The Annotators Agree To Not Agree On The Fine-grained Annotation of Hate-speech against Women in Algerian Dialect Comments

Imane Guellil[1], **Yousra Houichi**[2], **Sara Chennoufi**[3], **Mohamed Boubred**[4], **Anfal Yousra Boucetta**[5], **Faical Azouaou**[6]
[1]University of Edinburgh, Edinburgh, United Kingdom
[2]EZUS, France,
[3]LIRIS, France,
[4]Capgemini, France
[5]Ecole Superieure d'Informatique D'Alger (ESI), Algeria,
[6]Ecole supérieure en Sciences et Technologies de l'Informatique et du Numérique (ESTIN), Algeria

Abstract

A significant number of research studies have been presented for detecting hate speech in social media during the last few years. However, the majority of these studies are in English. Only a few studies focus on Arabic and its dialects (especially the Algerian dialect) with a smaller number of them targeting sexism detection (or hate speech against women). Even the works that have been proposed on Arabic sexism detection consider two classes only (hateful and non-hateful), and three classes(adding the neutral class) in the best scenario. This paper aims to propose the first fine-grained corpus focusing on 13 classes. However, given the challenges related to hate speech and fine-grained annotation, the Kappa metric is relatively low among the annotators (i.e. 35%). This work in progress proposes three main contributions: 1) Annotation of different categories related to hate speech such as insults, vulgar words or hate in general. 2) Annotation of 10,000 comments, in Arabic and Algerian dialects, automatically extracted from Youtube. 3) Highlighting the challenges related to manual annotation such as subjectivity, risk of bias, lack of annotation guidelines, etc.

Keywords: Sexism detection, hate-speech detection, corpus construction, manual annotation

1. Introduction

Hate speech is commonly defined as a language to express hatred against a specific person or a group based on certain key characteristics such as religion, gender, race, sexual orientation, and various disability forms (Shannaq et al., 2022). The excessive use of social media leads to the rise of antisocial behaviours illustrated in the spread of online hate speech, offensive language and cyberbullying (Shannaq et al., 2022). Authorities in many countries are recognizing hate speech as a serious problem as it can lead to depression which hurts people's health and relationships. It can also lead to suicide in more serious scenarios (Boucherit and Abainia, 2022).

With the online proliferation of hate speech, a significant number of research studies focusing on how to classify and detect this kind of speech have been presented in the last few years. The majority of these studies detect general hate speech (Caiani et al., 2021; Pamungkas et al., 2018; Almatarneh et al., 2019; Kalaivani and Thenmozhi, 2021) and only a few studies (de Paula et al., 2021) focused on the detection of hate speech against women (only by distinguishing between hateful and non-hateful comments). However, almost all studies are dedicated to English. This is mainly due to the lack of resources (lexicons and corpora that are constructed for other languages such as Arabic). To bridge the gap, the role of this paper is to propose a fine-grained manually annotated corpus including 10,000 YouTube comments and 13 classes: 0 (no hate), i (insult), v (vulgar), h (hate), s (without relationships with women), b (positive), p (a problem in the annotation), e (emojis only), c (passage from Coran, Muslims book), iv (insults and vulgar in the same time), ih (insult and hate in the same time), vh (vulgar and hate in the same time), ivh (insult, vulgar and hate in the same time). This corpus will be freely available to the research after its publication. The main conclusion from this work was that annotators tend to disagree more frequently when they have to deal with different annotation classes.

2. Arabic Hate Speech in social media: Challenges

Arabic is a language spoken by more than 330 million people as a native language. It is the fifth most spoken language in the world. Modern Standard Arabic (MSA) is usually the official language used in school whereas the classical is used in the Holy Qur'an (Muslim's book) (ESI, 2016; Guellil et al., 2020b). Another form of Arabic is the Arabic di-

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 133–139
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

alects which are used in daily life conversations. Also, Arabic in social media can be written either by using Arabic letters or Arabizi (Latin letters) (Guellil et al., 2021). 55% of the text in social media was written in Arabic (2017) (Haddad et al., 2020). Arabic Natural Language Processing (NLP) applications have to deal with several complex challenges in addition to the common challenges related to any NLP problems(Guellil and Faical, 2017; Guellil et al., 2018).

Arabic is known for its challenges, scarcity of resources and complexity. Detecting hate speech for Arabic content is a complex task (Husain, 2020). Different challenges can be raised when detecting hate speech in Arabic text: 1) The informal language using short forms and slang. 2) The use of dialects (Boucherit and Abainia, 2022). 3) The diversity of the Arabic language dialects (Husain, 2020). 4) The use of Arabizi (Guellil et al., 2020a)

3. Related Work on Arabic hate-speech corpora creation

Some papers focused on resources constructions dedicated to hate-speech detection (Albadi et al., 2018; Mubarak et al., 2022; Alsafari et al., 2020; Mubarak et al., 2020; Chowdhury et al., 2020; Almanea and Poesio; El Abboubi et al., 2020; Boucherit and Abainia, 2022; Guellil et al., 2022). (Albadi et al., 2018) aims to detect religious hate speech in the Arabic language on social media[1]. The authors started with constructing their dataset by collecting tweets and annotating them manually. For this purpose, They first collected 6,000 Arabic tweets referring to different religious groups and labelled them using crowdsourced workers. After this, they analysed the labelled dataset and reported the main targets of religious hatred in the Arabic Twitter space.

In the paper of Mubarak et al. (2022), the authors present an automated emoji-based approach of collecting tweets that have a much higher percentage of malicious content, without having any language dependency. From a collection of 4.4M Arabic tweets between June 2016 and November 2017, they extracted all tweets having any of the used emojis. An annotation job was created on the Appen crowdsourcing platform to judge whether a tweet is offensive or not. Annotators from all Arab countries were invited.

The role of the paper described by Alsafari et al. (2020) was to create a reliable Arabic textual corpus. The Data was extracted from Twitter based on a list of Arabic keywords related to each of the four categories under study: religion, ethnicity, nationality and gender. The authors randomly selected 200,000 posts for each category, with a total of 800,000 samples. The annotation has been carried out by three Gulf native speakers, two females and one male.

The paper of Mubarak et al. (2020) is adding an additional class to those which are generally studied, where these authors also identify vulgar comments in addition to comments including hate. The Twitter APIs were used to collect 660k Arabic tweets between April 15 – May 6, 2019. The tweets were annotated, ending up with 1,915 offensive tweets. Each tweet was labelled as offensive, which could additionally be labelled as vulgar and/or hate speech, or Clean.

The main idea of Chowdhury et al. (2020) was to introduce a new dialectal Arabic news comment dataset, collected from multiple social media platforms, including Twitter, Facebook, and YouTube. From 2011 to 2019, over 100k comments from different social media platforms were collected. The contents from each platform were collected through its own API (YouTube, Facebook, and Twitter). Data annotation (Amazon Mechanical Turk (AMT), a crowdsourcing platform, was used to obtain manual annotations. The comments were annotated for hate speech and vulgar (but not hate) categories. The authors analyzed the distinctive lexical content along with the use of emojis in offensive comments.

The aim of Almanea and Poesio was to introduce an Arabic misogyny and sexism dataset (ArMIS) characterized by providing annotations from annotators with different degrees of religious beliefs and providing evidence that such differences do result in disagreements. The authors discussed proof-of-concept experiments showing that a dataset in which disagreements have not been reconciled can be used to train state-of-the-art models for misogyny and sexism detection; and considered different ways in which such models could be evaluated.

The aim of El Abboubi et al. (2020) was to discuss both the impact of possible sex-based differences and the awareness and recognition of sexist attitudes in Moroccan Arabic. The findings of this study are based on quantitative data. The patterns analyzed are the following: sexist attitudes, self-assessment, sources of pressure to use or change sexist language, and recognition of sexist language. A questionnaire was designed to measure attitudes. The questionnaire is divided into two parts: one in which five questions are asked to reflect the respondents' attitudes towards Moroccan Arabic as a sexist language; and a second part in which statements are presented to respondents who rate them considering the extent to which they are sex-

[1] https://github.com/nuhaalbadi/Arabic_hatespeech

ist, and if those same statements are appropriate or not.

The paper of Boucherit and Abainia (2022) addresses the problem of detecting offensive and abusive content in Facebook comments, where the focus is on the Algerian dialectal Arabic. The authors have built a new corpus regrouping more than 8.7k texts manually annotated as normal, abusive and offensive (where 10,258 comments have been initially collected from public pages and groups related to sensitive topics).

In the paper of Mulki et al. (2019), the authors introduced the Levantine Hate Speech and Abusive (L-HSAB) Twitter dataset to be a benchmark dataset for automatic detection of online Levantine toxic contents. Three annotators manually labelled the tweets following into 3 categories: Normal, Abusive and Hate. Waseem et al. (Waseem and Hovy, 2016) manually annotated the dataset containing 16,914 tweets where 3,383 tweets were for sexist content, 1,972 for racist content, and 11,559 for neither sexist nor racist. For dataset generation, the authors used Twitter API to extract tweets containing some keywords related to women. The work of Waseem et al. (Waseem and Hovy, 2016) is considered a benchmark by many researchers (Al-Hassan and Al-Dossari, 2019; Pitsilis et al., 2018; Kshirsagar et al., 2018).

Finally, our recent work Guellil et al. (2022), also considered YouTube for constructing a corpus of 5,000 comments dedicated to sexism detection. However, we only considered two labels for annotating their dataset: Hateful and non-hateful comments.

4. Data collection and annotation

4.1. Data collection

Youtube comments related to videos about women are used. A feminine adjective such as: جميلة meaning beautiful, جَايحة meaning stupid or كلبة meaning a dog are targeted. A video on YouTube is recognised by a unique identifier (*video_id*). For example, the video having an id equal to *"TJ2WfhfbvZA"* handling a radio emission about unfaithful women and the video having an id equal to *"_VimCUVXwaQ"* advises women to become beautiful. Three annotators manually reviewed the obtained video from the keyword and manually selected 335 *video_id*. We used Youtube Data API[2] and a Python script to automatically extract comments of each *video_id* and their replies. In the end, we were able to collect *373,984* comments extracted, we call this corpus *Corpus_Youtube_women_10000*.

[2]https://developers.google.com/youtube/v3/

4.2. Data annotation

For the annotation, we randomly select 10,000 comments containing MSA and Algerian dialects written in Arabic and Arabizi. This corpus also contains some comments in French and others in English (As most of the Arabic people are bilingual). The annotation was done by three annotators, native speakers of Arabic and its dialects (2 women and one man). The annotators were separated and they had 3 weeks to manually annotate the selected comments using different labels. An annotation guideline was prepared for this purpose and it was shared with the annotators. The main points of this guideline are:

- The value of the column *hate* can be given multiple values: 0 (comment containing no hate, no insult, no vulgar word), i (if the comment contains insults, for example, *ya kalba meaning dog*, *ya hmara meaning donkey*, etc,), v (if the comment contains vulgar words), h (if the comment contains hate, for example *allah ya3tik elmoutI want meaning that you die*, or *we will dance on your grave, etc)*.
- If it has a comment that contains several characteristics at the same time, they had to mention it. For example, if a comment contains hate and vulgarities, you had to put vh (and not hv), in the same order i, v then h- had to be kept.
- The authors were asked to be as objective as possible for this annotation and not incorporate their personal feelings.
- As the comments were extracted automatically, it is also possible to find some comments with no relationship with women As an example *four lghounia hadi kho meaning this song is amazing bro*, They were asked to put the letter s (without interest)
- They were asked to put the letter p (problem) When they were facing a situation where they could not decide what to put. However, they were asked to use this option only when it is necessary.
- As we plan to use this corpus for sentiment analysis purposes as well, the annotators were also asked to put a b for the positive comments.
- The comments including only emojis without text should be annotated with the label e
- The comments including punctuation only should be annotated with the word "po"

Table 1: Agreement of the three annotators on the different labels

Label	rater1	rater2	rater3	Agreement
0	6723	3310	1206	695
i	1109	749	1843	285
v	266	366	338	107
h	106	1012	128	41
s	222	1583	3912	121
b	1027	2199	1809	869
p	103	8	129	0
e	82	25	0	0
c	8	1	8	0
iv	138	72	219	9
ih	115	402	151	27
vh	7	70	25	1
ivh	12	131	40	2

- The comments including only some text from the Coran should be annotated as c

4.3. The constructed corpus

The table illustrates the number of labels given by each annotator. Rater1 and Rater2 are women. Rater3 is a man.

Table 1 illustrates the agreement of the three annotators on the different labels. From this table, we observe that the annotators tend to more agree on the positive reference or the non-hateful references than they agree on the hateful comments. We also observe that the disagreement is higher for comments including more than one hateful class (such as comments including insults and vulgar words simultaneously). Finally, we can also observe that three tendencies of annotations are among our annotators. We have the careful annotation (Rater1) when the annotator does not assign a label only when she is sure. We have the extreme annotation (Rater2) when the annotators assigned the majority of the labels and we have the moderate annotator (Rater3) who tends to be in between the two previous annotators.

In order to highlight the inter-agreement among annotators we also present Figure 1 illustrating the Kappa-agreement between each two annotators. We observe that this rate is especially low between rater1 and rater3 (19%). The best agreement is between Rater2 and Rater3. We observe that fine-grain annotation with many classes returns a low kappa (illustrated in Figure 1). One of the reasons behind this is the typing errors related to some labels. This is also caused by the non-application of the guideline. For example, one of the annotators created another label ("other") when he should have used the label p for the problems. Another cause of conflicts is when the authors have to attribute different labels to the same comments. We observe a lack of consistency where some annotators misplaced the labels.

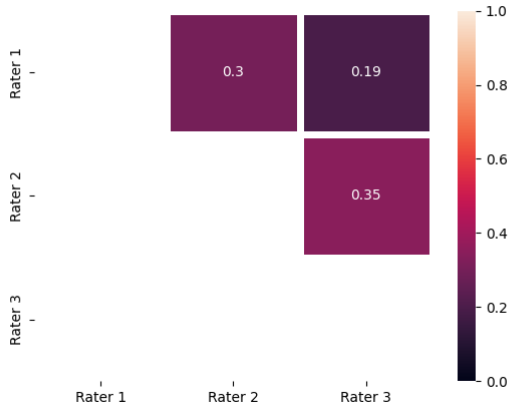

Figure 1: Intra-agreement among annotators

5. Discussion

In total 10,000 comments that were randomly selected were annotated by three annotators. However, we can observe that the inter-agreement among annotators (Kappa) was really low. This highlights how complicated is the annotation with many labels. In total, the 3 annotators agree on 2,157 (22%) comments from the 10,000 that they initially reviewed.

The main goal of this paper is to propose a resource for fine-grained hate speech detection. However, this resource can also be used for binary classification (when the research aims to only detect hate speech against women). In order to do that, we need to first separate the labels into two categories to distinguish between hateful and not-hateful comments. We decide to recognise the labels 0, s, b, p, e, c as non-hateful and the others (i, v, h, iv, ih, vh, ivh) as hateful. We also resolve some obvious annotation errors such as the one related to the tag "other" that we recognise as non-hateful. In that case, we observe that the three annotators agree on 1165 hateful comments and on 6219 non-hateful comments (a total of 7384 comments). The intra-agreement among annotators is illustrated in Figure 2. We observe in this figure that Kappa significantly improves, especially between the second and the third annotators where Kappa with two classes is up to 0.68 (considered to be a good degree of agreement (Salkind, 2010)). Hence, in all cases, we observe that Rater2 is providing the highest agreement.

The main challenge when annotating a corpus with many labels is the consistency of annotation guide-

lines. The annotators have different questions at the start of the annotation phase. The best way to do this would be to have an annotation pilot by selecting only a few documents (around 20) having them annotated by the three annotators and having a discussion for resolving the disagreements before starting the annotation. Another issue is the lack of consistency among the annotators. Some annotators created new classes when others did not respect the annotation format. One way to resolve this would be to automatically detect this incoherence and have it reviewed manually again by the annotators.

This corpus may be used in different ways. The first one would be to train a binary classifier for detecting hateful and not hateful comments. We can observe that the agreement for the binary classification is pretty good. However, the main aim of this corpus is to train a multi-class classifier in order to automatically distinguish among hate, insult and vulgar comments used against women in social media. The main challenge behind this would be the imbalance of the different classes. We can consider the augmentation of some classes. We can also consider algorithms dedicated to handling imbalanced corpora such as the Synthetic Minority Oversampling Technique (SMOTE)

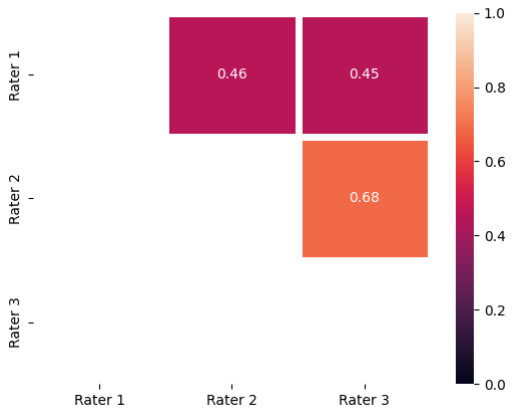

Figure 2: Intra-agreement among annotators

6. Conclusion

We constructed in this paper the first fine-grained corpus for Arabic/Algerian dialect hate speech against women detection. We focus on Arabic/Algerian dialect but we plan to extend this construction to other dialects such as Morrocan or Tunisian. We plan to extend this construction to other African languages as well. This corpus includes 14 labels and is distinguished among the general hate, insults and vulgar comments. Our future would be to automatically review some disagreements related to the mismatch of labels, upper-case, etc. We also plan to have this annotation reviewed by a fourth annotator who will have access to the different assigned labels in addition to the comments. We also plan to use the constructed corpus in order to train ML algorithms for fine-grained classification.

Acknowledgement

We would like to thank the Edinburgh Futures Institute (EFI)[3] for supporting the fees related to paper presentations and research trips leading to such quality papers. The purpose of the Edinburgh Futures Institute (EFI) is to pursue knowledge and understanding that supports the navigation of complex futures.

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. *Computer Science & Information Technology (CS & IT)*, 9(2):83.

Albadi, Nuha and Kurdi, Maram and Mishra, Shivakant. 2018. *Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere*. IEEE.

Dina Almanea and Massimo Poesio. Armis-the arabic misogyny and sexism corpus with annotator subjective disagreements.

Sattam Almatarneh, Pablo Gamallo, Francisco J Ribadas Pena, and Alexey Alexeev. 2019. Supervised classifiers to identify hate speech on english and spanish tweets. In *International Conference on Asian Digital Libraries*, pages 23–30. Springer.

Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. 2020. Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, 19:100096.

Oussama Boucherit and Kheireddine Abainia. 2022. Offensive language detection in under-resourced algerian dialectal arabic language. *arXiv preprint arXiv:2203.10024*.

Manuela Caiani, Benedetta Carlotti, and Enrico Padoan. 2021. Online hate speech and the radical right in times of pandemic: The italian and english cases. *Javnost-The Public*, 28(2):202–218.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J

[3]https://efi.ed.ac.uk/

Jansen, and Joni Salminen. 2020. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the 12th language resources and evaluation conference*, pages 6203–6212.

Angel Felipe Magnossão de Paula, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. *arXiv preprint arXiv:2111.04551*.

Zineb El Abboubi, Ahmadou Bouylmani, and Mohammed Derdar. 2020. Sexism in moroccan arabic: Gender differences in perceptions and use of language. *Journal of Applied Language and Culture Studies*, 3:215–230.

Karima Benatchba Prof President ESI. 2016. *A Sentiment analysis approach for Arabic dialects texts analysis based on automatic translation: Application to the Algerian dialect.* Ph.D. thesis, ESI Algeria.

Imane Guellil, Ahsan Adeel, Faical Azouaou, Mohamed Boubred, Yousra Houichi, and Akram Abdelhaq Moumna. 2022. Ara-women-hate: An annotated corpus dedicated to hate speech detection against women in the arabic community. page 68–75.

Imane Guellil, Faical Azouaou, Fodil Benali, Ala Eddine Hachani, and Marcelo Mendoza. 2020a. The role of transliteration in the process of arabizi translation/sentiment analysis. *Recent Advances in NLP: The Case of Arabic Language*, pages 101–128.

Imane Guellil, Faical Azouaou, Fodil Benali, alaeddine Hachani, and Houda Saadane. 2018. Approche hybride pour la translitération de l'arabizi algérien : une étude préliminaire. In *Conference: 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN), May 2018, Rennes, FranceAt: Rennes, France*. https://www.researchgate.net/publication ....

Imane Guellil, Faical Azouaou, and Francisco Chiclana. 2020b. Arautosenti: automatic annotation and new tendencies for sentiment classification of arabic messages. *Social Network Analysis and Mining*, 10:1–20.

Imane Guellil and Azouaou Faical. 2017. Bilingual lexicon for algerian arabic dialect treatment in social media. In *WiNLP: Women & Underrepresented Minorities in Natural Language Processing (co-located with ACL 2017)*.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Bushr Haddad, Zoher Orabe, Anas Al-Abood, and Nada Ghneim. 2020. Arabic offensive language detection with attention-based deep neural networks. In *Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection*, pages 76–81.

Fatemah Husain. 2020. Arabic offensive language detection using machine learning and ensemble machine learning approaches. *arXiv preprint arXiv:2005.08946*.

Adaikkan Kalaivani and Durairaj Thenmozhi. 2021. Multilingual hate speech and offensive language detection in english, hindi, and marathi languages.

Rohan Kshirsagar, Tyus Cukuvac, Kathleen McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. *arXiv preprint arXiv:1809.10644*.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2022. Emojis as anchors to detect arabic offensive language and hate speech. *arXiv preprint arXiv:2201.06723*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. Automatic identification of misogyny in english and italian tweets at evalita 2018 with a multilingual hate lexicon. In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Neil Salkind. 2010. Cohen's kappa. pages 188–189.

Fatima Shannaq, Bassam Hammo, Hossam Faris, and Pedro A Castillo-Valdivieso. 2022. Offensive language detection in arabic social

networks using evolutionary-based classifiers learned from fine-tuned embeddings. *IEEE Access*, 10:75018–75039.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.