Literature Survey

Problem Statement

Twitter aims to provide a safe space for all people irrespective of their caste, creed, country, gender, culture to put forth their views and opinions in a respectful manner without spewing hateful content and harming the sentiment of others.

The primary problem statement that I wished to tackle in this broad domain is to **gauge the impact of this hateful content on the user**. There are two primary aspects to the hateful content

a) The content of the hateful tweet itself
b) The intended individual/group at whom the tweet is targeted at and the impact on them.

It is the second aspect, the target of the hate speech that I wish to explore.

The crux of the Idea

Most of the literature focuses on the first aspect, and those that do focus on the target of hate speech[1][2] , approach it from linguistic and psycho-linguistic perspectives. They try to identify the linguistic features that differentiate directed hate speech (targeting a user) with generalized hate speech (targeting a community).

However, to understand another aspect of this **we can focus on the replies** to the particular tweet. This will give us a more nuanced understanding of the demographics of who is impacted, and to what degree by this hate tweet. For example, if a sexist tweet demeaning women is posted, the comments on the thread will tell the proportion of men who actively speak against it. Additional analysis of ethnicity, location, age, etc can also be done.

[1] "Hate Lingo: A Target-based Linguistic Analysis of Hate ...." https://arxiv.org/abs/1804.04257.
[2] "Analyzing the Targets of Hate in Online Social Media - AAAI." https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/download/13147/12829.

Relevant Research Works

The two most relevant pieces of research work , that covers similar domains are :-

Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media (ICWSM-2018)
https://arxiv.org/abs/1804.04257

The research paper focuses on the target of hate speech. It focuses on directed and generalized hate speech. The research mainly focuses on lexical analysis (comparing named entities , prevalent hashtags in directed hate speech as opposed to generalised), psycholinguistics analysis (scores for different LIWC categories) and semantic analysis (concepts evoked by different types of hate speech).

The research does focus on the target of the hate , but does so by analysing the content of the original post.

Analyzing the hate and counter speech accounts on Twitter
https://arxiv.org/abs/1812.02712

The research paper focuses on the replies of hate speech , with primary focus on account level analysis of hate speech and counterspeech accounts. (A tweet is 'counterspeech' if the tweet is a direct reply to a hateful tweet.) It focuses on the different means with which counter-speech accounts reply to hate speech (Humor,Pointing out hypocrisy,Hostile language etc).

Further creation dates , user activity , personality analysis were done to compare and contrast hate accounts and counterspeech accounts and they also proposed a classifier for prediction of counterspeech and hateful accounts.

Relevance and Novelty of Proposed Idea

The main idea is to identify the target of hate speech , quantitatively measure the impact on them based on the replies to the particular tweet. For a concrete example, let's take a misogynistic tweet. Based on the replies , we can gauge multiple things.

- Who is the targeted community (based on tweet content) and contrast with the people are speaking against it. (based on counter-replies)
- What is the tone used in the replies , which is a measure of how much it has impacted the user.
- Who are the bystanders (just liking a counter-reply as opposed to actually speaking against the argument).

There are some indicators apart from the tweet content which can give an inclination of the extent a particular tweet has impacted a person , things like number of replies , frequency of replies to other tweets versus this one , stance of the user regarding the particular topic etc.

Thus overall it seems a promising direction of research which has a good balance of previous literature to build upon as well as novel approaches to diversify into.