HASOC 2021
Hate Speech and offensive Content Identification in Multiple Languages

(../index.html)

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

Overview

Social media sites like Twitter and Facebook, being user-friendly and a free source, provide opportunities for people to air their voices. People, irrespective of age group, use these sites to share every moment of their lives, making these sites flooded with data. Apart from these commendable features of social media, they also have downsides as well. Due to the lack of restrictions set by these sites for their 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, which is not apparent just from a single comment or the reply to a 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

A conversational thread can also contain hate, offensive, and profane content, which is not apparent from a standalone or single tweet or comment or the reply to a comment, but can be identified if given the context of the parent content.

The above screenshot from Twitter describes the problem at hand effectively. The parent/source tweet, which was posted at 2:30 am on May 11th, expresses hate and profanity towards Muslim countries regarding the controversy happening during the recent Israel-Palestine conflict. 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 support the hate/profanity expressed in the parent tweet. So those comments are labelled as hate/offensive/profane.

This sub-task focused on the binary classification of such conversational tweets with tree-structured data into:

- (NOT) Non Hate-Offensive - This tweet, comment, or reply does not contain any Hate speech, profane, offensive content.
- (HOF) Hate and Offensive - This tweet, comment, or reply contains Hate, offensive, and profane content in itself or supports hate expressed in the parent 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…
- **Translation :** Modi ji (PM of India) was asking for ideas to solve the covid situation of India. My idea to him is to resign.
- **The Comment:** Doctors aur Scientists se manga hai. Chutiyo se nahi. Baith niche. [HOF]
- **Translation:** They have asked Doctors and Scientists. Not fuckers. Sit down. [HOF]
- **The reply:** You totally nailed it, can't stop laughing. [HOF]

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.

ICHCL Dataset

The sampling and annotation of social media conversation threads is very challenging. We have chosen controversial stories on diverse topics to minimize the effect of bias. We've hand picked controversial stories from the following topics that have a high probability of containing hate, offensive, and profane posts.

The controversial stories are as follow:

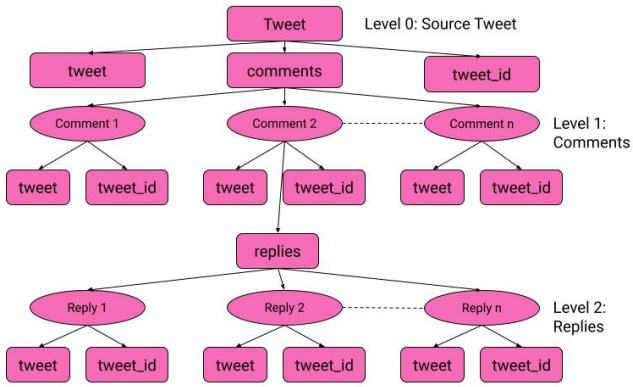
- Twitter Conflicts with the Indian Government on new IT rules.
- Casteism controversy in India
- Charlie Hebdo posts on Hinduism
- The Covid-19 crisis in India 2021
- Indian Politics

- The Israel-Palestine conflict in 2021
- Religious controversies in India
- The Wuhan virus controversy

The directory structure of data directory :

```
.
└── Data/
    ├── Story1/
    │   ├── Tweet_id1/
    │   │   ├── data.json
    │   │   └── labels.json
    │   ├── Tweet_id2/
    │   │   ├── data.json
    │   │   └── labels.json
    │   ├── .
    │   ├── .
    │   └── Tweet_id-N/
    │       ├── data.json
    │       └── labels.json
    ├── Story2/
    │   ├── Tweet_id1/
    │   │   ├── data.json
    │   │   └── labels.json
    │   ├── Tweet_id2/
    │   │   ├── data.json
    │   │   └── labels.json
    │   ├── .
    │   ├── .
    │   └── Tweet_id-N/
    │       ├── data.json
    │       └── labels.json
    ├── .
    ├── .
    └── .
```

The structure of data.json :
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.

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, click here (https://github.com/hasocfire/hasocfire.github.io/tree/master/hasoc/2021/ichcl/baseline)

Note: baseline model is just to give you a basic idea of our dir. structure and how one can classify context based data, there are no restrictions on any kind of experiments

Acknowledgment

We are a thankful anonymous reviewer of Expert System and application who inspired us to formulate this problem during the reviewing process of our paper Modha et al. 2021.

Reference

- Modha, Sandip; Majumder, Prasenjit; Mandl, Thomas; Mandalia, Chintak (2020): Detecting and Visualizing Hate Speech in Social Media: A Cyber Watchdog for Surveillance. Expert Systems With Applications (ESWA) https://doi.org/10.1016/j.eswa.2020.113725 (https://doi.org/10.1016/j.eswa.2020.113725)
- Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE '19). Association for Computing Machinery, New York, NY, USA, 14–17. https://doi.org/10.1145/3368567.3368584 (https://doi.org/10.1145/3368567.3368584)

Contact us

- hirenmadhu16@gmail.com - Hiren Madhu :- IISC Banglore, Banglore, India
- shreysatapara@gmail.com - Shrey Satapara :- LDRP-ITR, Gandhinagar, India