Odium Revelio! - Detecting subtle hate speech in online conversations

AAAI Press
Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303

Abstract

While social media can serve to bring together people in harmony, it also has the power to trigger hatred, abuse and toxicity if misused. While there has been considerable research focus on combating hate speech online, it has been widely acknowledged that this continues to remain an unsolved problem. Automated approaches for detecting hate speech and toxic comments have been quite effective in detecting direct unambiguous hate speech. However nuanced and subtle expressions of hate speech are often missed by the state of the art hate speech classification systems. In this paper, we propose a new approach for hate speech classification which can help identify subtle and nuanced hateful comments. We evaluate our approach for hate speech classification on an updated FoxNews comments dataset and show that it is effective.

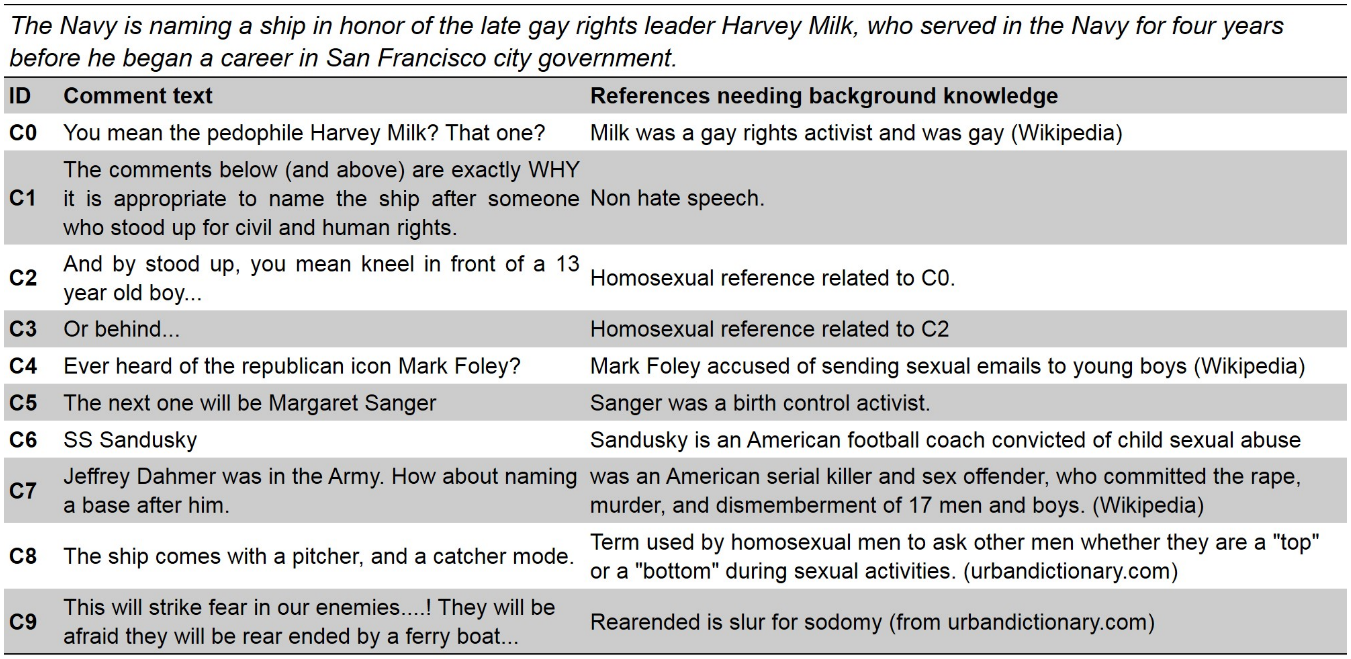
Introduction

While social media can serve to bring together people in harmony, it also has the power to trigger hatred, abuse and toxicity if misused. On one hand, social media platforms have helped form strong social connections between people far-flung in the world, while on the other hand, these very same platforms have also been the silent culprits in fanning the flames of hate speech and toxic content directed across people from different socio-ethnic groups. Of late, there has been considerable research focus on detecting and combating hate speech and toxic content present on online social media platforms. However it has been widely acknowledged that this continues to remain an unsolved problem. The challenges of analysing and addressing offensive content online are multi-fold. There are open questions around *what constitutes offensive content?* and whether imposing such controls on online content can lead to infringement of the basic freedom of speech. In addition to these, there are significant technical challenges in detecting toxic content efficiently due to the innate inability of machines to comprehend implicit/fine grained emotions, beliefs, opinions in human expressions.

Automated approaches for detecting hate speech and toxic comments have been quite effective in detecting toxic content containing explicit profanity and hate speech. Unfortunately human beings have the expressive power to express toxicity in subtle and implicit ways which fall below the radar of explicit hate speech detectors. Nuanced and subtle expressions of hate speech are often missed by the state of the art hate speech classification systems. We illustrate this with an example shown in Figure-1 below:

These comments are real-life examples posted on a Fox News Article. At first glance, each of these comments appear innocuous and non-toxic. None of the existing hate speech classification systems mark any of these comments as hate speech since there is nothing on the surface form of the text to indicate that it contains toxic comment. However comments 2-9 are examples of subtle hate speech. Human beings with the relevant background knowledge of the topic being discussed, would be easily able to infer the subtle toxic content hidden in those comments. We show this in the last column of Figure-1 where each of the comments is accompanied by additional background information showing why it is a toxic comment. For instance, the comment C2 does not contain any explicit hate speech, but it subtly talks about the child abuse charges levelled against the person `Harvey Milk` after whom the navy is naming the ship by using the words *kneel in front of a 13 year old boy*. Comment C3 seems quite innocous as it just says *Or behind.* From the surface form of comment C3, it is not a hateful comment. It has to be seen along with the previous comment in context to realize that it is a subtle hateful comment against the pedophile charges levelled against Harvey Milk. Comment C6 is a subtle reference to the baseball coach Jerry Sandusky accused of sexual abuse of young boys. Comment C8 contains the phrase textitpitcher and catcher. While the surface form of C8 does not contain any clues to its being offensive, we need the background information that textitpitcher and catcher is a subtle reference to gay persons as seen from external knowledge source *Urbandictionary.com*. These examples show that for identifying subtle and ambigous hateful comments, additional background information on the article, previous comments and external world knowledge are needed.

This background and topical knowledge available to humans makes it easy for them to detect such subtle nuances of hateful comments. On the other hand, the subtle and implicit hateful comments are extremely challenging for automated approaches to classify correctly and as we pointed out, most of the existing state of the art hate speech classifiers get all



The Navy is naming a ship in honor of the late gay rights leader Harvey Milk, who served in the Navy for four years before he began a career in San Francisco city government.

ID	Comment text	References needing background knowledge
C0	You mean the pedophile Harvey Milk? That one?	Milk was a gay rights activist and was gay (Wikipedia)
C1	The comments below (and above) are exactly WHY it is appropriate to name the ship after someone who stood up for civil and human rights.	Non hate speech.
C2	And by stood up, you mean kneel in front of a 13 year old boy...	Homosexual reference related to C0.
C3	Or behind...	Homosexual reference related to C2
C4	Ever heard of the republican icon Mark Foley?	Mark Foley accused of sending sexual emails to young boys (Wikipedia)
C5	The next one will be Margaret Sanger	Sanger was a birth control activist.
C6	SS Sandusky	Sandusky is an American football coach convicted of child sexual abuse
C7	Jeffrey Dahmer was in the Army. How about naming a base after him.	was an American serial killer and sex offender, who committed the rape, murder, and dismemberment of 17 men and boys. (Wikipedia)
C8	The ship comes with a pitcher, and a catcher mode.	Term used by homosexual men to ask other men whether they are a "top" or a "bottom" during sexual activities. (urbandictionary.com)
C9	This will strike fear in our enemies....! They will be afraid they will be rear ended by a ferry boat...	Rearended is slur for sodomy (from urbandictionary.com)

Figure 1: Subtle and Ambiguous Hate Comments

these examples wrongly classified as non-toxic. In this paper, we focus on the problem of detecting subtle and nuanced hateful comments. Most of the earlier work on hate speech classification have focussed on direct unambiguous hate speech detection. They do not call out the distinction between direct unambiguous hate speech vs subtle nuanced hate speech and do not propose techniques which can address the challenge of detecting subtle hate speech.

As we pointed out earlier, almost all the state of the art hate speech classifiers perform very poorly on subtle hate speech detection, while humans are able to perceive the toxicity in these comments. We hypothesize that human beings have the required background knowledge associated with the topic on hand, to detect such subtle toxic comments, which the automated approaches lack. Based on this assumption, we hypothesize that models detecting subtle nuanced hate speech would benefit from the addition of background knowledge associated with the topics discussed in the parent post as well as the associated comments.

As seen earlier in Figure-1, the additional side-information which contains the distilled background knowledge related to these comment topics makes the subtle and nuanced hate speech more direct and explicit. This lends weight to our claim that utilizing such additional background information will enable automatic approaches to help classify subtle and ambiguous toxic comments correctly. Based on the above premise, we propose a new approach for hate speech classification which can help identify subtle and nuanced hateful comments by utilizing background information. Background information can come from different sources such as the article on which the comments are being made, world knowledge about the external entities and phrases referenced in the toxic comment and also from the context of the previous comments in the thread. We hypothesize out that such additional background knowledge needs to be modelled in the context of the comment being classified and instead of being considered in isolation. Hence we propose a neural network based approach which models the interaction of this additional background knowledge with the textual content of the comments, thereby enabling it to detect subtle and ambiguous toxic comments.

Existing datasets do not focus on subtle and ambiguous hate speech. A large majority of the existing hate speech datasets contain typically explicit and direct hate speech. We speculate that this is probably due to the fact that these datasets were collected by explicit hate key words or topics as trigger words for collecting hateful posts/comments on online social media platforms. Hence these datasets typically do not reflect the subtle and ambiguous toxic comments present online. Nor do these datasets contain the side-information related to the background knowledge needed for resolving the ambiguous references.

We found one existing dataset (Gao and Huang 2017) which contain significant amount of subtle and ambiguous hate comments since they were created by collecting all comments related to an article/post instead of collecting only comments containing specific hate key words/hash tags/topics. This dataset is a collection of user comments from FoxNews website, which we refer to as FoxNews dataset. We augmented it with external background information for classifying subtle nuanced hate speech. We plan to share the augmented dataset publicly. We evaluate our approach for hate speech classification on this dataset and show that our approach achieves an accuracy of 85%.

Prior work on hate speech detection has focussed more on developing rich feature based classifiers. While neural network models have been proposed for hate speech classification, they have not modeled the problem of subtle and ambiguous hate speech (Davidson et al. 2017; Schmidt and Wiegand 2017; Park and Fung 2017; Vigna et al. 2017). The closest in spirit to our work is the work by Gao et al. (Gao and Huang 2017) where they propose context aware models for hate speech classification. However their contextual model only models the article title and user name and is extremely limited. They model the context independently from the comment representation, by feeding the independent encoded representations of comment, post and title and they do not model the cross-sentence interactions between comment and title. Their model is limited to intra-attention on the comment post and they do not model inter-attention across comment and article title. More significantly, they do not model any form of external background knowledge needed to resolve the ambiguous and subtle hateful references contained in the comments. Our work addresses all these aspects.

In summary, this paper makes the following contributions:

- To the best of our knowledge, we make the first detailed description of subtle hate speech by pointing out real life instances of it and discuss the challenges in detecting it. We also point out that background knowledge needs to be modelled by automated approaches for detecting such subtle hateful comments.
- We enrich an existing dataset with additional background information for the problem of subtle hate speech classification. We develop a neural network based approach for this problem and experimentally evaluate our approach on this dataset.

The rest of the paper is organized as follows. In Section 2, we describe our approach. In Section 3, we discuss our experimental setup and results. In Section 4, we provide a brief overview of related work and conclude in Section 5.

Our approach

In this section, we discuss our approach to hate speech classification, in particular towards handling subtle and ambiguous hateful comments. We first provide a brief overview of our approach and then describe the neural network classifier we developed for this problem.

Overview of our approach

Standard supervised machine learning techniques can classify unambiguous direct explicit hate speech with reasonable accuracy. These approaches typically either use a rich feature driven representation or a neural representation of the text to drive the classification. This works well when the surface form of the text is representative and contains explicit and direct hate speech. On the other hand, ambiguous and subtle hate speech content does not contain representative hateful words in its surface forms. The underlying toxicity of subtle hate speech gets revealed only when seen in combination with external background knowledge beyond the comment text itself.

As we discussed earlier, this background knowledge is available inherently to human readers of the comments, either as part of the world knowledge/commonsense knowledge they possess, or through the contextual information of the article on which the comments are being made as well as from the previous comments in the same conversation thread. In order for the automated approaches to identify the subtle and ambiguous hate speech, this background knowledge is needed as an additional input, as we had illustrated through the examples in Table-1. We emphasize that analyzing the standalone representations of the input comment text and background knowledge in isolation may not help in revealing the hidden hateful nuances in the comment text. Instead it is necessary to consider the cross-text interactions between the ambiguous comments and the background knowledge to surface out the underlying nuances. Based on this intuition supported by the motivating examples we had shown in Figure-1, we propose a neural network classifier for hate speech, which utilizes the combined cross-text interactions of the comment text and the background knowledge.

High level block diagram of our proposed approach is shown in Figure-2. The input comment text and the relevant background knowledge are first encoded into a neural sentence embedding representation. Cross-sentence analysis module captures the cross-text interactions between the comment text and the background knowledge. The combined representation is then fed to a standard Multi-layer Perceptron (MLP) classifier. Unlike explicit hate speech which has strong surface features, ambiguous hate speech does not have strong surface features. Hence instead of an explicit feature driven classifier, we use a neural network classifier for this task which can learn the latent features associated with subtle and ambiguous hate speech. Next we provide a detailed description of the various components of our classifier.

Detailed Description

Our hate speech classifier is intended to be effective against subtle and ambiguous hate speech. Explicit and direct hate speech typically is characterised by hate words in the surface form of the input text so that classifiers can learn to distinguish them easily by these strong markers. However such markers are absent in subtle hate speech. Instead the input comment text typically contains only hints or references which need to be analyzed in the context of additional background information of the article on which the comments are made, the previous comments on the post and any additional commonsense and world knowledge information. Given this need, our approach includes a background information retriever module for retrieving the background information, then appropriately encoding the retrieved background information using an background knowledge encoder module.

The cross-text interactions between the encoded back-

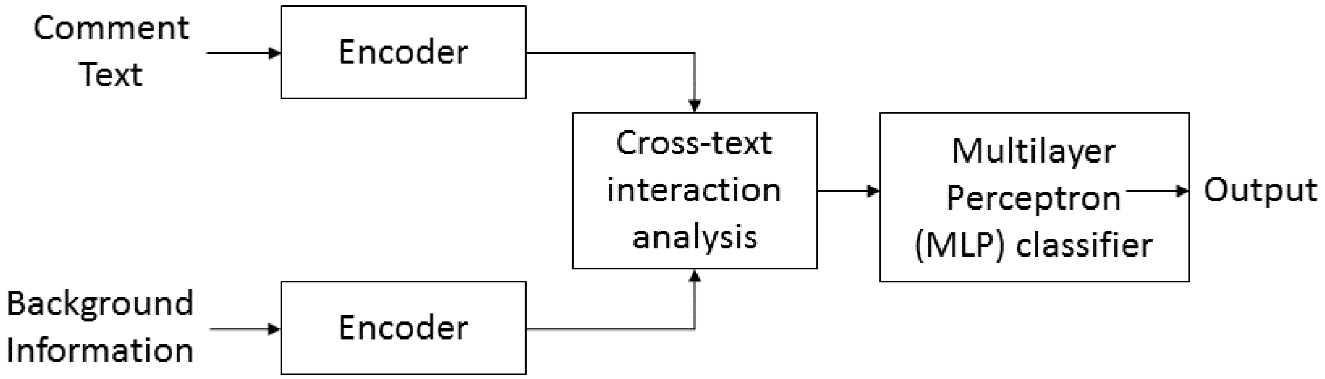

Comment Text
Encoder
Cross-text interaction analysis
Multilayer Perceptron (MLP) classifier
Output
Background Information
Encoder

Figure 2: High Level Block Diagram

ground knowledge representation and the comment text encoded representation are analyzed using inter-attention mechanism to capture their interactions. This inter-attended representation is then combined with standalone encoded representation of the background knowledge in the aggregator module. The output of the aggregator module is fed to the standard Multi-Layer Perceptron classifier. The whole neural network classifier is trained using the labelled data end-to-end, using the standard cross-entropy loss as the error signal.

The functional components of our neural network classifer is shown in Figure-3 . It consists of the standard components of comment text Embedding Module, Comment text Encoding module and the standard Multi-Layer Perceptron (MLP) Classifier module, which can be found in such typical neural classifier pipeline. The additional components are the background knowledge retriever module, background knowledge encoder module, cross-text interaction module and the aggregator module.

The input comment text is passed to the embedding module. The task of the embedding layer is to generate the sentence matrix given a textual sentence. In the comment embedding module, the word embedding sentence matrix is built for the input comment text using word embeddings (Mikolov et al. 2013). The comment text is then encoded using a recurrent neural network based encoder. As is common practice in several NLP tasks, we use Recurrent neural networks (RNNs) with long short-term memory (LSTM) [6] units as our encoders for the basic sentence representation. Hence the input comment text is hence encoded using a LSTM (Hochreiter and Schmidhuber 1997).

The key advantage of LSTM is that it contains memory cells which can store information for a long period of time and hence does not suffer from the vanishing gradient problem. LSTMs contain memory cells that can remember previous state information as well as three different types of gates namely input gates (Equation 2), forget gates (Equation 3) and output gates (Equation 4) which control how much of the information is remembered. Given an input vector x_t at a time t, and the previous output as h_{t-1} and previous cell output c_{t-1} , the current cell state and output are computed by the following equations:

$$H = \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix} \tag{1}$$

$$i_t = \sigma(W^i H + b^i) \tag{2}$$

$$f_t = \sigma(W^f H + b^f) \tag{3}$$

$$o_t = \sigma(W^o H + b^o) \tag{4}$$

$$o_t = \sigma(W^o H + b^o) \tag{5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W^c H + b^c) \tag{6}$$

$$h_t = o_t \odot tanh(c_t) \tag{7}$$

In neural sentence representation methods, attention mechanism has been shown to be effective in improving classification performance by assigning higher weightage to relevant words of the sentence (Bahdanau, Cho, and Bengio 2014). Attention can be intra-attention (also known as self-attention), wherein attention weights are learnt from the same input sentence which is getting represented or it can be inter-attention mechanism wherein attention weights for an input text encoding are learnt from the encoded representations of a related text (Bahdanau, Cho, and Bengio 2014; Vaswani et al. 2017). As part of the comment text encoding, we also model intra-attention on the comment text.

As mentioned before, background information includes the article summary on which the comments are made, the previous comment in the comment thread as well as external knowledge on the entities/key phrases mentioned in the sentence. The background information retriever module retrieves the relevant information related to the current comment. While the article summary and previous comment are typically readily available, the information on entities and key phrases mentioned in the comment text are obtained by considering two external knowledge sources namely the Wikipedia and Urban Dictionary.

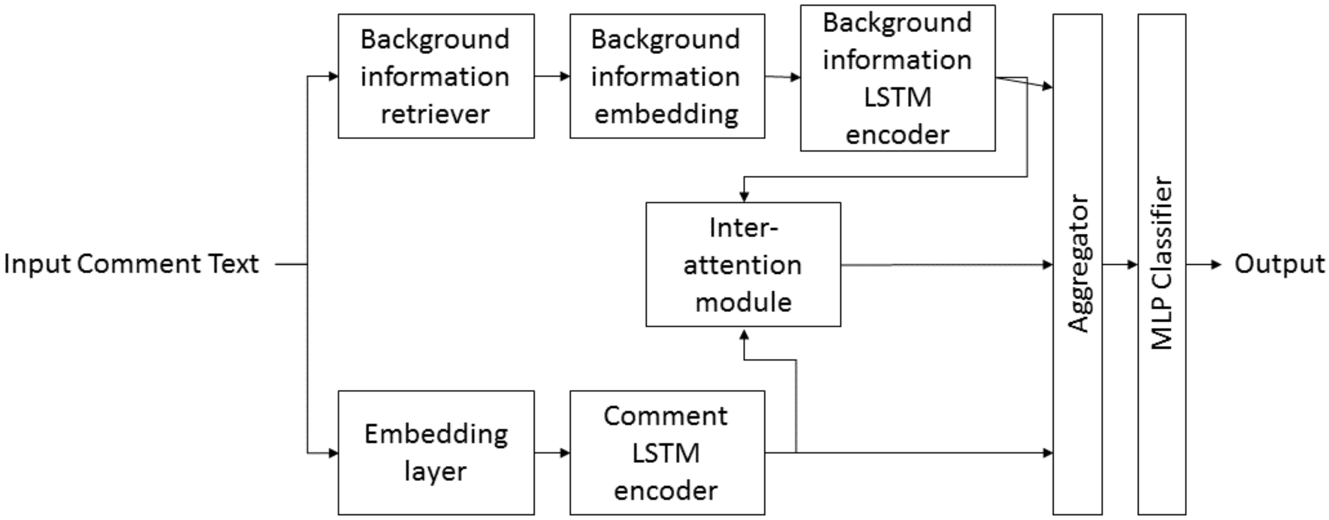


Figure 3: Functional Components of our Approach

The retrieved background information is also passed through an embedding layer to obtain the background information embedding matrix and then encoded using the background information encoder LSTM module. As in the case of comment text encoding, we use intra-attention on the background information embedding as well. We denote the outputs of comment text encoder LSTM matrix as Y_c and background information encoder LSTM output Y_B respectively.

The comment text encoding representation is then conditioned using the background knowledge representation for capturing their cross-text interactions. This conditioning is done in the cross-text interaction module by using inter-attention mechanism. Our cross-text interaction module is implemented using inter-attention mechanism between the input comment text and background information. Let Y_C be the output matrix produced from comment text LSTM and Y_B be the output matrix produced from background information encoder. The output of the cross-text interaction module is generated as follows:

$$M_1 = tanh(W_1 Y_c + W_2 Y_B) \tag{8}$$

$$\alpha = softmax(W^T M_1) \tag{9}$$

$$O_1 = \alpha Y_c \tag{10}$$

At the aggregator module, the conditioned output of the comment text conditioned on the background information, namely O_1 is concatenated with encoded representations of the background information. The output of the aggregator module is then fed to the standard MLP classifier. Hence the input to MLP classifier would be (Y_B, O_1). The MLP classifier is a standard multi-layer fully connected feed forward network, with the last layer being a softmax layer with 2 units. During our evaluation, we report results on different variants of this approach, wherein we consider combinations of using (a) article summary (b) using previous comment (c) externally retrieved knowledge as background information.

Experimental Results

In this section, we discuss our dataset and experimental results.

Dataset

Most of the existing hate speech datasets gathered from social media platforms have been collected by using explicit set of trigger words associated with hate and offensive language to collect an initial set of samples which are then manually annotated. Most of the prior work have adopted this approach of data collection and annotation. This is due to the fact that only an extremely small percentage of posts contain hateful or offensive language and it becomes cumbersome to look for hate speech in random samples of social media posts. However collecting samples using trigger words for hate speech leads to the dataset more focused towards explicit surface forms of hate speech and does not lend itself to a dataset which contains subtle and ambiguous hate speech.

We found one dataset consisting of comments made on articles of controversial topics made on FoxNews website (Gao and Huang 2017), in which all comments in the threads were collected without explicitly looking for any specific hate word or offensive triggers. The dataset contains the internal background information on the title of the article on which comments are made, the summary of the article. We augmented the data set by adding the previous comment in the conversation thread for each comment as well as manually annotated external background information for each comment. The external background information consists of information on named entities and phrases mentioned in the comment text, related to the comment post and manually added by human moderators. To illustrate examples of external background information, consider the comment SS Sandusky. The term Sandusky refers to Jerry Sandusky, a former baseball coach convicted of sexual abuse against young boys and this information is added as external background infor-

Table 1: Experimental Results for Fox-News Dataset

Method	**Accuracy**	**Precision**	**Recall**	**F-Score**	**AUC**
Baseline svm	0.68	0.61	0.68	0.60	0.52
baseline gradient-boost	0.70	0.67	0.70	0.59	0.51
comment only	0.73	0.55	0.36	0.43	0.71
comment + title	0.75	0.58	0.46	0.51	0.74
comment + summary + title	0.76	0.60	0.47	0.67	0.83
comment + external	0.70	0.50	0.59	0.54	0.71
comment +previous	0.74	0.59	0.53	0.55	0.73
comment + prev + title	0.74	0.59	0.50	0.54	0.76
comment + title + external	0.75	0.65	0.42	0.51	0.78
comment + summary(intra) + external (intra)	0.77	0.60	0.72	0.66	0.84
comment + external (intra) + title (intra)	0.81	0.73	0.63	0.68	0.86
comment + title (inter) + external (intra)	0.82	0.81	0.82	0.81	0.76
comment + summary (inter) + external (intra)	0.81	0.80	0.81	0.80	0.76
comment + prev(inter) + external(intra)	0.82	0.81	0.82	0.81	0.76
comment + external (intra) + prev	0.85	0.89	0.56	0.69	0.86
comment + summary (inter) + external (intra) + prev	0.85	0.84	0.85	0.84	0.79

mation. We used this augmented FoxNews comment dataset for experimental evaluation and we plan to make this publicly available.

Experiments

Our implicit neural classifiers were implemented using the Tensorflow library (Abadi 2016). The model training objective is cross-entropy loss. We use an Adam Optimizer with a learning rate of 0.001 and batch size of 32 for both our schemes. We used standard pre-trained word embeddings which were not modified during training. Out of vocabulary words were handled by randomly initialized embeddings generated by sampling values uniformly from -0.05 and 0.05. The comment sequence was pruned to a sequence of length 150, selected as part of hyper-parameter tuning. The hidden size for all LSTMs was 150 and the MLP classifier consists of a fully connected layer of dimension 150 followed by softmax. The baseline classifier was modelled on the standard hate speech classification classifiers used in prior work, with features being standard bag of words and presence of hate trigger words from hate speech lexicons from [1]hatebase.org. We report two baseline classifier model results, one using gradient boosting and other using Support Vector Machine.

As mentioned before, we consider different types of background information namely article title, article summary, previous comment in the article, and external knowledge. We also considered applying intra-attention mechanism on the background information concatenated with neural sentence encoding of the comment post as well inter-attention mechanism between comment post and background information. The sequence length for external information was set to 184, and for article title and summary, it was set to 60 and 150 respectively. We report the results of our experimentation in Table 1.

We find that while baseline classifers exhibit an accuracy of 68% with Support vector machine classifier and 70% with gradient boosting classifer, our neural network classifiers with background information are able to achieve accuracies of upto 84% (obtained with inter-attention between comment post and article summary, intra-attention on external information and previous comment encoding aggregated). We find that adding background information using article summary and previous comment achieves an accuracy of 73% and 74% respectively, without applying any attention mechanism. Adding external information in standalone only achieves 71% accuracy. We find that article title, summary and previous comment provide stronger and relevant background information compared to external information. Adding intra-attention mechanism on comment post along with external information achieves an accuracy of 76%. This shows that intra-attention helps to give higher weightage to relevant words in the comment text for classification. Adding inter-attention mechanism for capturing cross-text interactions between comment text and article summary improves performance to 81%. The addition of previous comment as background information improves accuracy further to 85%.

However we also find that all of background information does not additively improve performance. While previous comment and external information independently improve performance when added as background information, their combination is not effective as seen in the Table above. We hypothesize that this may be due to the noisiness of the added background information hampering correct prediction. We plan to investigate this as part of future work.

For external knowledge on the entities and phrases in the comment text, we had two options. One is to use the external information manually annotated in the dataset, and other is to automatically retrieve the external knowledge from sources such as Wikipedia and UrbanDictionary. We used

[1] https://www.hatebase.org/

the [2]wikipedia and [3]urban dictionary APIs to retrieve the respective pages corresponding to the entities and phrases mentioned in the comment text. We pruned this to a summary by considering semantic relatedness between the retrieved information and the article summary using simple document-vector similarity. However we found that automatically extracted external background knowledge resulted in very noisy information which was not usable. Hence in all the experiments above, we had used the manually annotated external knowledge background information available in the dataset we had created. However we note that not all datasets would have this external knowledge available and hence this is a disadvantage in our current approach. Hence we plan to investigate efficient automatic retrieval of relevant background external knowledge for entities and phrases in comment text as part of future work.

We find that while our approach can correctly classify many of the ambiguous and subtle hateful comment examples we illustrated in Table-1, it fails to classify correctly some of the following examples: *just like Milk, this ship will be full of sea-men.*. The implicit reference is to the offensive term *semen* which sounds alike to the actual word *sea-men* used in the comment text. *The USS Qeerbarge*. The implicit reference is to the offensive term *queer* referring to homosexuals, which sounds alike to the word ***qeer**barge* used in the comment text. None of our current background information is able to help such ambiguous references since these are *sound alike* references. We plan to address such soundalike subtle hate references as part of future work.

Related work

There has been considerable prior work on detecting offensive content, hate speech and profanity in online content (Schmidt and Wiegand 2017; Davidson et al. 2017; Warner and Hirschberg 2012; Waseem and Hovy 2016; Wulczyn, Thain, and Dixon 2016; Burnap and Williams 2015). Davison (Davidson et al. 2017) proposed classifying hate speech and offensive language in tweets using Logistic Regression classifier, using explicit set of features such as TF-IDF counts of unigrams, bigrams and trigrams, sentiment, binary and count indicators for hashtags, mentions, retweets, and URLs, as well as features for the number of characters, words, and syllables in each tweet. Their dataset consisted of tweets of hate speech and offensive language which were collected by looking for the presence of explicit terms present in a well known hate speech lexicon available at Hatebase.org. There has been a number of works which had used rich set of features for hate speech classification. Such features include standard bag of word features, word and character n-grams, part of speech tags, dependency relations, sentiment, word generalization features, word embedding vectors etc. However these features have typically depended on explicit surface forms of hate speech being present in the input text.

Of late, a number of deep learning methods have also been proposed for hate speech classification (Park and Fung 2017; Vigna et al. 2017; Schmidt and Wiegand 2017; CNN). A detailed survey of various hate speech classification methods can be found in (Schmidt and Wiegand 2017). However most of the prior work has focused on detecting direct and explicit hate speech and offensive content. Davison (Davidson et al. 2017) in fact point out in their paper that tweets without explicit hate keywords are in fact hard to classify. Also the datasets that have been used in many of these methods have collected hate speech samples by looking for the presence of specific keywords/triggers from hate speech lexicon. This makes these classifiers learn explicit hate speech patterns well, but makes them ineffective against subtle and ambiguous hate speech and offensive content. On the other hand, our work is focused on detecting subtle and ambiguous hate speech and hence is complementary to most of these earlier works.

The closest in spirit to our work is the prior work by Li Gao et al. (Gao and Huang 2017) where they propose context aware models for hate speech classification. However their contextual model only models the article title and user name and is extremely limited. They model the context independently from the comment representation, by feeding the independent encoded representations of comment, user name and title. They do not model the cross-sentence interactions between comment and title as their model is limited to intra-attention on the comment post. More significantly, they do not model any form of external background knowledge needed to resolve the ambiguous and subtle hateful references contained in the comments. Our work addresses all these aspects.

Conclusion

In this paper, we proposed a new approach for online hate speech classification, focusing on improving the detection of subtle and ambiguous hate speech. We performed detailed experimental evaluation of our approach and showed that our approach is effective. We found that using background information from external knowledge sources can be noisy and plan to address this issue in future work.

References

Abadi, M. e. a. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, 265–283. Berkeley, CA, USA: USENIX Association.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Burnap, P., and Williams, M. L. 2015. Cyber hate speech on twitter : An application of machine classification and statistical modeling for policy and decision making.

Using convolutional neural networks to classify hate-speech.

Davidson, T.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. *CoRR* abs/1703.04009.

[2] https://pypi.org/project/Wikipedia-API/

[3] https://github.com/bcyn/urbandictionary-py

Gao, L., and Huang, R. 2017. Detecting online hate speech using context aware models. *CoRR* abs/1710.07395.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. 9:1735–80.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Park, J. H., and Fung, P. 2017. One-step and two-step classification for abusive language detection on twitter. *CoRR* abs/1706.01206.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *SocialNLP@EACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *CoRR* abs/1706.03762.

Vigna, F. D.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on facebook. In *ITASEC*.

Warner, W., and Hirschberg, J. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, 19–26. Stroudsburg, PA, USA: Association for Computational Linguistics.

Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*.

Wulczyn, E.; Thain, N.; and Dixon, L. 2016. Ex machina: Personal attacks seen at scale. *CoRR* abs/1610.08914.