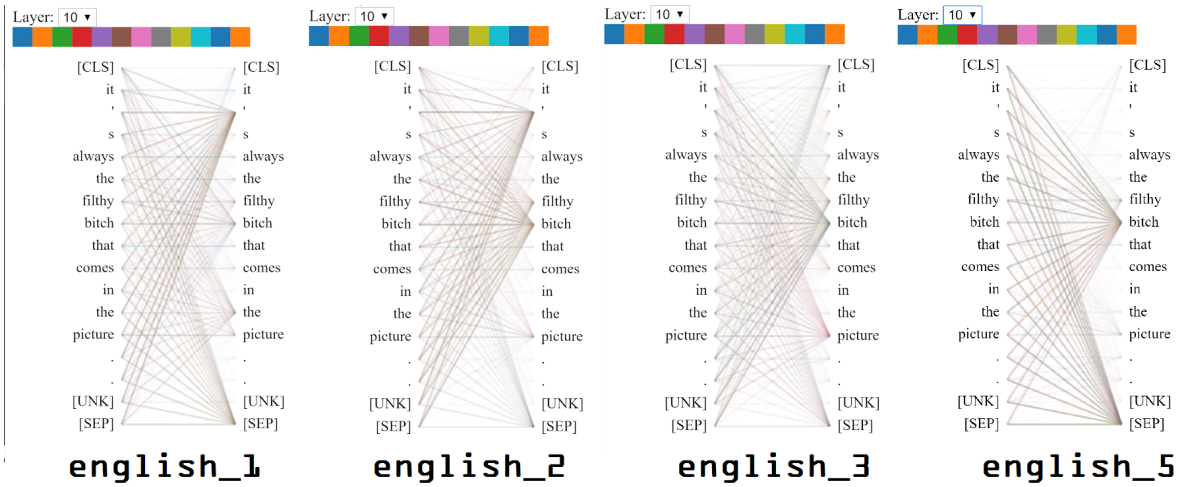

Figure 2: The differences in L10 attention allocation for the abusive input sentence "It's always the filthy bitch that comes in the picture.."

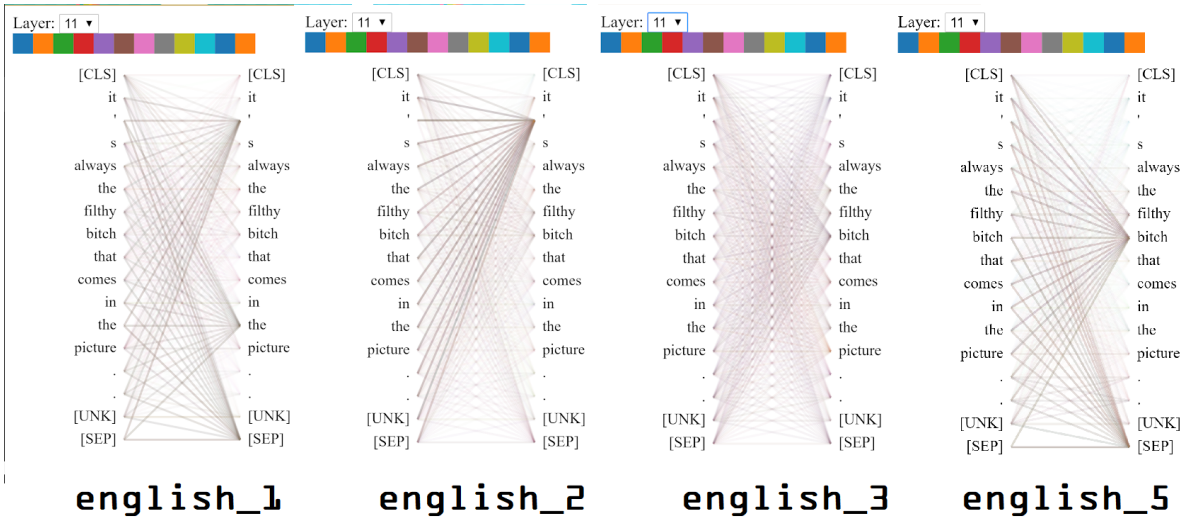

Figure 3: The differences in L11 attention allocation for the abusive input sentence "It's always the filthy bitch that comes in the picture.."

Table 4: Results for baselines and BERT-based models.

	Acc.	Prec.	Rec.	F1
Arabic				
Baseline NB (Mulki et al., 2019)	0.88	0.86	0.71	0.74
BERT-ml	0.81	0.80	0.81	0.79
English				
Baseline NB	0.39	0.66	0.39	0.31
Baseline RNN (Founta et al., 2019)	0.84	0.85	0.85	0.85
BERT-English	0.82	0.81	0.82	0.82
Indonesian				
Baseline RF (Ibrohim and Budi, 2019)	0.76	-	-	-
BERT-ml	0.88	0.88	0.88	0.88
Multilingual				
BERT-ml	0.83	0.82	0.83	0.82

the word "bitch" and didn't focus as clearly on less offensive profanity found in other input sentences (Figure 6).

A potential explanation is that this head focuses on "severe" profanity and does not allocate attention to less-profane words. We then test this same head on input data that contains equally profane, but different, words.

4.4.2 Similar Levels of Profanity

Input tweets with "equally" (subjectively chosen by us) profane texts were compared to test if L11H0 focuses on a specific severity of profanity. (Figure 7).

The word "bitch" is still focused on more than the other profane words selected ("fucking" and "shit") - which seem to be attended to as much as "ass" and "idiot" from the previous examples. We then test if this head specifically focuses on the word "bitch".

See Figure 8. Though in the first example the model attends to the apostrophe in "don't", attention is still moderately allocated to "bitch" to a moderate degree, and is strongly allocated strongly to "bitch" in the second two examples. We believe that this head may specifically focus on the word "bitch", and use the presence of the word as an indicator for abusive text.

5 Discussion

5.1 Stopwords and Punctuation

While removing punctuation, stopwords, and casing features in multi-lingual models can be detrimental to model performance, it has no impact on the English model. The larger, mono-lingual English BERT model may render these details re-

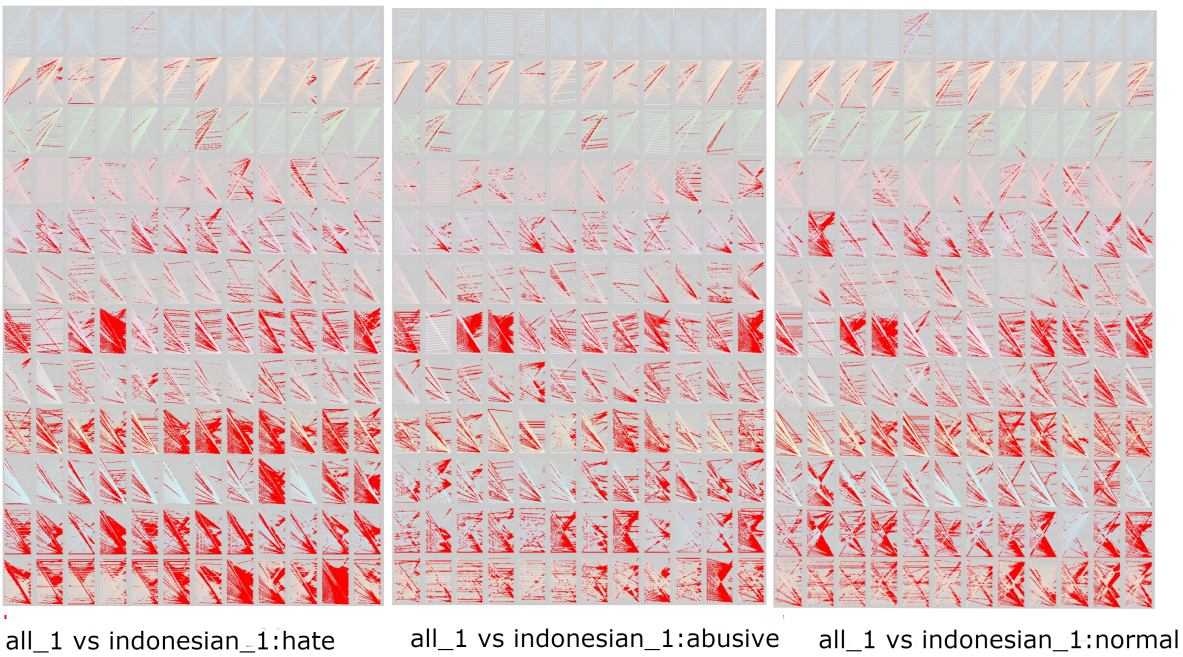

Figure 4: Classifier based on Multilingual BERT, trained on a concatenated multilingual dataset versus an Indonesian-only dataset. The red lines highlight the differences in attention allocation, given the same Indonesian input text.

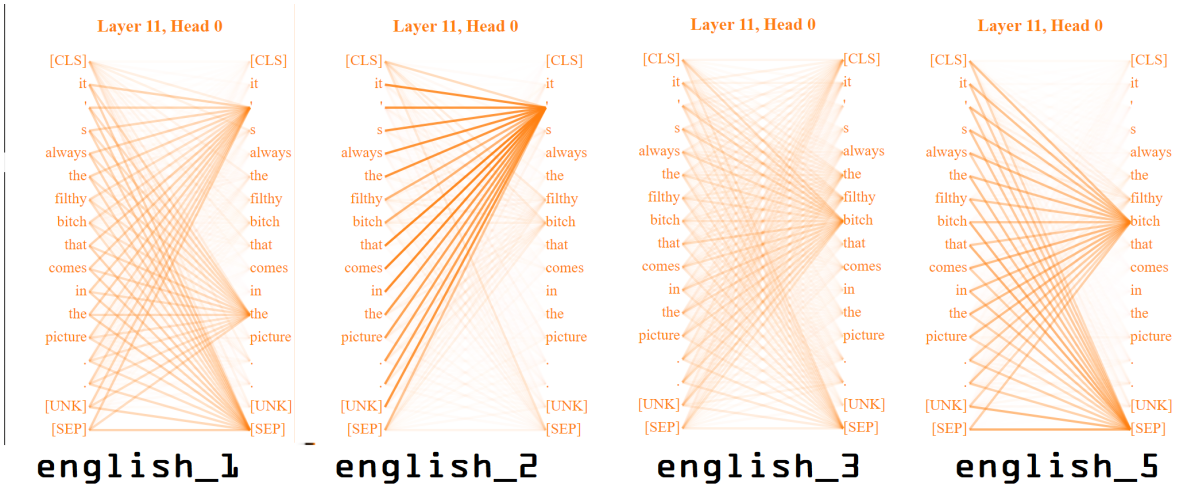

Figure 5: The differences in L11H0 attention allocation for the abusive input sentence "It's always the filthy bitch that comes in the picture.."

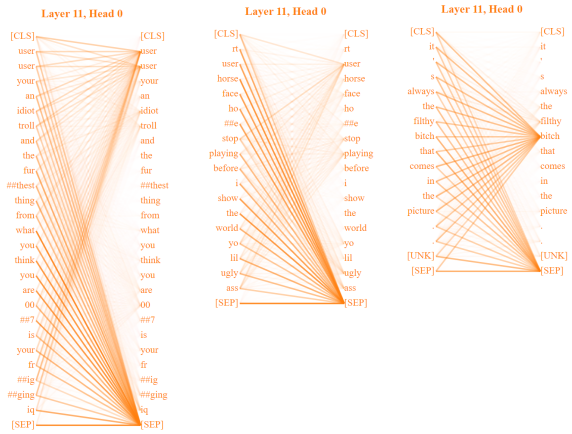

Figure 6: The differences in L11H0 attention allocation for the abusive input sentences that vary in profanity.

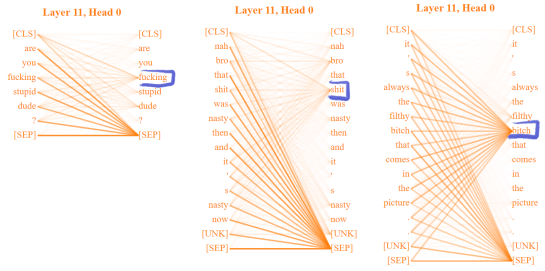

Figure 7: The differences in L11H0 attention allocation for the abusive input sentences that all have more significant profanity.

dundant for the hate speech classification task.

For English data, we see that BERT does attend to punctuation and stopwords when they are present in training, but attends less when they are not. BERT's attention heads in higher layers, which are theorized to be more responsible for encoding semantic information, may not be attending strongly enough to these tokens for the token-removal to have a meaningful impact on the overall model output. This may be explained by Rogers et al. (2020), who mention that [CLS], [SEP], and punctuation tokens are highly attended to by merit of their high frequency in the training data, but that this attention could be a "no-op" rather than attention to a sentence-level representation. They note that in higher layers, [SEP] tokens are more attended to but also less important for predictions (Rogers et al., 2020). The same may be true of punctuation tokens in the higher semantic-focused layers of BERT.

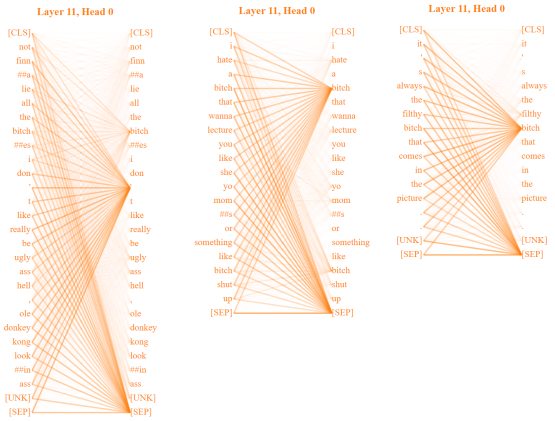

Figure 8: The differences in L11H0 attention allocation for the abusive input sentences that all have the same profane word.

5.2 Semantic Hate Speech Attention Heads

We find that specific BERT attention heads attend to tokens indicative of hate speech.

First, we examine the model trained on multi-

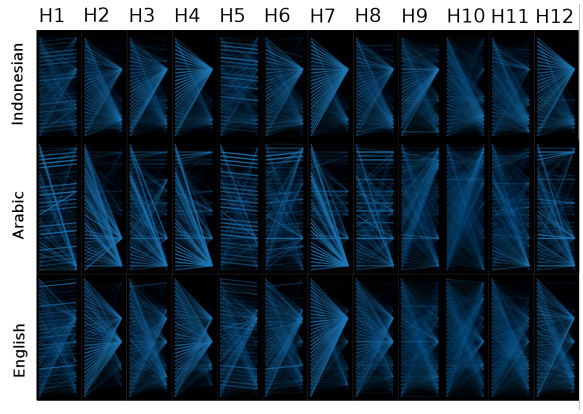

Figure 9: BERT layer 10 attention heads hate-speech sentences in Indonesian, Arabic, and English

lingual data. In Figure 9, we see that Layer 10 attends to semantically relevant tokens, regardless of input language. As hypothesized, this indicates that particular attention heads (especially L10H2) attend to semantic information *common across multiple languages*. The attention head captures some generalized notion of "cross-lingual semantics". If, instead, the model had encoded two separate concepts ("English semantics" and "Arabic semantics"), then we would expect to see one attention head attend to semantically-relevant English input, and another attend to semantically-relevant Arabic input.

Second, we find that while some BERT attention heads attend to semantically relevant hate speech tokens in general, other BERT attention heads are trained to attend to a specific subset of profane tokens, as detailed in Section 4.4 These attention heads are essentially keyword detectors: they attend to specific words or phrases. Thus, while BERT might be sophisticated enough to detect hate speech subtleties, it might also be fooled into predicting false positives by the presence of triggering keywords (e.g. "bitch"). This is problematic for hate speech classification, since an ideal hate speech classifier could detect when a speaker is *discussing* hateful keywords like "bitch", rather than *employing* hateful keywords for hate or abuse.

6 Conclusion

We find that the attentional patterns exhibited by BERT on mono- and multilingual tweets indicate a cross-linguistic encoding of hateful or abusive words, with multiple heads around layer 10. These results are unsurprising, given that such activations require input at the level beyond individual words.

6.1 Limitations

Model performance could be improved by tuning hyperparameters, e.g. training over more than 2 epochs. Visualizations could be improved by visualizing individual neurons, rather than only attention heads. We could test model robustness by feeding it inputs designed to elicit false positives. Likewise, we might improve the model by including more negative samples (i.e. label = "None") that contain hateful keywords.

6.2 Future Work

Employing the lottery ticket hypothesis or other dropout experiments could aid in generating a pruned network with less activations but similar performance. Additionally, we could consider *larger* models to investigate if `bert-large` can encode higher-order features (e.g. pragmatics) or classify different types of pejoratives.

We could improve the transferability of our classifier: we could train on general sentiment labels, rather than hate speech labels, and see if the system achieves similar results. Or, we could improve the transferability of our pre-trained model: if we pretrained on tweets as opposed to Wikipedia (as BERT does), the model may work better with our datasets (Raffel et al., 2019).

Our experiments were performed on tweets in isolation, but providing additional context could better inform the model to the presence of abusive or hateful speech. Users also commonly obfuscate hateful language by character omission or substitution with non-alphabetical symbols, often to elude machine detection. Accounting for this would make BERT more robust to misspellings and syntactic errors. Additionally, given that the nature of internet language is dynamic, fine-tuning BERT on swiftly evolving social media data could potentially improve performance. It is worth investigating whether BERT can learn pragmatic information to detect phenomena such as slur reclamation.

References

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Wallace Chafe. 1988. Punctuation and the prosody of written language. *Written communication*, 5(4):395–426.

Sophia Chan and Alona Fyshe. 2018. Social and emotional correlates of capitalization on twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 10–15.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

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*.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 105–114.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Jeffrey T Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.

Chris McCormick and Nick Ryan. 2019. Bert fine-tuning tutorial with pytorch. https://mccormickml.com/2019/07/22/BERT-fine-tuning/.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

Sebastian Ruder, Matthew Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. https://docs.google.com/presentation/d/1fIhGikFPnb7G5kr58OvYC3GN4io7MznnM0aAgadvJfc/edit#slide=id.g569f436ced_0_26.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.