HASOC

Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL)

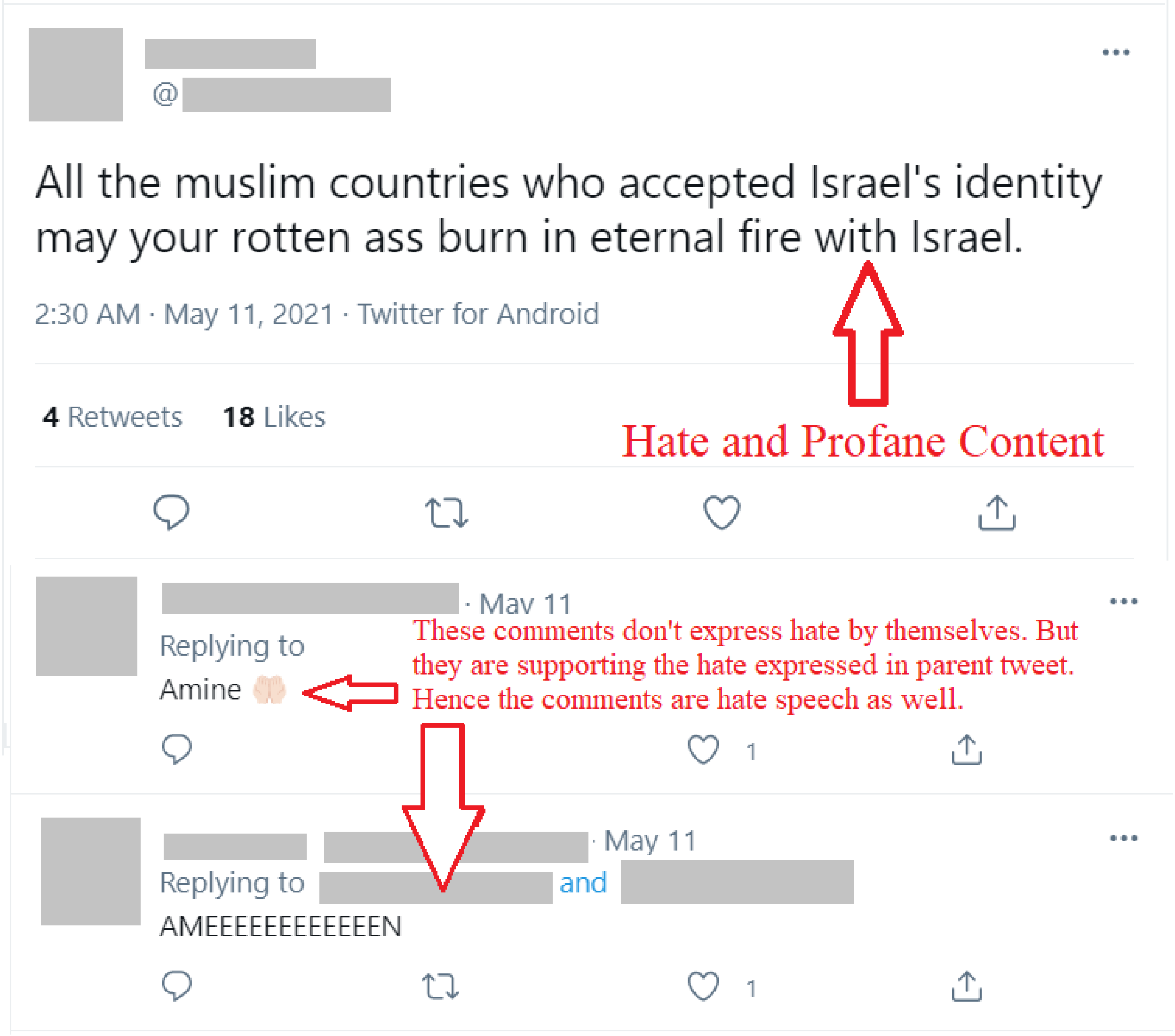
June 14, 2021

Overview

Social Media sites like Twitter, Facebook, being user-friendly and a free source, provide opportunities to people to air their voice. People, irrespective of the age group, use these sites to share every moment of their life making these sites flooded with data. Apart from these commendable features, these sites also have down sides as well. Due to the lack of restrictions set by these sites for its users to express their views as they like, anyone can make adverse and unrealistic comments in abusive language against anybody with an ulterior motive to tarnish one's image and status in society. A conversational thread can also contain hate and offensive content in it which is not apparent just from the single comment or the reply to comment but can be identified if given the context of the parent content. Furthermore, the contents on such social media are spread in so many different languages including code-mixed languages such as Hinglish. So it becomes a huge responsibility for these sites to identify such hate content before it disseminates to the masses.

Problem

The problem we're trying to address here is about identifying such conversational hate speech content that requires the context of the parent tweet in order to be justifiable.



[redacted]
@[redacted]

All the muslim countries who accepted Israel's identity may your rotten ass burn in eternal fire with Israel.

2:30 AM · May 11, 2021 · Twitter for Android

4 Retweets **18** Likes

Hate and Profane Content

[redacted] · May 11
Replying to
Amine 🤲

These comments don't express hate by themselves. But they are supporting the hate expressed in parent tweet. Hence the comments are hate speech as well.

1

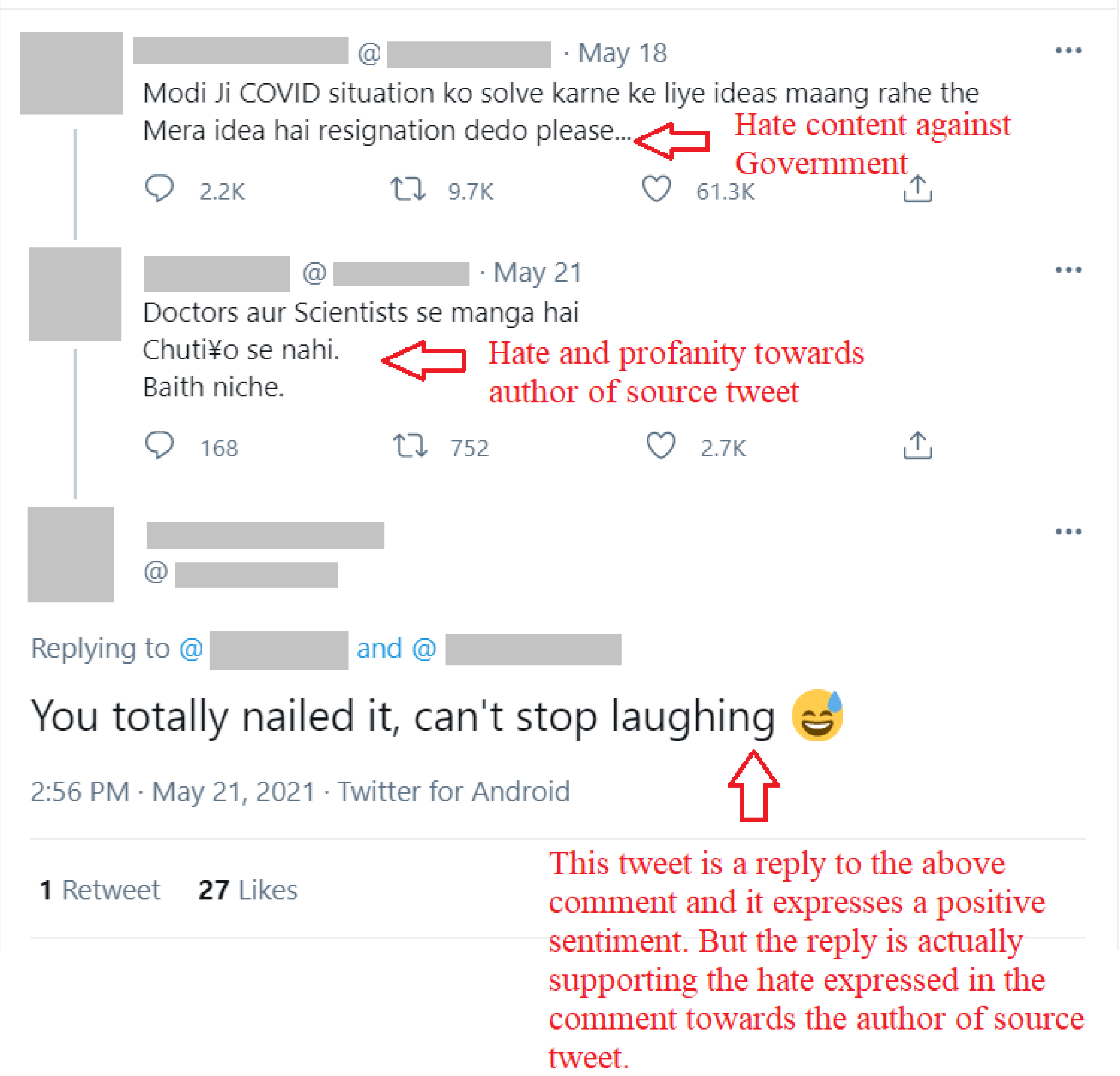
[redacted] [redacted] · May 11
Replying to [redacted] and [redacted]
AMEEEEEEEEEEEEN

1

The above screenshot from twitter describes the problem at hand effectively.

The parent tweet which was done at 2:30 am on May 11th is expressing hate and profanity towards muslim countries regarding the controversy happening in the israel at the time.

The 2 comments on the tweet have written “Amine” which means Trustworthy or Honest in arabic. If the 2 comments were to be analyzed for hate or offensive speech without the context of the parent tweet they wouldn’t be classified as hate or offensive content. But if we take the context of the conversation then we can say that the comments are supporting the hate expressed in the parent tweet. So those comments are hate speech as well.



[redacted] @[redacted] · May 18

Modi Ji COVID situation ko solve karne ke liye ideas maang rahe the
Mera idea hai resignation dedo please...

Hate content against Government

[redacted] @[redacted] · May 21

Doctors aur Scientists se manga hai
Chuti¥o se nahi.
Baith niche.

Hate and profanity towards author of source tweet

[redacted]
@[redacted]

Replying to @[redacted] and @[redacted]

You totally nailed it, can't stop laughing 😅

2:56 PM · May 21, 2021 · Twitter for Android

1 Retweet 27 Likes

This tweet is a reply to the above comment and it expresses a positive sentiment. But the reply is actually supporting the hate expressed in the comment towards the author of source tweet.

Another such example with code mixed text.

The Source Tweet: Modi Ji COVID situation ko solve karne ke liye ideas maang rahe the. Mera idea hai resignation dedo please… [Hate]

Translation : Modi ji (PM of India) was asking for ideas to solve the covid situation of India. My idea to him is to resign. [Hate]

The Comment: Doctors aur Scientists se manga hai. Chutiyo se nahi. Baith niche. [Hate]

Translation : They have asked Doctors and Scientists. Not fuckers. Sit down. [Hate]

The reply: You totally nailed it, can't stop laughing. [Hate]

The reply has a positive sentiment. But it is positive in favour of the hate expressed towards the author of the source tweet in the comment. Hence, it is supporting the hate expressed in the comment. Hence, it is also hate speech.

This is the type of problem we're aiming to solve via this shared task.

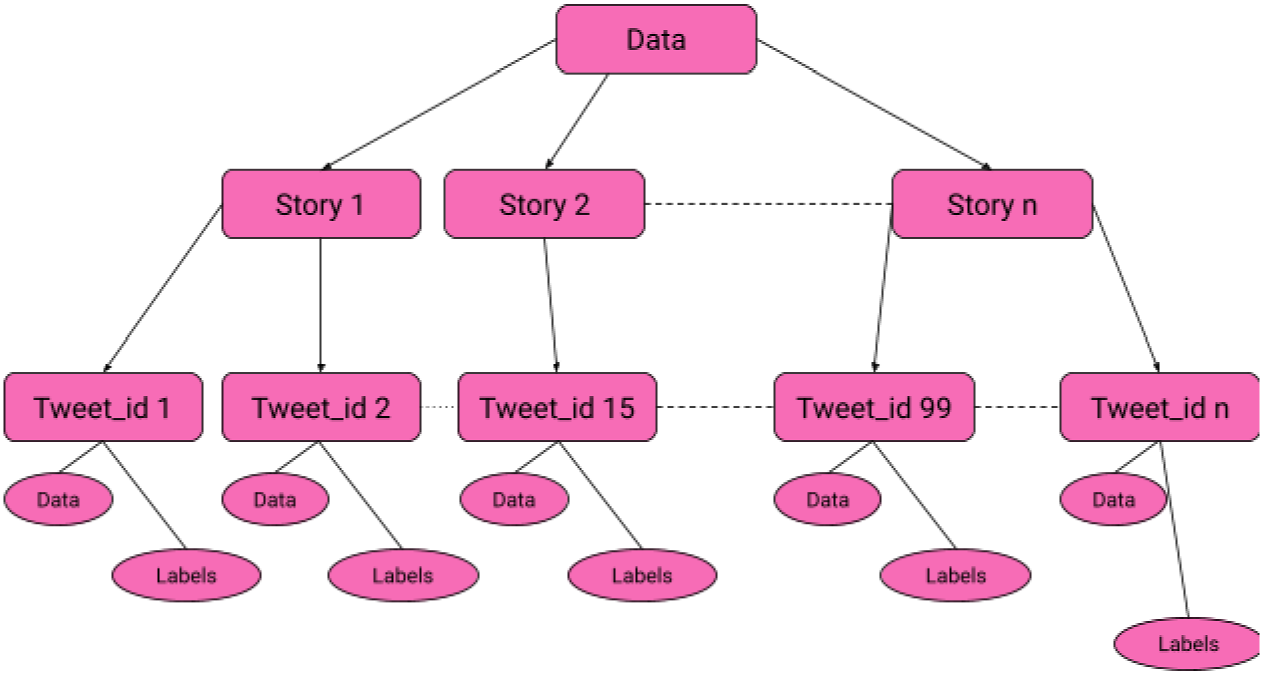
Data

The collection and annotation of data is a crucial task. As collecting tweets from just one or two twitter handles can lead the classifier to overfitting those two twitter users. Additionally, the problem we're encountering can not be accommodated to csv files as the structure of text data is not linear like traditional NLP problems.

To make the data diverse, we've hand picked a few controversial stories which can contain hate and offensive language. They are as following:

- Kunal Kamra: Kunal Kamra's satirical hate for Indian Government
- Cowdunk: Controversies regarding people in rural areas applying cowdunk on their bodies for immunity against corona. Some people hate such a mentality and some people hate news channels for spreading such news and hurting India's international image.
- BJP: Hate spreading on BJP tweets or tweets by population spreading hate for BJP
- Others: (still yet to be decided)

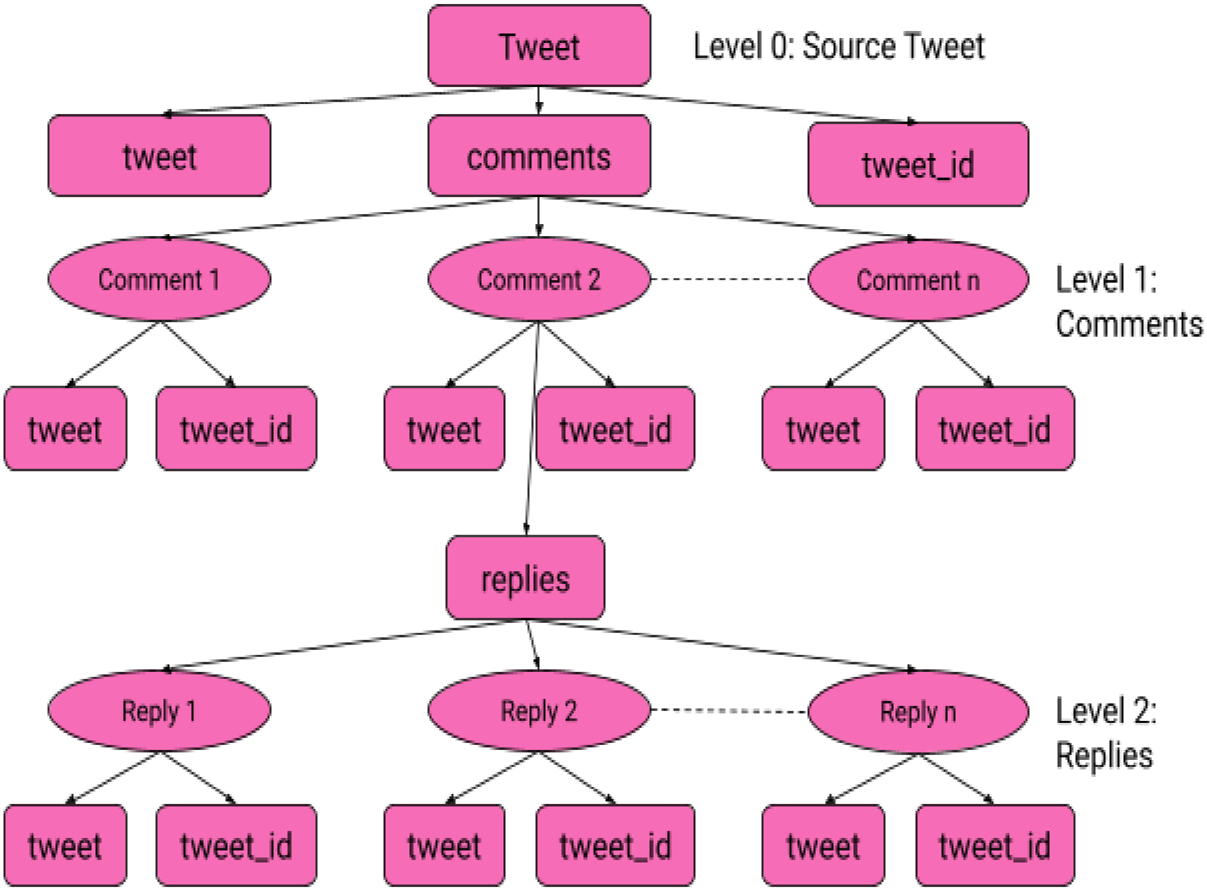
The directory structure of data is:



The rectangles represent directories and ovals represent files. All the files are in JSON format.

To tackle the dataset structure, we've used JSON as the data format for the dataset. JSON allows us to have a free structure that we can use to our needs.

The structure of data.json file is:



The rectangles are keys and ovals are elements of array represented by the parent key.

The contents of various keys are as follow:

- tweet: the text that is contained in the tweet
- tweet_id: a global tweet_id generated by twitter
- comments: array of comments that a tweet has
- replies: array of replies that a comment has

The structure of labels.json is linear. labels.json contains no nested data structure. It only contains key-value pairs where the key is the tweet id and value is the label for the tweet with the given tweet id.

There are 2 labels:

- HOF: Hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar. Degrading, dehumanizing or insulting an individual. Threatening with violent acts. Unacceptable language in the absence of insults and abuse. This typically concerns the usage of obscenity, swearwords and cursing.
- NONE: Harmless hateless content

The train set will be ready by 30th June. (Tentative)

Baseline Model

We understand that FIRE hosts so many beginner friendly workshops every year and this problem might not seem like beginner friendly. So, we've decided to provide participants with a baseline model which will provide participants with a template for steps like importing data, preprocessing, featuring and classification. And the participants can make changes in the code and experiment with various settings.

The code for baseline model is available at : github link for repo

Register

You can register on the following link: link to google form

If you are interested in participating but have any doubts you can email us with your queries at:

mandl@uni-hildesheim.de - Prof. Thomas Mandl
sjmodha@gmail.com - Prof. Sandip Modha

hirenmadhu16@gmail.com - Hiren Madhu

shreysatapara@gmail.com - Shrey Satapara