Subjectivity Mining - Assignment 1

Automatic hate speech detection

Payanshi Jain - pja204

September 8, 2023

Paper A

Question 1: Do you have clarifying questions? Did you understand everything?

Theoretically, the paper is well-crafted, and the annotation scheme is quite comprehensible to me. However, when viewed from a practical standpoint, distinguishing between Overtly Aggressive and Covertly Aggressive seems somewhat subjective, potentially leading to variations among annotators. Additionally, the inclusion of "counter attack" within the Defend category feels a bit misplaced; it have a potential to be a separate category.

Question 2: Give a short overview of the annotation guidelines presented in the paper

The paper outline a scheme for annotating the corpus with information regarding aggression levels and the specific types of aggression exhibited. The tagset comprises three top-level tags: Overtly Aggressive (OAG), Covertly Aggressive (CAG), and Non Aggressive (NAG). Each aggressive level (OAG and CAG) contains two attributes: Discursive Role and Discursive Effect. Discursive Effects are based on a typology of aggression and encompass ten categories, which encompass all sub-types of aggression and abuse. Discursive Roles define three roles a person might play in an aggressive discourse: Attack, Defend, and Abet.

Furthermore, annotation is performed at the document level, covering complete posts, comments, or any unit of the discourse. Annotators were given specific instructions, including allowing the marking of multiple discursive effects if a comment exhibits more than one type of aggression. Additionally, if a comment contains any form of abuse, it must be marked as such, along with at least one more effect. Comments marked as exhibiting General Non-threatening Aggression cannot be marked for any other effect. Finally, if a tweet or comment is in a language other than English or Hindi (or is not understood by the annotator), it should be marked as non-aggressive.

Question 3: What is the motivation for the paper?

The motivation for the paper is the increasing incidents of aggression and related behaviors such as trolling, cyberbullying, flaming, and hate speech on the internet. The paper highlights that while these behaviors existed before the internet, the internet's wide reach and influence have given them unprecedented power to affect the lives of billions of people. Therefore, the motivation is to take preventive measures to safeguard people using the web and ensure that the internet remains a viable medium of communication and connection.

Question 4: What is the research question?

The paper discusses the development of an aggression tagset and an annotated corpus of Hindi-English code-mixed data from social media platforms like Twitter and Facebook. The primary focus appears to be on the creation of this dataset to further research in the field of aggression analysis.

Question 5: What did they find?

The final annotated dataset reveals a significant distinction in communication styles between Facebook and Twitter. Although the length of Facebook comments is mostly less than 150 characters, similar to Twitter's character limit of 140, a comparison of aggression levels on both platforms indicates a notable difference. Facebook users tend to be more overtly aggressive, while Twitter users display a more subtle and covert approach to expressing aggression. Notably, a significant portion of both tweets and Facebook comments in the dataset revolve around political aggression. Additionally, the data highlights an intriguing interaction between code-mixing and aggression. It shows that a majority of code-mixed comments and tweets are aggressive, whereas for posts in Hindi, aggression is

evenly distributed, and for posts in English, it tends to be largely non-aggressive.
Question 6: What is their conclusion?

The paper introduces a novel annotation scheme for categorizing aggression levels, a first of its kind. The annotated dataset is a valuable resource for studying and automatically detecting aggression, trolling, and cyberbullying on social media. Despite initial attempts at automated identification, achieving high accuracy remains a challenge, indicating the intricacy of aggression classification.
Question 7: What are -according to you- interesting aspects of the paper?

In my opinion, what stands out most in this paper is the adept handling of code-mix comments, coupled with the multi-level tagging system that enables a nuanced understanding of genuinely aggressive comments.

Paper B

Question 1: Do you have clarifying questions? Did you understand everything? I comprehend the authors' intent in the paper. However, despite the three hierarchical levels, the categories still seem quite broad. Although they make an effort to identify targets as Individual, Group, and Other in Level C, the definition of "Other" appears somewhat vague.
Question 2: Give a short overview of the annotation guidelines presented in the paper

The annotation guidelines for the OLID dataset employ a hierarchical schema with three distinct levels. Level A focuses on offensive language detection, distinguishing between "Not Offensive" posts that lack profanity or offense, and "Offensive" posts containing unacceptable language, veiled or direct targeted offenses, insults, threats, or profanity. Level B categorizes offensive language into "Targeted Insult" posts with explicit insults/threats towards individuals or groups, and "Untargeted" posts featuring non-specific profanity. Level C identifies the targets of insults/threats, classifying them as "Individual" (targeting specific individuals), "Group" (targeting collective entities based on shared characteristics), or "Other" (targets not fitting the previous categories).
Question 3: What is the motivation for the paper?

The motivation for this paper stems from the pervasive presence of offensive content on social media platforms. Previous research focused on detecting specific types of offensive content, such as hate speech, cyberbullying, or cyber-aggression. However, this paper aims to address the issue comprehensively by targeting various forms of offensive content.
Question 4: What is the research question?

The primary research question addressed in this paper revolves around the comprehensive identification of offensive content on social media. Specifically, the paper focuses on modeling the task hierarchically, aiming to not only detect offensive content but also distinguish between different types and targets of offensive messages. Additionally, the paper explores the similarities and differences between the newly compiled OLID dataset and pre-existing datasets used for hate speech identification, aggression detection, and similar tasks. The authors also conduct experiments to compare the performance of different machine learning models on the OLID dataset.
Question 5: What did they find?

The experiments conducted with various machine learning models yielded significant insights. The linear SVM, known for its effectiveness in text classification tasks, demonstrated competent performance in distinguishing between offensive and non-offensive posts. The bidirectional Long Short-Term-Memory (BiLSTM) model, adapted from sentiment analysis, outperformed the SVM, achieving a macro-F1 score of 0.80. The Convolutional Neural Network (CNN) model, based on a well-established architecture, excelled in discriminating between targeted insults/threats and untargeted offenses, outperforming the BiLSTM. All three models exhibited commendable results in offensive target identification, with a slight advantage for the neural models. Notably, the challenge of classifying the "Others" category was evident due to its heterogeneous nature and limited training instances, resulting in lower performance for this class. Overall, the models demonstrated promising capabilities in accurately detecting and categorizing offensive language in social media posts.
Question 6: What is their conclusion?

The OLID dataset, particularly in the OffensEval context, treats each level of annotation as an independent sub-task. As far as our knowledge extends, this dataset stands as the initial instance to incorporate annotation specifying both the type and target of offenses in social media, offering intriguing avenues for future research. Our baseline experiments, utilizing both SVMs and neural

networks, have demonstrated the challenging yet achievable nature of this task. Looking ahead, we aim to conduct a comparative analysis between OLID and datasets annotated for related tasks like aggression identification and hate speech detection. Additionally, we intend to extend this hierarchical annotation approach to create similar datasets for other languages.

Question 7: What are -according to you- interesting aspects of the paper?

In my opinion, the most compelling aspect of this paper lies in its adept handling of hierarchical tagging, particularly in Level B, which involves categorizing offensive language into Targeted and Untargeted insults, and Level C, where the identification of Individual, Group, or Other targets is conducted.

Paper C

Question 1: Do you have clarifying questions? Did you understand everything?

The paper's annotation scheme is clear, yet it leaves me wondering about its ability to capture the intensity or severity of hate speech, as well as whether it targets individuals or vulnerable groups. While machine learning models may yield positive outcomes, I believe the annotation lacks granularity in this regard.

Question 2: Give a short overview of the annotation guidelines presented in the paper

The annotation guidelines were meticulously crafted through the collaborative efforts of all annotators to ensure a shared understanding of hate speech. The guidelines categorize sentences into four types: HATE, NOHATE, RELATION, and SKIP. HATE includes sentences containing deliberate attacks directed at a specific group based on identity. NOHATE encompasses sentences devoid of hate speech. RELATION applies when multiple sentences together convey hate speech, while SKIP is for non-English sentences or those lacking classifiable content. To maintain consistency, annotators engaged in extensive discussions and modified guidelines based on initial annotations. Inter-annotator agreement was calculated to ensure reliability. Context played a crucial role, with annotators often seeking additional information to make informed decisions. The resulting dataset, although unbalanced, provides a robust foundation for hate speech analysis.

Question 3: What is the motivation for the paper?

The motivation behind this paper stems from the increasing prevalence of hate speech in online user-generated content, particularly on social media platforms. The authors recognize the significance of addressing this issue due to its escalating impact on society. They aim to contribute to the field by presenting a meticulously annotated dataset focused on hate speech, derived from the supremacist forum, known for white supremacist discussions.

Question 4: What is the research question?

The primary research question addressed in this paper revolves around the identification and analysis of hate speech within online content. Specifically, the authors aim to investigate the characteristics and prevalence of hate speech in sentences sourced from the supremacist forum. They also aim to evaluate the effectiveness of various classification models in identifying hate speech. Additionally, the study assesses the potential impact of contextual information on the labeling process.

Question 5: What did they find?

The introduced dataset and the broader task of hate speech annotation present several noteworthy considerations. The source of the content, originating from the supremacist forum, inherently contains elements of racism and hate. However, discerning what constitutes hate speech amidst racist expressions is a nuanced and subjective matter touching on themes of free speech and civility. The paper meticulously outlines the criteria for hate speech annotation, acknowledging the complexity of the process. The annotation guidelines underwent multiple iterations to address inconsistencies among human annotators. Yet, certain criteria, such as defining a deliberate attack, remain open to interpretation and warrant further clarification. Additionally, the choice to label at the sentence level, as opposed to full comments, introduces a level of granularity that prompts discussion. The inclusion of the RELATION label for interdependent sentences is a notable feature, although its utilization has been infrequent. Finally, the necessity of additional context for accurate labeling, as highlighted in the error analysis, emphasizes the importance of understanding context dependency in hate speech detection. As annotators gain experience, they tend to rely less on contextual cues for accurate annotations

Question 6: What is their conclusion?

This paper introduces a meticulously labeled hate speech dataset sourced from the supremacist

forum, comprising around 10,000 sentences categorized as hate speech or not. The annotation criteria are elaborated due to the nuanced nature of hate speech. Various aspects of the dataset, including the need for contextual cues, and vocabulary distribution in hate speech examples, are examined. Baseline experiments with automatic classifiers, especially on challenging cases, are conducted. The dataset is publicly accessible, offering a foundation for further research. Future work could explore integrating world knowledge and context, studying the impact of RELATION-labeled sentences, and delving deeper into dataset characteristics like timelines, user behavior, and targets of hate speech. Addressing class imbalance through advanced labeling methods like active learning is also suggested.
Question 7: What are -according to you- interesting aspects of the paper?

In my view, the most intriguing aspect of this paper lies in the inclusion of an extra annotation called "Relation" alongside Hate/NoHate/Skip. This addition acknowledges the nuanced spectrum of emotions, bridging the gap between hate and non-hate expressions. It also prompts thoughtful reflection on Type 1 and Type 2 errors.

References

[1] Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, Tushar Maheshwari(2018) Aggression-annotated Corpus of Hindi-English Code-mixed Data. In: Proceedings of LREC-2018, Miyazaki, Japan.

[2] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal Noura Farra, Ritesh Kumar (2019) Predicting the Type and Target of Offensive Posts in Social Media.

[3] O. de Gibert, N. P erez, A. Garc ıa-Pablos, M. Cuadros, 2018. Hate Speech Dataset from a White Supremacy Forum. In ALW2: 2nd Workshop on Abusive Language Online.