"To Target or Not to Target": Identification and Analysis of Abusive Text Using Ensemble of Classifiers

Gaurav Verma, Niyati Chhaya, Vishwa Vinay
{gaverma, nchhaya, vinay}@adobe.com
Adobe Research, India

Abstract

With rising concern around abusive and hateful behavior on social media platforms, we present an ensemble learning method to identify and analyze the linguistic properties of such content. Our stacked ensemble comprises of 3 machine learning models that capture different aspects of language and provide diverse and coherent insights about inappropriate language. The proposed approach provides comparable results to the existing state-of-the-art on the Twitter Abusive Behavior dataset (Founta et al. 2018) *without* using any user or network-related information; solely relying on textual properties. We believe that the presented insights and discussion of shortcomings of current approaches will highlight potential directions for future research.

Introduction

Inappropriate language on social media has raised concerns among the users as well as the moderators. Multiple manifestations of online inappropriateness, involving abuse, hate speech, sexism, racism, cyberbullying, and harassment, further amplifies the problem of its categorization given the complex interactions between each of these categories. This has encouraged a lot of research on the topic, including defining the multiple "faces" of inappropriateness and curating crowdsourced datasets (Founta et al. 2018), building machine learning (ML) models to identify abusive and hateful content (Founta et al. 2019), and highlighting the risk of deploying such models trained on biased datasets (Sap et al. 2019). An important direction that is relatively less explored is the analysis of inappropriate content that is facilitated by the learned patterns of various ML models. We believe that this is an important direction because the ML models can not only identify inappropriate language, but can also aid in highlighting aspects of language that make it so. The understanding of these aspects can be used to further improve systems to perform more robust and bias-free identification.

In this work we primarily focus on two aspects relating to inappropriate language: (1) training ML models to *identify* online inappropriateness, (2) *analyzing* the learnings of trained ML models to gain insights into inappropriate language. Furthermore, we believe that some of our findings can give insights into possible directions for future research.

In spirit of the Task 2 of ICWSM'20 Data Challenge, we conduct experiments on the Twitter Abusive Behavior dataset (Founta et al. 2018) wherein we consider four categories – normal, spam, abusive, and hateful. We train an ensemble of three sufficiently diverse machine learning models to classify tweets into these categories: (a) a logistic regression classifier on pyscholinguistic features, (b) a bag of n-gram based classifier, and (c) an attention-based bidirectional LSTM classifier. The choice of these models is governed by their ability to provide not only good identification/classification results, but also interpretable insights based on their learned parameters. We then concatenate the predictions of these three models to learn a stacked ensemble (again, a logisitic regression classifier). Our models give results on par with the state-of-the-art (Founta et al. 2019) *without* using network or user metadata; solely relying on linguistic properties contained within the tweet. In the following sections we discuss these classification models and the associated insights. We end with a discussion about the difficulty of learning subtle differences in abusive and hateful tweets, while giving some possible future directions.

Classification Experiments

We conduct classification experiments on Twitter Abusive Behavior dataset (Founta et al. 2018). The dataset comprises of $\sim 100,000$ tweets classified into four categories: normal (53.85%), spam (27.15%), abusive (14.04%), and hateful (4.96%). We preprocess the tweets by converting all letters to lowercase. To limit the vocabulary size, we drop the hashtag symbol (#) while keeping the following tag-word as it often carries crucial information. We replace all user mentions with a common token (e.g., `@TheTweetOfGod` is replaced by `user_tag`); similarly web links are replaced by the token `web_link`. Additionally, we remove all non-alphanumeric characters from the tweets (except spaces and punctuation). Following this preprocessing, we train machine learning models to solve the multiclass classification task. For training and validating the models, we split the entire dataset into train, validation, and test sets in $0.8 : 0.1 : 0.1$ ratio. For consistency, these sets are kept the same while training and evaluating all the models. Next, we discuss the models and the insights they offer.

Model	Accuracy
LR on LIWC features	0.78
N-gram based Classification	0.80
Attention-based BiLSTM	0.81
Stacked Ensemble	0.83

Table 1: Classification accuracy on the test set.

Logistic Regression on Linguistic Features

LIWC (Pennebaker, Francis, and Booth 1999) is a text analysis software[1] that allows categorization of words into psychologically meaningful categories. Prior works have demonstrated its ability to capture aspects related to "attentional focus, emotionality, social relationships, thinking styles, and individual differences" expressed in language (Tausczik and Pennebaker 2010). We use all the 94 scores obtained for each tweet as its feature representation to train a logistic regression (LR) model. We standardize the input data and remove highly correlated features (i.e., pearson correlation coefficient > 0.9). In Figure 1 we show the top-10 learned coefficients based on their absolute values.

From Figure 1 we can infer that personal pronouns (e.g., I, me, mine, you, we) are not strong indicators of spam, abusive, or hateful content whereas they occur frequently in normal content. Similar observation holds true for words indicating time or duration (until, now, season). This shows that text related to abusive behavior tends to be aloof from the author, the ownership or specifics are avoided in order to disassociate themselves from the message. Abusive and hateful content have a high concentration of swear words, as well as express different forms of affect (use of words such as 'happy', 'cried', 'hurt','ugly','nasty') and emotions (tone), in turn indicating expressions and opinions. Hateful language uses well-formed words (dic) whereas abusive tweets allude to the use of ill-formed jargon. This composition indicates that expressive, well-formed content may typically be hateful, whereas expressive emotional (tone) content is likely to be abusive. From Table 1 we see that LR gives a decent classification accuracy; Figure 3 presents the confusion matrix for a closer look at label-wise predictions.

N-gram based Classification

Joulin et al. (2017) proposed fastText – a simple and efficient baseline for text classification that uses bag of n-gram features to capture partial information about the local word order. We use the fastText library[2] to train a classifier for our task. We train the model for 10 epochs with a learning rate initialized at 0.1. We set the maximum length of n-grams to 3. The trained model provides vector representations (or embeddings) of the sentences as well as for the words in the vocabulary. The word embeddings allow us to execute nearest neighbor queries as well as perform analogy operations. For instance, we find that the nearest neighbors for offensive words like 'fu*king' or

[1]https://liwc.wpengine.com/

[2]https://github.com/facebookresearch/fastText

Normal	Spam	Abusive	Hateful
business	hoodies	jack*ss	ret*rds
gather	advertise	fu*king	spitt*ng
snapped	online	bruh	n*zi
holds	store	di*khead	ch*ke
freaking	horoscopes	fat*ss	b*tch

Table 2: Some of the most attended words for each class by the attention-based BiLSTM.

'fu*k' are also offensive, yet *diverse*[3], in nature – a*sholes, bullsh*t, su*ks, pen*s, dumba*s, sh*tty, etc. We believe that the nearest neighbor querying that this approach enables can be used to expand on the dictionary of offensive words. Furthermore, we observe interesting word-level analogies like (word2vec (Mikolov et al. 2013) responses are given in parentheses for reference):
(a) fu*king − abuse + normal = boring (w2v: f_**_king)
(b) fata*s − hate + normal = pathetic (w2v: sh*thead)
(c) b*tch − hate + normal = nasty (w2v: haters)

In absence of the context these words are used in, it is difficult to interpret these analogies. However, there is a clear shift *away* from inappropriate expression *toward* more acceptable words while preserving the broad meaning. These analogous words can have a potential use in suggesting "milder" words to the users as they express themselves on digital platforms or to do counterfactual modeling of abusive and hateful tweets. In Figure 2 we show the t-SNE plot of tweet embeddings obtained from the trained model. In inset, we highlight the tendency of this model to classify hateful tweets as abusive tweets; an observation that is reasserted by the numbers presented in the confusion matrix in Figure 3. Given that this tendency is consistent across all the models under consideration, we revisit this observation later.

Attention-based Bidirectional LSTM

Zhou et al. (2016) proposed a bidirectional LSTM model with attention. Being bidirectional in nature, the model can encode both the left and right sequence context in natural language. The attention module allows the model to "attend" to input words while performing classification or generation tasks. The attention weights are often used to interpret which input words were crucial for a given prediction – higher the attention weight, larger is the contribution of that word toward the prediction. Owing to the remarkable modeling capabilities of LSTMs and interpretability of attention module, we train the attention-based BiLSTM model to perform our classification task. We initialize the word embeddings using GloVe representations trained on Twitter corpus (Pennington, Socher, and Manning 2014) and train the model[4] for 4 epochs with a learning rate initialized at 10^{-3}. Following training, we compute the class-wise average of attention weights for all the words in train, validation, and test examples. Table 2 shows some of the "highly attended"

[3]For reference, word2vec (Mikolov et al. 2013), widely used word embeddings, lists fu@kin, f_ck, f_*_cking, friggin, freakin, fu@ked, (censoring by *us*, in *this particular* list, is done by using '@' symbol) etc. as nearest neighbors of the word fu*king.

[4]https://github.com/TobiasLee/Text-Classification

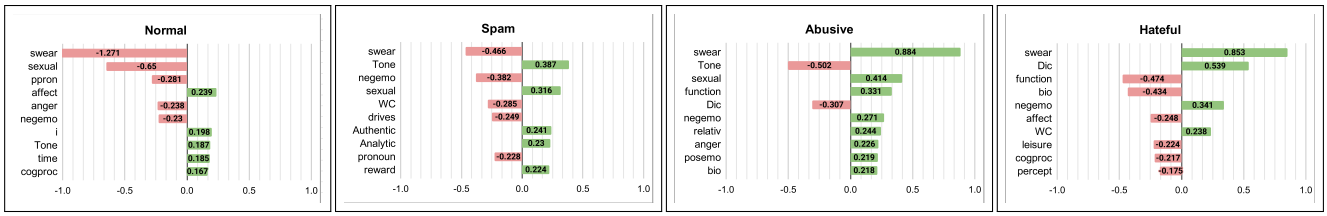

Figure 1: Logistic Regression on LIWC features: for each class, we depict the top-10 learned coefficients based on their absolute values and the corresponding features. The figure is best viewed on screen with zoom.

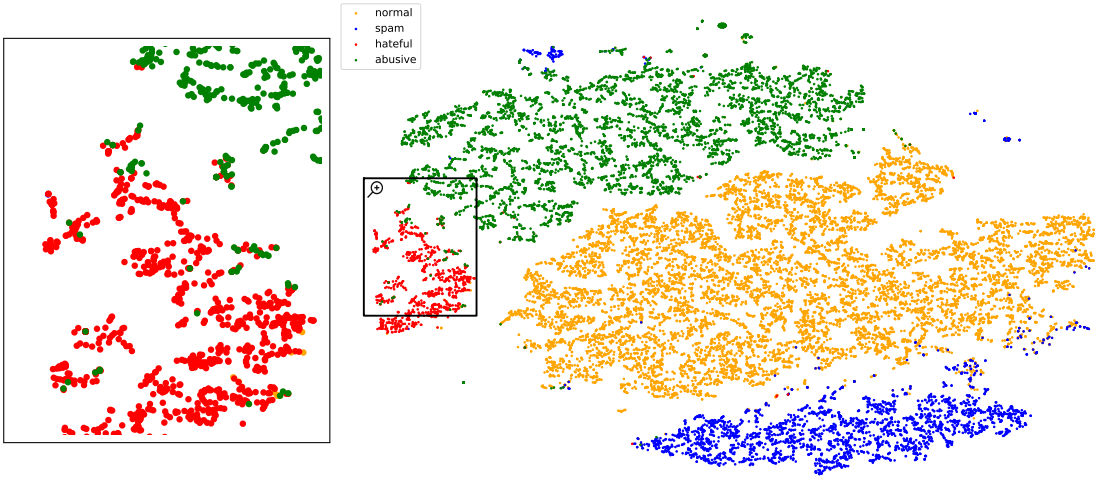

Figure 2: t-SNE plot of sentence embeddings obtained from the n-gram based classifier. In inset, we show that many abusive tweets have similar embeddings as hateful tweets.

words for each of the 4 classes and gives an estimate of the words that are considered important by the model for making a certain prediction – in our case, classifying into one of the 4 categories. For instance, 'freaking' – a euphemism for 'fu*king' is considered important for classifying tweets as `normal`. Even though the words under `abusive` and `hateful` categories seem similar in nature, there are certain subtle differences. Founta et al. (2018) claim that hateful tweets often contain a well-defined description of the target group(s) whereas abusive tweets do not. This is reflected in Table 2 as words like 'ret*rds' and 'n*zi' are often used to target groups unlike words like 'jacka*s' and 'fata*s'. Interestingly, the words under `spam` category are also often encountered in advertisements and posts by Twitter bots.

Stacked Ensemble

Model stacking for ensemble learning involves taking the probability estimations of *base models* and using them as features for training a *meta model*. The general practice is to take diverse and complex base models that make sufficiently different assumptions to solve the predictive task, and then train a simple meta model to "interpret" these predictions. We treat the above three models as base models and use their predictions over the train (and validation) examples to train (and validate) a logistic regression model (i.e., the meta model). We use the predictions over the test set to evaluate the meta logistic regression model. As we note in Table 1, the stacked ensemble performs better than all the base models. This reaffirms the diverse modeling assumptions argument that we presented earlier. One consistent shortcoming of the base models, as it is evident from the confusion matrices in Figure 3, is the tendency to incorrectly classify tweets that are actually `hateful` as either `abusive` or `normal`. It is encouraging to see that the stacked ensemble has a better capability to distinguish between these classes.

Discussion and Future Directions

This section discusses the presented work in light of the past works and points out some potential directions of future research. We start with a comparison of proposed methods with the existing state-of-the-art, and then discuss some of the shortcomings of our methods.

Overall accuracy: Founta et al. (2019) proposed a unified deep learning architecture that uses textual, user, and network features to detect abuse. It is encouraging to see (Table 1) that our stacked ensemble performs on par with their method *without* using *any* user or network features[5]. Our methods solely rely on textual properties of the tweets. While this speaks for the competitiveness of our approach, more importantly, it highlights a potential future direction to boost the classification performance of our models.

Class-wise predictions: As we mention earlier, the tendency to incorrectly classify `hateful` tweets as `abusive`

[5]Founta et al. (2019) perform their experiments only on 3 classes: `normal`, `abusive`, and `hateful` and report 0.84 accuracy. Whereas, our models, when trained to classify only among these 3 classes give $\sim$ 0.89 accuracy.

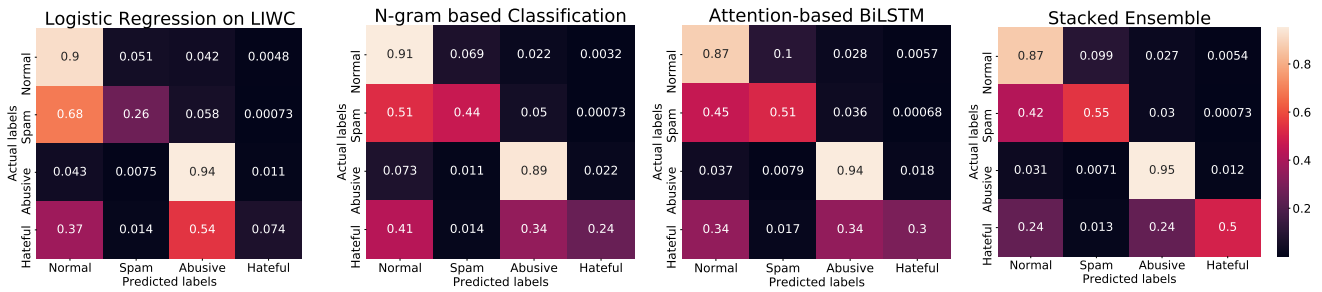

Figure 3: Confusion matrices: the values indicate the fractions of examples in the test set that were classified as any given label.

tweets is consistent across all the models. Even though the stacked ensemble mitigates this to a reasonable extent, the problem still persists. Similarly, the tendency to incorrectly classify `spam` tweets as `normal` tweets is consistent across all the models; stacked ensemble still being better than others. We believe that the `spam`-`normal` confusion can be handled well by incorporating user and network features – spam tweets usually come from bot accounts and few users tend to spam repeatedly. However, since the `hateful`-`abusive` confusion is central to our discussion around linguistic properties of inappropriate tweets, we present some insights in the following paragraphs.

Abusive or Hateful?

When crowdsourcing the data, Founta et al. (2018) found that even though the the label `hateful` frequently co-exists with other labels like `abusive`, `offensive`, and `aggressive` (all of which are later merged into the `abusive` category), it is not significantly correlated with any other label. Their definition of hateful tweet emphasizes on the well-defined description of the target groups. As mentioned earlier, this is also indicated in the highly attended words we present in Table 2. Owing to this fundamental difference in hateful and abusive tweets, which is largely captured in the linguistic properties of the tweet (as opposed to user or network-related properties), we consider it important to be able to distinguish between these two categories.

The tendency of our models to confuse between these two categories can be, in part, attributed to the training data. There is a clear imbalance in data – hateful tweets only account for $\sim$ 5% of examples. While random oversampling of minority class examples helps in mitigating the consequences of this imbalance, the results are still far from great. Furthermore, the average number of annotators that agreed on `hateful` label for the tweets is lowest when compared to other labels – i.e., 2.95 out of 5 for `hateful` in comparison to 3.90, 3.47, and 3.53 for `normal`, `spam`, and `abusive`, respectively. This indicates a lack of agreement among the annotators. However, from a linguistic modeling perspective, many hateful tweets are very similar to abusive tweets. For instance, "i'm fu*king done with twitter" (`hateful`) vs "i'm fu*king done" (`abusive`); "when a n*gga got you fu*ked up web_link" (`hateful`) vs "b*tch you got me fu*ked up web_link" (`abusive`); "some women need to grow the hell up. it's so pathetic." (`hateful`) vs "some people are so pathetic and need to grow the fu*k up!" (`abusive`). These examples further illustrate the point that there's a well-defined target group within hateful tweets and explicitly incorporating that information might be a promising direction for future research.

Conclusion

In this work we discussed three ML models to classify tweets as `normal`, `spam`, `abusive`, and `hateful`. We then discussed a meta logistic regression model that uses the predictions of these classifiers as features to solve the classification task. Our three base-models provide valuable insights regarding inappropriate tweets: the logistic regression model trained on LIWC features provides insights into psycholinguistic patterns that are emergent in such tweets, the n-gram based classifier provides word embeddings that are tuned with respect to abusive and hateful behavior, and the attention-based BiLSTM highlights some of the important words that influence its predictions. Furthermore, our stacked ensemble provides classification accuracy that is comparable to the state-of-the-art by only using textual properties. Lastly, we discussed the shortcomings of our proposed approaches as well as the subtle linguistic differences in abusive and hateful tweets with a hope that it will influence future research on the topic.

References

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*.

Founta, A. M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2019. A unified deep learning architecture for abuse detection. In *ACM Conference on Web Science*.

Joulin, A.; Grave, É.; Bojanowski, P.; and Mikolov, T. 2017. Bag of tricks for efficient text classification. In *EACL*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.

Pennebaker, J.; Francis, M.; and Booth, R. 1999. Linguistic inquiry and word count (liwc).

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*.

Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *ACL*.