Aggression-annotated Corpus of Hindi-English Code-mixed Data

In this academic article, which focuses on distinguishing between approved and unapproved aggressive behaviors in order to make monitoring and intervention methods more effective, it is mentioned that the classification of aggressive behaviors is theoretically weak and the behaviors often overlap with each other.

Everything in the article is explained in an understandable way. However, the only thing that is not understood is that non-Hindi and non-English posts are labeled as "Non-Aggressive". I would find it more logical to remove these shares from the data set because I could not understand the reasons for their inclusion in the data set. (*If the tweet/comment was in a language other than English or Hindi (or something that the annotator did not understand), it was to be marked as non-aggressive*).

Verbal aggression is divided into 2 classes overt and covert according to its expression, and 4 classes and subclasses (Identity Threat has 6 subclasses) according to its target. While classifying aggression, the corpus is made according to the level of aggression and information about the types of aggression it exhibits. The tag set contains 3 levels of Aggression, and each of these levels includes 2 sub-tags, discursive role, and discursive effects. Discursive roles divide a person's attitude towards comments involved in any sharing and interaction into 3 classes: Offensive, Defense, and Provocation.

Anchors are allowed to flag more than one Discursive Effect as needed. If the comment or Tweet contains harassment/abuse, it is required to be flagged and at least 2 Discursive Impacts to be flagged in addition. Finally, if the host marks the post as exhibiting General Non-Treaty Aggression, that post cannot be flagged for another Discursive Impact.

Although it is not a very effective research question, it can be addressed because it is a subject that is mentioned in the article. "How does the performance of detecting aggression vary with the number of discursive effects annotated?".

As a result of the study, it was seen that the hate language in two different social media applications differed in expressing aggression. For example, posts on Twitter are more subtle and private, while posts on Facebook show a more aggressive profile.

The similarity and complexity in the classification of the attack were one of the points that caught my attention. Because asking the annotators to classify this situation that we constantly encounter on social media, and the fact that these attacks are labeled differently for each, reveals how such attacks change from person to person and the difficulty of detecting them.

Predicting the Type and Target of Offensive Posts in Social Media

This academic paper addresses the key similarities and differences between OLID and pre-existing datasets for hate speech identification, aggression detection, and similar tasks, and compares the performance of different machine learning models on this dataset. In the academic paper, three levels of hierarchy are used to distinguish whether the language in the dataset is offensive, its type, and its target. To evaluate the proposed set of tags and to ensure data retrieval method and quality, a trial description was made with experts using nine keywords. As a data source, extreme left and right news accounts were used for containing political comments and insults. When examining these comments, two annotations were taken for each sample. In case of disagreement, a third review was requested and the final decision was taken by majority vote. Announcers tagged tweets by level.

The research question could be "How does the performance of detecting the target of aggression differ with different machine learning models?". The results obtained in the academic paper are satisfactory and future studies aim to make a cross-sentence comparison of annotated datasets and OLID for similar tasks such as aggression identification and hate speech detection. All three models yielded similar results far exceeding random baselines, with a slight performance advantage for the neural models. Although there is no previous study investigating the target of offensive language, examining the success of detecting the target of hate speech on models was the most interesting point of the academic paper.

Hate Speech Dataset from a White Supremacy Forum

This article describes a hate speech dataset of thousands of sentences manually labeled as containing hate speech, addressing the subjectivity and difficulty inherent in labeling hate speech following strict guidelines. In order to have the same understanding of hate speech, the annotators created and discussed the guidelines together, and 4 labels were created. In order for a sentence to be categorized as a HATE tag, three premises must be true, such as being an intentional attack and targeting a specific group of people. The

NOHATE tag is used to classify sentences that do not contain hate speech. The RELATION tag is meant that sentences in a post do not contain hate speech on their own, but contain a combination of several sentences. The SKIP tag is for sentences that are not written in English or contain no information that would be classified as HATE or NOHATE. Also, by using a web-based tool developed by the authors and viewing all the sentences from the same post at the same time, the reviewers were able to better understand the intent of the post's author to the commenter.

Research question "How does the hate speech detection performance change with a custom disclosure tool developed to perform the manual tagging task?" could be. As a result, future studies are needed to obtain stronger hate speech automatic classifiers, emphasizing the importance of vocabulary and the whole conversation. In general, the LSTM-based classifier achieves better results, but even simple SVM using word bag vectors can distinguish classes reasonably well.

I found this academic paper much more detailed and understandable than the other two papers. Everything was detailed, from the explanation of the differences between the Relevant Studies section and the current study described in the paper to the statistical information used. One of the interesting points made in the academic paper is that Hatebase proved by Saleem et al. that it is successful in identifying keyword-based hate speech, but that hateful phrases cannot be distinguished from clean ones when shared with different intentions. Another interesting point I haven't heard of is Perspective7, a tool developed by Google and Jigsaw that measures the "toxicity" of comments. Also, the Related Studies section was a section that really caught my attention as it showed me how this technology is developing, so I had access to a lot of resources to review.