Subjectivity Mining: Assignment 1

Individual

Noa van Mervennée (2710633)

Artificial Intelligence, Vrije Universiteit, 1081 HV Amsterdam
n.van.mervennee@student.vu.nl

1 Paper A: Kumar et al.

The first paper analyzed is called *Aggression-annotated Corpus of Hindi-English Code-mixed Data.*

Give a short overview of the annotation guidelines presented in the paper The tagset contains 3 tags at the top-level (overtly/covertly/non-aggressive) and each of the the two aggressive levels contains 2 attributes – discursive role and discursive effects The discursive roles are divided into attack, defend and abet. The discursive effects are the following: physicial threat, sexual / gendered / racial / communal / casteist / political / geographical / general non-threatening aggression, curse/abuse. Approximately 500 test instances are annotated at document-level (comment, complete post .. etc.) by four annotators in the first round. In the second round of agreement experiments 1100 test instances were annotated by three annotators.

What is the motivation for the paper? Due to the increasing growth in the availability and use of the internet, hate speech is growing along with it. As a consequence; aggression, criminal-activities and unratified verbal behaviour is expanding. The motivation of the paper is to recognize, and subsequently deal with, the aggressive behavior automatically or semi-automatically.

What is the research question? How can an intelligent system learn to automatically or semi-automatically distinguish incidents of both overt as well as covert aggression and at the same time gain an understanding of the difference between ratified and unratified aggressive behavior?

What did they find? For the top-level annotation, the inter-annotator agreement score was 0.49. Certain changes had to be made to the annotation guidelines to increase the score and therefore they conducted a second round of agreement experiments. This resulted in the inter-annotator agreement score for the top-level at slightly above 0.72. The agreement for the 10-class annotation of discursive effect was approximately 0.57. With these results they continued the experiment. Moreover they found out that approximately 2/3 of the Facebook comments are of less than 150 characters (which is approximately equal to Twitter's restriction of 140 characters). In addition, people are more vocal and overtly aggressive on Facebook in comparison to Twitter where people are more subtle and covert in expressing aggression [2]. Another observation is that it seems that majority of the tweets and facebook messages concern political aggression. At last, the data shows that a majority of code-mixed comments and tweets are aggressive, while for posts in Hindi, it is equally distributed and for posts in English, it is largely non-aggressive [2].

What is their conclusion? They conclude that, as far as they know, it is the first dataset to be annotated with different levels and kinds of aggression. In addition they state that the dataset could be prove to be an invaluable resource for understanding as well as automatically identifying aggression and other related phenomenon like trolling and cyberbullying over the web, especially social media platforms [2].

What are -according to you- interesting aspects of the paper? It is interesting to see the combination of various aggression levels and discursive roles/effects rather than focusing on just one subtask. Moreover, it is interesting to see the high increase of the inter-agreement score after changing the guidelines.

Do you have clarifying questions? Did you understand everything? I doubt the relevance of the comparison between the posts in Hindi versus the posts in English due to the differences in context of the posts and difference in culture. Could it not be the case that an hate-full tweet in Hindi is considered as non hate-full in English due to their cultural differences? Moreover, they changed the guidelines after a low inter-agreement score. Would it not be the case that you are steering too much in one direction with so few annotators? As a consequence, several posts that are considered as non hate-full might be considered as hate-full by many others.

2 Paper B: Zampieri et al.

The second paper analyzed is called *Predicting the Type and Target of Offensive Posts in Social Media.*

Give a short overview of the annotation guidelines presented in the paper The paper describes to use a hierarchical annotation schema split into three levels: offensive / not-offensive (A), the type (targeted / untargeted insult) (B), the targets (individual / group / other) (C). By using keywords that are often included in offensive messages, they first annotated 300 instances with six experts using nine keywords. The goal of the trial annotation was (i) to evaluate the proposed tagset, (ii) to evaluate the data retrieval method, and (iii) to create a gold standard with instances that could be used as test questions to ensure the quality of the annotators for the rest of the data, which was carried out using crowdsourcing [3]. During the full annotation task they included more political keywords as they contain more offensive content. They annotated the data using crowdsourcing. The paper describes to ensure data quality by (i) only hiring annotators who were experienced in the platform, and (ii) using test questions to discard annotations by individuals who did not reach a certain threshold [3]. Each instance in the dataset was annotated by multiple annotators and inter-annotator agreement was calculated at the end [3]. They acquired two annotators for each instance, and in the case of disagreement, a third annotator is requested with the majority vote chosen.

> We first acquired two annotations for each in- stance. In the case of disagreement, we requested a third annotation, and we then took a majority vote.

What is the motivation for the paper? Previous work on the topic of identifying potentially offensive messages did not consider the problem as a whole, but focused on detecting very specific types of offensive content. This paper aimed to model the task to target different kinds of offensive content hierarchically by identifying the type and the target. Therefore, the motivation of the paper is to combine the type and target to improve previous studies that propose techniques that (semi)-automatically identifies offensive content.

What is the research question? How will the Offensive Language Identification Dataset (OLID), that makes use of a hierarchical annotation schema, improve the identification and characterization of offensive language?

What did they find? The inter-annotator score of five annotators on 21 tweets was 0.83 for the first layer (offensive or not). Moreover, political keywords they tend to be richer in offensive content. For the full-dataset, approximately 60% of the time, the two annotators agreed. Furthermore, one of the main challenges they state, was to find enough data for each class. Moreover, all three models perform significantly better than chance by identifying level A, with the neural models performing substantially better than the SVM [3]. For level B, all models perform better at identifying targeted insult compared to untargeted insult. At last, for level C, all three models achieved similar results, far surpassing the random baselines. The performance of all models for the other-class is 0, which can be explained the minority of training-instances and by the fact that the model has to learn many varieties and thus has difficulty identifying it.

What is their conclusion? They mention that, to the best of their knowledge, the OLID is the first dataset to contain annotation of type and target of offenses in social media, and it opens interesting research directions.

What are -according to you- interesting aspects of the paper? The guidelines for annotations are very detailed, giving us a good idea of the reproducibility of this experiment. In addition, the splits of the targets are quite interesting to include any type of offensive messages, whereas other papers focus on specific types. However, many other types of abusive messages have a large overlap with each other.

Do you have clarifying questions? Did you understand everything? The paper states to have used *six experts* for the first trial of annotation. Questions arise, why are these annotators experts? Are they diverse enough and understand offensive language for various cultures? Moreover, the inter-agreement score for five annotators in the first trial of 21 tweets 0.83. Isn't the number of tweets too small to already draw conclusions about a high-agreement?

3 Paper C: de Gibert et al.

[1] The second paper analyzed is called *Predicting the Type and Target of Offensive Posts in Social Media.*

Give a short overview of the annotation guidelines presented in the paper 10.568 sentences have have been extracted from Stormfront and annotated in batches of 10 to control the process. The annotation procedure and guidelines were progressively refined and adapted for the first two batches. They classified the data as conveying hate speech or not and into two other auxiliary classes: relation (one sentence does not contain hate speech on its own, but the combination of several sentences does) and skip (sentences that are not written in English or that do not contain information as to be classified into hate or no-hate). To make it possible re-build the conversations these sentences belong to, a post identifier and the sentence's position in the post, a user identifier, a sub-forum identifier is included. Furthermore, the number of previous posts the annotator had to read before making a decision over the category of the sentence is also given [1].

What is the motivation for the paper? The increase of cyberbullying and cyberterrorism, and the use of hate on the Internet, make the identification of hate in the web an essential ingredient for anti-bullying policies of social media This paper releases a new dataset of hate speech to further investigate the problem of the increase of cyberbullying and cyberterrorism, as a consequence to identify hate on the web and limit the consequences of hatefull content. They present the first public dataset of hate speech annotated on Internet forum posts in English at sentence-level . The annotation work was carried out using a tool developed by the authors.It displays all sentences belonging to the same post for a better understanding of the post.

What is the research question? How well can Hate Speech be semi-automatically classified with the use of detailed annotation guidelines from which a dataset is composed on a sentence level? This paper describes a hate speech dataset com- posed of thousands of sentences manually la- belled as containing hate speech or not.

What did they find? The inter-agreement scores on the batches 1 and 2 have an average of 91.03% and 90.97%. Furthermore, context appears to be of great importance when annotating hate speech. The sub-forums that contain more hate belong to the category of news, discussion of views, politics, philosophy, as well as to specific countries (i.e., Ireland, Britain, and Canada) [1], whereas the sub-forums that contain more no-hate sentences are about education and homeschooling, gatherings, and youth issues.. Moreover, by contructering a hate score (HS) they fonud out that the most hateful words are derogatory or refer to targeted groups of hate speech. On the other hand, the least hateful words belong to the semantic fields of Internet, or temporal expressions, among others. This shows that the vocabulary is discernible by category, which in turn suggests that the annotation and guidelines are sound [1]. In addition, the dataset has been contrasted against the English vocabulary in Hatebase. 9.28% of hate-vocabulary overlaps with Hatebase, a higher percentage than for NOHATE vocabulary, of which 6.57% of the words can be found in Hatebase [1]. Moreover, the results in terms of accuracy of the generated models that classified Hate, No-hate and All are lower when including sentences that required additional context. The results also show that No-hate sentences are more accurately classified than Hate sentences.

What is their conclusion? They state that the constructed dataset provides a good starting point for discussion and further research.

What are -according to you- interesting aspects of the paper? I find it interesting that the paper is not only focussing on sentences individually, but incorporate the whole context around them. Moreover they explain the process of collecting data and annotating very detailed before the actual classification, which makes the experiment, again, very reproducible.

Do you have clarifying questions? Did you understand everything? I wonder who the annotators were and why they were chosen. What makes them good annotators? Moreover, they only focus on groups when classifying a message as hate-speech. They state that the concept of hate-speech is excessively subjective. But wouldn't it be the case that the classification procedure can be better at detecting hateful comments when individuals are also included? Perhaps if they flagged instances that meet at least one requirement out of those three that would make a comment an hate-speech, it could increase the chances of messages containing hate-speech.

References

1. de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate speech dataset from a white supremacy forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 11–20. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-5102, https://aclanthology.org/W18-5102
2. Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of Hindi-English code-mixed data. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1226
3. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1415–1420. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1144, https://aclanthology.org/N19-1144