CSE556: Natural Language Processing Final Report
Hate Speech Implication and Generation

Atishay Gupta 2021241 | Mahir Yadav 2021263 | Naman Sharma 2021266 | Vianshu Shalyan 2021298

Abstract

***Warning**: this paper contains content that maybe offensive or upsetting.*
Social media can be really good for sharing information, but it also makes it easy for people to say hurtful things to others. This is called hate speech, and it's a big problem because it can really hurt people's feelings and cause problems in society. So, it's important to stop hate speech because it can have bad effects on both the community and individuals.Therefore, it is not only important for models to detect these speeches, but to also output explanations of why a given text is toxic. While plenty of research is going on to detect online hate speech in English, there is very little research on the explainability aspect of hate speech. Providing descriptions of internal stereotypical bias as an explanation of hate posts makes a hate speech detection model more trustworthy. To do this, we're using a tool called ConceptNet. ConceptNet helps the system understand the text better by adding common sense knowledge.

1. Introduction

As more people use social media the amount of text being created and shared has grown incredibly fast. Among this text hate speech is a major problem. Hate speech includes any messages that hurt or insult people because of their race, gender, religion, or other personal traits. Stopping hate speech is important because it can lead to violence and increase divisions in society. Because there's so much online content it's too hard for people to check everything manually. That's why we need automated tools to help find and manage hate speech. However many of the current tools don't explain why something is considered hate speech also everyone is doing hate speech detection which makes it difficult for users to understand or trust the results.We are generating stereotype targets which users can trust on. This project aims to make an automated system that not only spots hate speech but also explains clearly why a piece of text is flagged. To do this, we're using a tool called ConceptNet. ConceptNet helps the system understand the text better by adding common sense knowledge, like how certain words or ideas are connected in real life. This should make the explanations more understandable and relevant. By improving how the system explains hate speech, we can help make social media safer and more welcoming for everyone.

2. Motivation

Problem of Scale: Millions of messages are shared on social media platforms everyday.It's impossible for human moderators alone to keep track of all potentially harmful content, including hate speech because of such large volume of messages. Automated tools are necessary to help with this task, but these tools need to be both effective and trustworthy.
Lack of Transparency and Trust: Most current tools that detect hate speech don't explain their decisions. This lack of transparency can make users skeptical about the detections. Users might wonder why certain posts are flagged while others are not which can lead to mistrust and confusion. Example: If a social media platform deletes a user's post for being "hate speech" without any explanation, the user may feel unfairly targeted or censored. They might not understand what they did wrong.
Need for Contextual Understanding: Simply detecting words that might be hateful isn't enough.The context in which words are used is also crucial. There are Certain words which can be considered offensive but could be used in a positive way among friends. Example: The word "queer" has been a slur against LGBTQ+ individuals but has been reclaimed by many within this community as a term of empowerment. Any automated system without context would have marked "queer" as hate speech.
Common Sense Knowledge: By using ConceptNet, the system can understand not just the words, but the meanings and relationships behind them. This helps in accurately identifying and explaining why something is considered hate speech by taking into account nuances and contexts that might not be obvious. Example: If someone posts "Go back to where you came from," this might not seem explicitly hateful without context. However with common

sense knowledge the system can recognize this as a derogatory remark often aimed at immigrants.

3. Related Work

Decoder only transformer models such as GPT-2 and GPT-3 that have been pre-trained on a large amount of text data that can generate fluent. Encoder-decoder Transformers like T5 have shown massive improvements and success in many NLP tasks such as generation, summarization and translation. Recently, there are many attempts to use these generative models in solving non-generational tasks, *Wang et al in A multi-task instruction-based generative framework for few-shot ner* used used the T5 model for solving named entity recognition as a generative problem.

Krishanu Maity et al in StereoHate: Towards identifying Stereotypical Bias and Target group in Hate Speech Detection has done a hate speech classifiaction task on HSES dataset i.e speech corpus in hindi understanding hate speech patterns in low-resource languages. The paper proposes the model CGenEx, is a commonsense-aware unified generative framework, combines generation (stereotypical bias) and classification (target group category) tasks, achieving superior performance. CGenEx uses everyday knowledge to help explain why something might be considered toxic or hateful online.

4. Methodologies

In this work, we have proposed a commonsense-aware generative framework for generating stereotypical bias to explain why an input post is hateful. Detailed descriptions of the proposed models are described below.

4.1. Commonsense-aware Generative Framework

Output Preparation For each input sentence which we made after adding concept net with [SEP] in between X_i.

$$Y_i = \{\langle St \rangle\} \tag{1}$$

where St represents the stereotypical bias explanation of the input post X_i. Special characters can be used to denote any specific boundaries or features within the explanation.
Modeling The problem is formulated as a text generation task:

$$Y = G(X) \tag{2}$$

where G is a generative model, leveraging large pre-trained sequence-to-sequence models to generate the explanation Y from the input X.

4.1.1 Sequence to Sequence Learning (Seq2Seq):

The problem of text-to-text generation, defined in Equation 2, can effectively be addressed using a Sequence to Sequence model, which consists of two models: 1) BiLSTM with attention 2) T5(text-to-text transformer) model.

4.1.2 T5 Model

Describe the architecture and pre-training objectives of the T5 model, emphasizing its flexibility and strength in handling a variety of text-to-text transformation tasks.

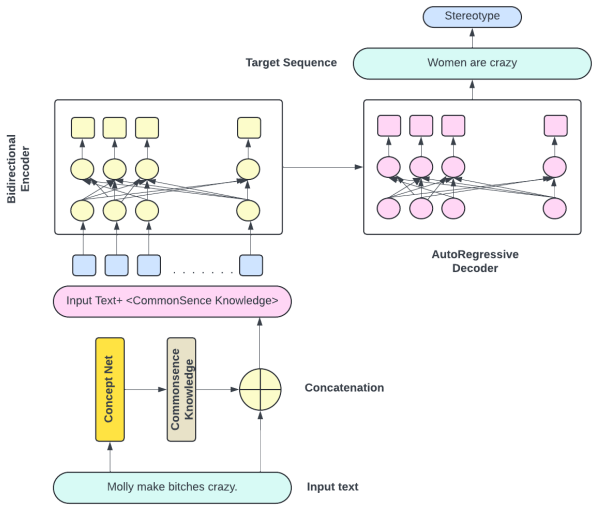

Figure 1. fine tuned T5 model with concept net

4.1.3 GPT-2 model

The model relies on a stack of transformer blocks of multi-headed attention and fully connected layers to encode the input tokens.Since GPT is a forward-only language model, the attention is only computed over preceding tokens.

4.1.4 BiLSTM with Attention Mechanisms

Details the workings and benefits of using BiLSTM networks, particularly focusing on how the attention mechanism enhances the capability of BiLSTM by focusing on important features in the input sequence.

4.2. Commonsense Extraction Module

We use ConceptNet as our knowledge base for the commonsense extraction module. At first, we feed the input text, Xi, to the Commonsense Extraction Module to extract the top 5 commonsense reasoning triplets using the strategy where a triplet consists of two entities and a connection/relation between these two entities which is then converted into a single sentence. Formally, to get the top 5 triplets from Conceptnet, we take the nouns, verbs, and adjectives from the input and search for related triplets in ConceptNet. Then, we sort them in order of the combination of

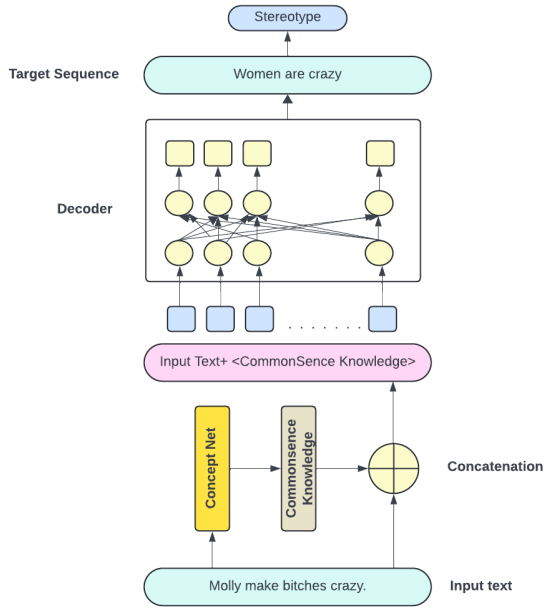

Figure 2. GPT-2

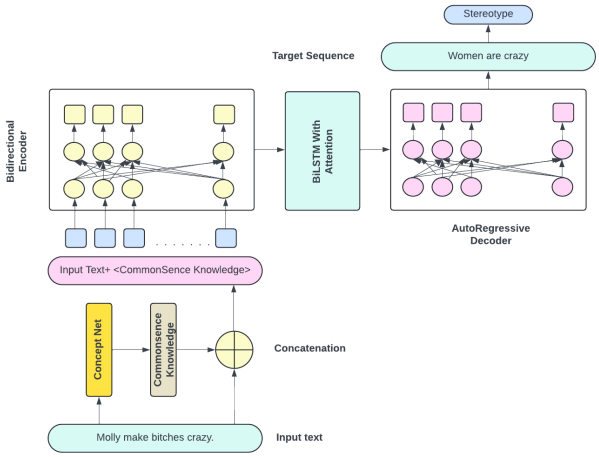

Figure 3. BiLSTM with Attention Mechanism

their IDF score and the edge weight of the triplets and then will select the top 5 triplets.

4.3. Inference

During the training process, we have access to both the input sentence (concept net words added) (Xi) and target sequence (Yi). Thus, we train the model using the teacher forcing approach, i.e. using the target sequence as the input instead of tokens predicted at prior time steps during the decoding process and for testing we just have to give Xi and it will give stereotype target.

5. Dataset

We are using Bias Frames dataset, available on Hugging Face for our research. The dataset contains over 150,000 structured annotations of social media posts.These annotations span approximately 34,000 implications related to a wide range of demographic groups.The curators included online posts from the reddit, r/darkJokes, r/meanJokes, r/offensiveJokes, Toxic language detection Twitter corpora ometime between 2014-2019. The dataset have predefined splits. The train splits have 112900 posts , validation splits have 16738 posts and test splits 17501 posts. The data fields are the same among all splits. We have used the following data fields in our model:

- whoTarget: a string, '0.0' if the target is a group, '1.0' if the target is an individual, and blank if the post is not offensive
- intentYN: a string indicating if the intent behind the statement was to offend. This is a categorical variable with four possible answers, '1.0' if yes, '0.66' if probably, '0.33' if probably not, and '0.0' if no.
- sexYN: a string indicating whether the post contains a sexual or lewd reference. This is a categorical variable with three possible answers, '1.0' if yes, '0.5' if maybe, '0.0' if no.
- sexReason: a string containing a free text explanation of what is sexual if indicated so, blank otherwise
- offensiveYN: a string indicating if the post could be offensive to anyone. This is a categorical variable with three possible answers, '1.0' if yes, '0.5' if maybe, '0.0' if no.
- sexPhrase: a string indicating which part of the post references something sexual, blank otherwise
- speakerMinorityYN: a string indicating whether the speaker was part of the same minority group that's being targeted. This is a categorical variable with three possible answers, '1.0' if yes, '0.5' if maybe, '0.0' if no.
- post: a string containing the text of the post that was annotated
- targetMinority: a string indicating the demographic group targeted
- targetCategory: a string indicating the high-level category of the demographic group(s) targeted
- targetStereotype: a string containing the implied statement

5.1. Data Preprocessing:

All punctuations marks were removed from texts to ensure raedibility and consistency. The text was tokenized into individual words, for easier manipulation and analysis of the dataset. Common stopwords such as 'is', 'and' , 'are' etc which often do not contribute much to the meaning of text were removed. User mentions, typically denoted by '@' symbols were filtered out. Top 5 common sense aware word are added using Comet. After the words were simplified, they were converted into final string which is used for training of the model.

6. Observation

6.1. Definitional Challenges:

7. Results and Analysis

We have used BLEU that is one of the earliest metrics to be used to measure the similarity between two phrases , ROUGE-L which measures the longest common subsequences between a pair of phrases , and BERTScore for comparing our models.

Table 1. Model Performance

Model	Acc	F1 Score	BLEU	Rouge-L
Fine-tune GPT-2 with ConceptNet	0.63	0.65	0.20	0.30
Fine-tune T5 small with ConceptNet	0.75	0.77	0.32	0.40
BiLSTM Attention model with ConceptNet	0.77	0.74	0.35	0.42

So the T5 small model and BiLSTM model are performing equally well.

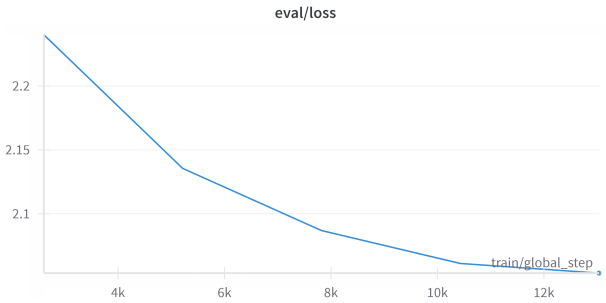

Figure 4. loss for T5 small

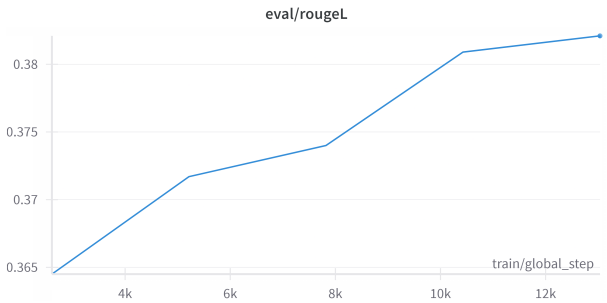

Figure 5. Rouge-L score for T5 small

8. Conclusion

In conclusion, this paper has presented a commonsense-aware generative framework for explaining hate speech using sequence to sequence models. By leveraging the power of the T5 model and the contextual understanding provided by ConceptNet, we can generate explanations that help in understanding the reason behind classifying a text as hate speech. This is a step towards creating more reliable and transparent AI-powered content moderation tools. .

8.1. Future Work

In our future work we will add two variations of commonsense aware encoder-decoder architecture (CGenEx-con and CGenEx-fuse) that are capable of incorporating Common Sense in their sequence-to-sequence learning process.We will also add reinforcement learning based training by using reward based objective function.

9. Contributions

We have all helped each other in various parts of the project, and the overall has been mainly a collective team effort.

Atishay Gupta: Data Collection, Exploratory Data Analysis, Data Preprocessing, Fine-tune GPT-2 with ConceptNet, Fine-tune T5 small with concept net, Report + Poster

Mahir Yadav: Data Collection, Exploratory Data Analysis, Data Preprocessing, Fine-tune GPT-2 with ConceptNet, Fine-tune T5 small with concept net, Report + Poster

Naman Sharma: Data Collection, Exploratory Data Analysis, Data Preprocessing, Fine-tune GPT-2 with ConceptNet, BiLSTM attention model with concept net, Report + Poster

Vianshu Shalyan: Data Collection, Exploratory Data Analysis, Data Preprocessing, Fine-tune GPT-2 with ConceptNet, BiLSTM attention model with concept net, Report + Poster

10. Refrences

1. https://arxiv.org/pdf/2203.03903
2. https://www.cse.iitb.ac.in/~pb/papers/nle23-stereohate.pdf
3. https://arxiv.org/pdf/1911.03891