Received 9 October 2021
Revised 29 January 2022
4 March 2022
Accepted 6 March 2022

Evaluating transfer learning approach for detecting Arabic anti-refugee/migrant speech on social media

Djamila Mohdeb
University of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria
Meriem Laifa
University of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria and
Laboratory of Informatics and its Applications of M'sila (LIAM), M'sila, Algeria, and
Fayssal Zerargui and Omar Benzaoui
University of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria

Abstract

Purpose – The present study was designed to investigate eight research questions that are related to the analysis and the detection of dialectal Arabic hate speech that targeted African refugees and illegal migrants on the YouTube Algerian space.

Design/methodology/approach – The transfer learning approach which recently presents the state-of-the-art approach in natural language processing tasks has been exploited to classify and detect hate speech in Algerian dialectal Arabic. Besides, a descriptive analysis has been conducted to answer the analytical research questions that aim at measuring and evaluating the presence of the anti-refugee/migrant discourse on the YouTube social platform.

Findings – Data analysis revealed that there has been a gradual modest increase in the number of anti-refugee/migrant hateful comments on YouTube since 2014, a sharp rise in 2017 and a sharp decline in later years until 2021. Furthermore, our findings stemming from classifying hate content using multilingual and monolingual pre-trained language transformers demonstrate a good performance of the AraBERT monolingual transformer in comparison with the monodialectal transformer DziriBERT and the cross-lingual transformers mBERT and XLM-R.

Originality/value – Automatic hate speech detection in languages other than English is quite a challenging task that the literature has tried to address by various approaches of machine learning. Although the recent approach of cross-lingual transfer learning offers a promising solution, tackling this problem in the context of the Arabic language, particularly dialectal Arabic makes it even more challenging. Our results cast a new light on the actual ability of the transfer learning approach to deal with low-resource languages that widely differ from high-resource languages as well as other Latin-based, low-resource languages.

Keywords Hate speech, Anti-migrant speech, Algerian dialectal Arabic, African migrants, Transfer learning, Arabic natural language processing

Paper type Research paper

1. Introduction

The rise of digital media supported by social networks and citizen journalism has changed the way audiences receive and react to information that is related to the concerns of public interest. Online social media platforms are no longer used for only communication and socialization, but rather the most important tools for influencing and shaping public opinion, the role that the traditional mass media such as newspapers and television have monopolized for decades.

The content produced by social media users has serious issues that are mainly related to the information credibility as well as the difficulty to control or influence the orientations of the content creators who enjoy a high degree of freedom as they are not subject to the restrictions imposed by traditional media institutions. That should bring us to the deviations caused by the misuse of freedom of expression in the online social space in which the spread of hate speech is its darkest face.

Aslib Journal of Information Management
Vol. 74 No. 6, 2022
pp. 1070-1088

2050-3806
DOI 10.1108/AJIM-10-2021-0293

Hate speech is a contested term that has a variety of definitions. We find the more comprehensive definition suggested by Fortuna and Nunes (2018) who explained hate speech as the "language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used".

Hate speech phenomenon is the undesirable outcome of what is called in social sciences the "Othering" process that creates the norms of the Us–Them dichotomy in a community (Burnap and Williams, 2015). This process is motivated to a certain extent by an exclusive character that defines who belongs to the in-group and accordingly who is part of the out-group (Burnap and Williams, 2015). "Othering" in a hateful discourse is based on "the conscious or unconscious assumption that a certain identified group poses a threat to the favored group" (Powell, 2017). The members of the out-group "are othered depending on how noticeable their differences are perceived as a threat in a specific context" (Powell, 2017).

According to Gagliardone *et al.* (2015), the majority of studied cases of online hate speech aim at individuals based on ethnicity and nationality. In this context, refugees and illegal migrants are amongst the most vulnerable targeted categories as long as they are perceived by host communities as a menace to their social, economic and cultural quality of life (Reidpath and Allotey, 2018).

The migration crisis which is generated by major conflicts and natural disasters in Asia, the Middle East and Africa is one of the most debated topics in the last decade on either mass media or online social media (Reidpath and Allotey, 2018). An anti-refugee discourse has overwhelmed the online social platforms triggering hate and fear towards refugees, migrants and forced migrants, causing concern amongst scholars and policymakers. This hateful rhetoric has marked online debates not only in Europe (Himmel and Baptista, 2020) but also in many countries that were obligated to host large number of refugees in spite of its political and economic instability (Özerim and Tolay, 2021).

Recent research pushed the question of online hate speech detection into new directions using automatized methods and systems. Automatic hate speech detectors are mainly based on machine learning and Natural Language Processing (NLP) techniques (Siegel, 2020). Nonetheless, the effectiveness of these techniques turns out to be even more problematic due to the controversy that surrounds the term "hate speech" and its relation with the freedom of opinion and expression. On this basis, social media that use automatic hate speech detection and removal are occasionally attacked and accused of bias and censorship during complex situations such as political conflicts, e.g. Israeli–Palestinian conflict and Black Lives Matter protests (Dwoskin and De Vynck, 2021), and elections, e.g. US presidential election 2020 (Clayton, 2021). Language dependency is another important reason for the restricted performance of hate detection techniques when dealing with languages other than English, especially non-Latin-based, low-resource languages. This could be justified from one side by the relatively linguistic complexity of some of these languages and from the other side by the lack of sufficient textual data resources for training and experiments (Pires *et al.*, 2019).

Relying on transferring knowledge across domains, transfer learning is a promising machine learning methodology for overcoming this latter challenge. To simplify, transfer learning is a technique in which a model that has been pre-trained on a specific source task can be re-utilized on a different but related target task. Thus, rather than collecting large amount of data to improve target learners' performance, a pre-trained model reuses the knowledge acquired from being already trained with a large volume of other data on the source task. This approach, with its broad application possibilities, can present practical solutions to the problem of automatic hateful content detection in low-resource languages.

In this paper, we investigate the transfer learning approach for automatically detecting the hateful aspect of the anti-refugee/migrant Arabic speech on social media using the pre-trained

language models mBERT (Devlin *etc al.*, 2018), XLM-R (Conneau *et al.*, 2019), AraBERT (Antoun *et al.*, 2020) and DziriBERT (Abdaoui *et al.*, 2021). In order to make the task more challenging, the chosen models were applied to a collection of multilingual hate-related texts that were mainly written in the Algerian dialectal form of the Arabic language. The harvested texts which were extracted from the comment sections in YouTube social platform discussed the subject of the "African refugee/migrant crisis" in Algeria during the period from 2014 to 2021.

Our main contributions can be summarized as follows:

(1) Collecting and annotating a new corpus of anti-refugee/migrant data.

(2) Analyzing the properties of the anti-refugee/migrant content and its actual prevalence on the YouTube Algerian space.

(3) Assessing the performance of monolingual and multilingual pre-trained language models with the aim to provide important insights for future work comparison on the Arabic hate speech detection task.

(4) Determining the impact of the lingual variation on the pre-trained language models.

(5) Evaluating the impact of the presence of lingual noise on the performance of transformers.

The remainder of this paper is organized as follows. In Section 2, we provide the background of the crisis of refugees and migrants in Algeria. In Section 3, we review the related work on the topic of hate speech detection. Our methodology steps are described in detail in Section 4. Experiments' results on the hate speech detection task are discussed in Section 5. Finally, we conclude the paper by deliberating the implications and limitations of this work.

2. Migration crisis in Algeria: a background

Since 2012, Algeria has lived through the repercussions of the massive movements of refugees generated by the political instability in the Middle East, the Sahel, Libya and sub-Saharan Africa. In this paper, we have limited our study to specifically the sub-Saharan African refugees and illegal migrants.

While it is commonly assumed that Algeria, for sub-Saharan illegal migrants, is a transit towards European countries, a large number of them consider it as their final destination (Musette, 2014). According to Algerian government's statistics dating back to 2017, every day, an average of 500 migrants irregularly enter its territory (Hafid, 2019).

In June 2017, an anti-refugee/migrant hashtag targeting sub-Saharan migrants went viral on social media platforms. The participants in the online debate asked for a speedy solution for this "uncontrollable" number of refugees and migrants in their country (Attia, 2017). The online mediated anti-refugee discourse was the same as the anti-refugee rhetoric in other countries of the world that were facing the same situation. The refugees were blamed for the charges the host community had to carry in regard to the negative implications of the crisis. Other accusations raised the fear of the public about the sociocultural differences of migrants and the possibility of influencing the demographic composition of the host society. Facing pressure, the Algerian government talked about a scheme to grant residency rights and job permits to illegal African migrants (Chikhi, 2017). On the other hand, it has resorted to regularly deporting a number of migrants to their countries of origin, a policy that was criticized by international human rights organizations (Attia, 2017) but supported by a relatively large segment of the Algerian public on social networks. Despite all the efforts that have been made, the problem still stands. The anti-refugee/migrant speech seemingly appears and disappears on social networks dividing Algerian society as well as the political class between sympathizers and rejectionists.

3. Literature review

Hate speech has been considered as a broad umbrella term for many types of targeted non-civil content. Fortuna and Nunes's survey (2018) have shown that hate speech was related to numerous concepts including cyberbullying, abusive language, offensive language, discrimination, profanity, toxicity, flaming, extremism and radicalization.

Over time, a large and growing body of research has been developed for better understanding and detecting the online hate content.

3.1 Automatic hate speech detection

While reviewing the literature from a computer science point of view, we found that automated hate speech detection is mostly based on NLP techniques combined with supervised machine learning (e.g. Logistic Regression, Support Vector Machine and Random Forests) (Burnap and Williams, 2015), deep learning (e.g. CNN, LSTM and GRU) (Mishra *et al.*, 2018) and recently transfer learning models (e.g. BERT) (Alatawi *et al.*, 2021).

The hate speech detection task is usually considered as a binary (Hewitt *et al.*, 2016), multi-class (Davidson *et al.*, 2017) or multi-label (Ousidhoum *et al.*, 2019) classification task. Surveys such as Fortuna and Nunes (2018) and Siegel (2020) have shown that the applied models use a variety of text mining features ranging from general features such as TF-IDF (Dinakar *et al.*, 2011), BoW (Greevy and Smeaton, 2004), N-Gram frequencies (Greevy and Smeaton, 2004), dictionaries (Liu and Forss, 2015), sentiment analysis (Ousidhoum *et al.*, 2019), topic classification (Agarwal and Sureka, 2017), word embeddings (Siegel *et al.*, 2017), distance metrics (Warner and Hirschberg, 2012), etc. to textual features that are specifically related to the language styles of the hate speech such as stereotyping (Alorainy *et al.*, 2018), othering language (ElSherief *et al.*, 2018), subjectivity (Gitari *et al.*, 2015) and perpetrator characteristics (Waseem and Hovy, 2016).

3.2 Transfer learning for hate speech detection

Recently, transfer learning is the state-of-the-art technique in NLP. Pre-trained language models have shown a more enhanced performance in different NLP tasks beating the classical machine learning and deep learning models in different settings (Zhuang *et al.*, 2020). In the context of online hate speech detection, pre-trained language models have been exploited in a number of studies revealing interesting performance.

Caselli *et al.* (2020) introduced HateBERT, a retrained BERT model for offensive and abusive language detection in English. HateBERT is the result of retraining the general English BERT base uncased model on a dataset that contains around 71.1 million tokens from banned communities in Reddit. It was evaluated on three different datasets, each representing a different language phenomenon: offensive language, abusive language and hate speech. While the in-dataset evaluation (i.e. evaluating the fine-tuned model on each of the three datasets) shows that the proposed fine-tuned model outperforms the general BERT, the cross-dataset evaluation (i.e. evaluating the fine-tuned model on all the combinations of the three datasets) suggests that the portability of HateBERT is influenced by the compatibility of the annotated phenomena.

The study of Bigoulaeva *et al.* (2021) that have addressed the task of automatic hate speech detection for low-resource languages noticed that using cross-lingual transfer learning to leverage already existing hate speech data from higher resource languages is an effective way of achieving good performance on low-resource target languages without collecting and annotating new data.

Beyond these papers, we have observed that the studies that have applied transfer learning combine usually fine-tuned, pre-trained language models with deep neural architecture for automatic detection tasks. We cite, for example, the work of Rizoiu *et al.* (2019) who proposed transfer Deep Hate (t-DeepHate), a novel pipeline for hate speech detection. The pipeline combines transfer learning methods with a bidirectional Long Short-Term Memory neural

network (bi-LSTM) deep architecture. The bi-LSTM architecture adapts pre-trained word and sentence embeddings to the hate speech domain and the transfer learning methods are utilized to construct a single set of hate speech embeddings while leveraging multiple datasets.

The Class Representation Attentive BERT neural model introduced by Zahiri and Ahmadvand (2020) outperforms the state-of-the-art BERT-based baseline by 1.89% on relative Macro F1. It leverages matching scores between trainable class representations and encoded input data to successfully detect online hate speech.

3.3 Detection of anti-refugee hate speech

Understanding the hateful aspect of anti-refugee/migrant speech is mandatory since the hate targets are perceived as an economic "burden" and a threat to the sociocultural homogeneity of the host community (Reidpath and Allotey, 2018). We notice that the category "refugee" is usually confused with the category "migrant" or "immigrant". Therefore, we should clarify that migrants and immigrants are people who willingly leave their countries of birth in search of better life chances. The "immigrant" settles permanently and legally in the host country, whereas the "migrant's" settlement is temporary and is restricted to a legal specific period. On the other hand, "refugees", according to the Office of the United Nations High Commissioner for Refugees (UNHCR), are people who are "fleeing armed conflict or persecution" and "for whom denial of asylum has potentially deadly consequences" (UNHCR, 2016).

Although significant literature has discussed the anti-refugee discourse in social sciences, far too little attention has been paid to anti-refugee speech detection in the computer science field. An early related work on this topic is that of Ross *et al.* (2017) who compiled the first corpus of German refugee-related hate speech from Twitter with the aim of measuring the reliability of hate speech annotations. The study reached different conclusions, finding that the ambiguity in the existing hate speech definitions negatively influences the quality of hate speech annotations.

Zhang *et al.* (2018) created RM, a dataset consisting of English tweets that discuss the topic of refugees and Muslims in the UK. The RM data have been used along with six other datasets to evaluate the performance of a deep neural network combining Convolutional and Gated Recurrent Networks. The proposed model outperformed the state-of-the-art baselines as it has had the ability to capture both word sequence and order information in short texts.

Köffer *et al.* (2018) investigated the potential value of automatic analytics of German texts to detect hate speech that has targeted refugees during the refugee crisis 2015/2016 in Europe. They collected their dataset from the comment section of mainstream journalistic news websites as well as websites of so-called alternative media. Annotation of texts as hateful or not hateful relied on a crowdsourcing rating system via an online survey. The authors have then used N-grams and distributional semantics features (Word2Vec, Doc2Vec) to detect hateful content with baseline classification methods.

A recently interesting study by Perifanos and Goutsos (2021) presented a new multimodal approach to hate speech detection by combining computer vision and NLP models for detecting hateful speech in Greek aimed at refugees and migrants. The proposed model which is a combination of a fine-tuned, pre-trained language model from BERT and Residual Neural Networks has been trained on a dataset that assembles hateful textual and visual data collected from Twitter. The model reported a consistently high accuracy in anti-refugee speech detection.

3.4 Arabic hate speech detection

Hate-related studies for the Arabic language have investigated different settings for hate speech detection. For instance, Mulki and Ghanem (2021) introduced LeT-Mi, the first Arabic Levantine Twitter dataset for Misogynistic language. LeT-Mi was exploited as an evaluation dataset for binary/multi-/target classification tasks performed by a number of baseline and

deep learning classifiers, along with a Multi-Task Learning (MTL) set-up. The obtained results indicate that employing MTL has improved the performance of the classification task.

Al-Hassan and Al-Dossari (2021) established a comparison between four deep learning models and SVM to detect Arabic hateful content in a dataset of 11 K tweets labelled with five classes. The results showed a good performance of deep learning models LSTM, CNN + LSTM, GRU and CNN-GRU compared to SVM baseline in the multi-classification of hate classes. However, the ensemble model of CNN + LTSM produced the best results.

Alshalan and Al-Khalifa (2020) investigated Convolutional Neural Network (CNN), Gated Recurrent Units (GRU) and CNN + GRU ensemble models in addition to the transformer multilingual BERT (mBERT) to detect hate speech in a public dataset of 9,316 Arabic tweets labelled as hateful, abusive and normal. They found that CNN outperforms the other models with a significant difference in the classification task and that mBERT fails to improve over the baselines and the other evaluated models.

Haddad *et al.* (2020) used CNN and Bidirectional Gated Recurrent Unit (Bi-GRU) augmented with attention layers to detect Arabic offensive language and hate speech. The attention added layer enabled the Bi-GRU model to effectively raise precision scores, thus achieving the highest results compared to other models.

3.5 Literature gaps

Despite the importance of studies in Arabic hate speech detection, the generalizability of published research on this issue is problematic. Many reasons for this drawback are possible. First, the research to date has tended to focus on Modern Standard Arabic (MSA) or Middle Eastern dialects (Antoun *et al.*, 2020), yet Arabic has a wide range of other dialectal variants including North African dialects which have their own diversities and complexities. As an example, to the best of our knowledge, no study has investigated hate speech detection in Algerian dialectal Arabic. Second, the existing hate speech corpora are mainely extracted from Twitter, i.e. other popular social media platforms or web content in general are neglected. Moreover, with exception to some studies (Mulki and Ghanem, 2021), the collected corpora are restricted to limited popular categories of hate targets. Finally, the existing literature typically investigated classical machine learning and deep learning models, while the transfer learning approach has not been adequately tested and evaluated despite its acknowledged effectiveness in low-resource languages.

Our work is an attempt to characterize the Algerian anti-refugee/migrant discourse on YouTube social media and to formulate an effective evaluation of transfer learning approach for Arabic hate speech detection with taking into consideration three important perspectives: language with low resource (i.e. dialectal Algerian Arabic), target of hate speech (i.e. refugees and migrants) in addition to the transfer learning approach (i.e. monolingual and multilingual pre-trained models).

In particular, this study aims to address the following research questions:

RQ1. Is there a hate discourse against African refugees and illegal migrants on the YouTube Algerian space?

RQ2. How prevalent is this discourse on the YouTube Algerian space?

RQ3. How dangerous is this discourse on the YouTube Algerian space?

RQ4. On what basis, Algerians accept or refuse the presence of African refugees and illegal migrants in their country?

RQ5. What is the dynamic of the anti-refugee/migrant speech on the YouTube Algerian space?

RQ6. Which pre-trained models are the best for Arabic hate speech classification, monolingual or multilingual models?

RQ7. What is the impact of the lingual variation on the multilingual transformer models?

RQ8. Assuming that the Arabizi variant of Arabic may be considered as noise, what is its impact on the performance of the language transformers?

4. Methodology

4.1 Data collection

Using YouTube API, we collected user-generated comments that were publicly available on YouTube platform in which the subject was "African refugees in Algeria" or "African migrants in Algeria" during the period (2014–2021). The extraction of data included mainstream news outlets channels as well as personal channels that have been created by regular users. Since we focused on collecting large amounts of data, we only considered videos with a reasonable number of user-generated comments.

Initially, as shown in Table 1, we identified a playlist of 38 videos. Among 18 channels, 05 channels are maintained by national news agencies in Algeria, 05 are affiliated with international outlets and 08 channels are affiliated with other users.

The retrieved comments have undergone multiple data cleaning functions in order to remove the out-of-subject comments and to filter out the non-textual, non-informative comments, and duplicated instances. The final dataset includes *4,681* comments that are related to the African refugee/migrant debate on the Algerian YouTube space. A summarized description of data is listed in Table 2.

Regarding the language used by Algerian commenters, the collected comments are distributed as follows:

(1) 3,806 comments in Arabic. This includes mainly the Algerian dialectal Arabic with a negligible fraction of MSA.

	Video author (v.author)	Category	# Videos	# Comments
National news agencies	EL BILAD TV	News and politics	02	*1,102*
	Ennahar TV	News and politics	09	651
	Numidia TV	Entertainment	01	129
	Dzaïr TV	News and politics	02	48
	Beur TV	News and politics	01	35
	EL BILAD TV Officiel	People and blogs	01	145
International news agencies	France 24 Arabic	News and politics	04	*331*
	Medi1TV	News and politics	05	264
	Al Araby TV	People and blogs	02	52
	Alghad TV	News and politics	01	14
	Sky News Arabic	News and politics	01	16
Usual users	User channel 01	People and blogs	01	*389*
	User channel 02	People and blogs	01	356
	User channel 03	People and blogs	01	323
	User channel 04	News and politics	01	237
	User channel 05	People and blogs	02	206
	User channel 06	News and politics	01	142
	User channel 07	Comedy	01	139
	User channel 08	News and politics	01	102
		Total	*38*	*4,681*

Table 1.
Data sources

(2) 434 comments in Algerian Arabizi. The Arabizi is a no-rules writing style that transliterates Arabic with Latin script and numerals to write Arabic text.

(3) 338 comments in French.

(4) 95 comments in Arabic-French mixed language.

(5) Only 8 comments in English.

Figure 1 shows the language distribution of the data.

4.2 Data preprocessing and annotation

In order to prepare data for annotation, we firstly eliminated symbols from the collected comments, emojis, hashtags, digits and URLs. Further preprocessing is applied to clean text from noise by converting non-Arabic text to lower case, normalizing Arabic letters, removing Arabic diacritics, punctuation marks and stop words.

The data have been manually labelled with the help of three annotators who were recruited for this purpose. A comment was classified exclusively within one of the following categories:

(1) *Incitement (I)*: a comment that incites and encourages violence against refugees and migrants.

Dataset	RED (antiRefugee/Emigrant dataset)
Number of instances	4,681
Number of attributes (columns)	16
Names of columns	videoid, v.title, v.published, v.description, video_url, v.duration, v.dislikes, v.likes, v.rating, v.category, v.author, author, text, repcount, commentlike, label
Data types	• *Textual* (videoid, v.title, v.description, video_url, v.category, v.author, author, text, label) • *Numerical* (v.duration, v.dislikes, v.likes, v.rating, repcount, commentlike) • *Date* (v.published)

Table 2.
Description of the dataset

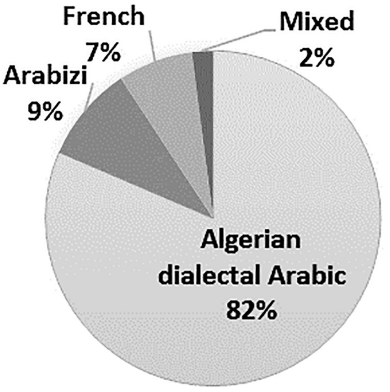

Figure 1.
Language distribution

(2) *Hate (H)*: a comment that expresses a rejectionist attitude towards refugees and migrants using abusive language that diminishes, humiliates, insults or dehumanizes the hate targets.

(3) *Refusing with non-hateful words (RNH)*: a comment that expresses a rejectionist attitude towards refugees and migrants without using abusive language.

(4) *Sympathetic (S)*: a comment that expresses a stance of "acceptance" or just "empathy" towards refugees and migrants.

(5) *Comment (C)*: a non-hateful, non-sympathetic comment, which does not express clearly the position of the commenter with regard to the subject of refugees and migrants.

The annotated dataset has a total of 1,407 comments in the "H" class, 520 in the "RNH" class, 979 in the "S" class, 1,722 in the "C" class and 53 in the "I" class.

The percentage distribution of different labels is illustrated in Figure 2.

4.3 Data analysis

In this subsection, we analyze our dataset using descriptive statistics in order to answer our first five questions:

RQ1. Is there a hate discourse against African refugees and illegal migrants on the YouTube Algerian space?

RQ2. How prevalent is this discourse on the YouTube Algerian space?

RQ3. How dangerous is this discourse on the YouTube Algerian space?

The pie chart in Figure 2 shows that there is indeed a hate speech targeting African refugees and illegal migrants in the Algerian space of YouTube social platform. However, the number of users who adopted this speech does not exceed 31%. This indicates that it is not the dominant discourse regarding the topic of the African migrants' crisis in Algeria. Moreover, its risk is very low as we noted only 1% of comments that explicitly incite violence against refugees and migrants making this violent language ineffective in reality.

Our observations are consistent with previous findings (Siegel, 2020) which have highlighted that the tendency to characterize social media platforms as dominated by hate speech is misleading and potentially problematic. The main reason for this inconvenience is

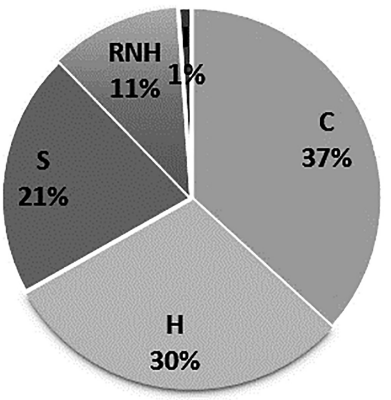

Figure 2.
Label distribution

that hateful content is more visible than other forms of content despite being limitedly produced (Saha *et al.*, 2019).

RQ4. On what basis, Algerians accept or refuse the presence of African refugees and illegal migrants in their country?

The tables of term frequency (Tables 3 and 4) show that Algerians sympathize with African migrants mostly from a religious background. We have noticed the common use of words such as God, mercy, Muslims and phrases such as "the land of God", and supplications such as "may God release from you", "be merciful to those on the Earth, and the One in the heavens will have mercy upon you" and "may God help you". This is not surprising since the sub-Saharan Africans are often affiliated with Islam, and this confirms once again the fact that the religious and cultural identity of refugees is usually exploited to arouse sympathy with them or to justify their rejection.

Top 10 words		Top 10 H		Top 10 S	
Word	Freq	Word	Freq	Word	Freq
Algeria الجزائر	704	Algeria الجزائر	302	God ربي	265
Africans الافارقة	467	Africans الافارقة	218	Earth ارض	127
God ربي	435	Our country بلادنا	122	Muslims مسلمين	97
Our country بلادنا	226	The country البلاد	106	Poors مساكن	93
Morocco المغرب	219	People الشعب	99	Africans الافارقة	84
People الشعب	199	State الدولة	88	Release يفرج	83
State الدولة	180	God ربي	82	Algeria الجزائر	81
France فرنسا	168	Threat خطر	81	Welcome مرحبا	53
Country البلاد	164	Morocco المغرب	65	They live يعيشو	38
Algerians الجزائريين	129	Algerians الجزائريين	59	The good الخير	35

Table 3. Top frequent words in the whole data, hateful data (H) and sympathetic data (S)

Top 10 words		Top 10 H		Top 10 S	
Bigrams	Freq	Bigrams	Freq	Bigrams	Freq
ربي يفرج God release (from you)	59	الشرق الأوسط Middle East	31	ربي يفرج God release (from you)	57
حقوق الانسان human rights	37	المجتمع الجزائري Algerian society	19	ارض ربي God's land	26
ارض ربي God's land	33	فوات الأوان It's too late	19	الارض يرحمكم Earth have mercy on you	21
الشعب الجزائري The Algerian people	31	شمال افريقيا North Africa	18	يرحمكم السماء Heaven have mercy on you	18
الشرق الأوسط Middle East	31	حقوق الانسان Human rights	17	مساكن ربي Poors God	14
شمال افريقيا North Africa	27	الشعب الجزائري The Algerian people	15	ارض واسعة Wide land	13
الارض يرحمكم Earth have mercy on you	22	اقرب وقت Shortest time	14	مرحبا بيكم You're Welcome	13
ربي يستر May God keep us	21	بلدان المغرب Maghreb countries	13	ربي يرزقهم May God help them	11
اقرب وقت Shortest time	20	الدولة الجزائرية Algerian state	12	ارحمو الارض Have mercy Earth	9
المجتمع الجزائري Algerian society	20	الافارقة الغزاة African invaders	12	لا فرق No difference	7

Table 4. Top frequent bi-grams in the whole data, hateful data (H) and sympathetic data (S)

On the other hand, the hateful comments show the concerns of the rejectionists about Algerian cultural identity and demography, security and economic consequences, and question the government's interest in the illegal migration issue. Besides, it is apparent from Table 5 that the haters in comments tend to carefully explain their stances towards the African migration crisis to justify it, to convince the opposants or to exacerbate more attention on the migration subject.

RQ5. What is the dynamic of the anti-refugee/migrant speech on the YouTube Algerian space?

As for the dynamic of the African migrants' topic on YouTube, data from Figure 3 reveal that it has mostly gained wide interaction in 2017. This is probably due to the anonymous hashtag #No_to_Africans_in_Algeria, which has generated substantial discussion and debates around the migration crisis on social media at the time (Chikhi, 2017). This hashtag was responded to with a counter-hashtag rejecting the policy of expelling refugees and calling for their assistance and protection (Attia, 2017). Furthermore, Table 6 illustrates that the interest

Label	Text length
C	14.17
H	*28.85*
I	19.40
RNH	*20.48*
S	19.04

Table 5.
Average text length for comment labels

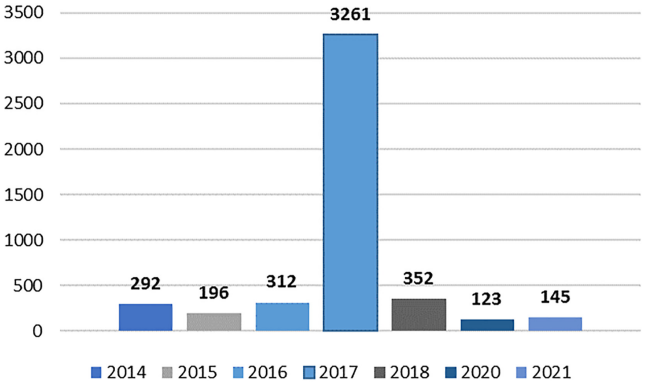

Figure 3.
Comment distribution

Year	C	H	I	RNH	S
2014	101	*103*	8	39	41
2015	*100*	53	/	34	9
2016	*150*	92	2	34	34
2017	*1,033*	*1,033*	*41*	*358*	*796*
2018	*159*	89	2	34	68
2020	*100*	8	/	8	7
2021	*79*	29	/	13	24

Table 6.
Label distribution per year

in this subject before and after 2017 seems weak, yet the distribution of users' stances towards African migrants maintains its stability, so that the category "C" leads most of the comments, followed by, respectively, "H", "RNH", "S" and "I" categories.

4.4 Hate speech detection

These experiments aim to examine the following three research questions:

RQ6. Which pre-trained models are the best for Arabic hate speech classification, monolingual or multilingual models?

RQ7. What is the impact of the lingual variation on the multilingual transformer models?

RQ8. Assuming that the Arabizi variant of Arabic may be considered as noise, what is its impact on the performance of the language transformers?

To answer these questions, we used the following five dataset combinations from the initial collected dataset RED, i.e. the anti*R*efugee/*E*migrant *D*ataset:

(1) *RED_All_Data*: consists of all the collected comments in four lingual variations: Algerian dialectal Arabic, Algerian Arabizi, French and English.

(2) *RED_All_No_Arabizi*: in this combination, we excluded the Arabizi comments from the dataset.

(3) *RED_All_Trans_Arabizi*: in this combination, the Arabizi texts were converted to Arabic script using the Buckwalter transliteration method. The combination consists of Arabic texts (i.e. Algerian dialectal Arabic and converted Arabizi) and non-Arabic texts (i.e. French and English).

(4) *RED_Arabic_TransArabizi*: in this combination, we kept only the collected comments which are in Algerian dialectal Arabic and the Arabizi texts that were converted to Arabic script.

(5) *RED_Arabic_only*: in this combination, we kept only the collected comments which are in Algerian dialectal Arabic.

For all datasets, we performed *k*-fold standard cross-validation to sample training set and test set. The number of folds was set to 05. Table 7 shows the description of all the dataset combinations.

Next, we developed four fine-tuned pre-trained language models using the multilingual transformers mBERT and croX lingual Language Model-RoBERTa (XLM-R) as well as the monolingual transformers AraBERT and DziriBERT. The experiments were run on Google Colab to use GPU processors for fast execution. The parameters are set to the recommended values mentioned in the original BERT paper (Devlin *et al.*, 2018). Especially, for all the three models, we chose 5e-5 as the learning rate, 256 as the max length of texts, 16 for the batch size and 8 for the number of epochs.

Dataset	Total instances
RED_All_Data	4,681
RED_All_No_Arabizi	4,247
RED_All_Trans_Arabizi	4,681
RED_Arabic_TransArabizi	4,240
RED_Arabic_only	3,806

Table 7.
Dataset combinations' description

Below, we describe the applied models:

(1) *mBERT*: BERT, which stands for *B*idirectional *E*ncoder *R*epresentations from *T*ransformers, is a deep architecture that relies on the concept of bidirectionality and unsupervised language representation. BERT pre-trains deep bidirectional representations from unlabelled text from Wikipedia by jointly conditioning on both left and right context in all layers (Devlin *et al.*, 2018). Without requiring significant changes on the principal architecture, the pre-trained BERT model may be fine-tuned with just one additional output layer to provide state-of-the-art models for a variety of NLP tasks, such as question answering and sequence classification (Devlin *et al.*, 2018).

mBERT, on the other hand, is a BERT model that has been pre-trained on 104 languages with two objectives: Masked language modeling and next sentence prediction. It was specifically trained on a large corpus of Wikipedia articles with a vocabulary that was shared across all languages. This model is primarily aimed at being fine-tuned on tasks such as sequence classification, token classification or question answering.

(2) *XLM-R*: It is a transformer-based masked language model that has been trained on 100 languages, using more than 2TB of filtered Common Crawl data (Conneau *et al.*, 2019). It was introduced by Facebook AI researchers in November 2019 achieving state-of-the-art results in cross-lingual transfer learning. *XLM-R* significantly outperformed mBERT on a variety of cross-lingual benchmarks such as classification, sequence labelling and question answering (Conneau *et al.*, 2019). XLM-R employs an approach of Translation Language Modelling and proves a good performance particularly on low-resource languages.

(3) *AraBERT*: It is a monolingual pre-trained BERT model for the Arabic language, namely MSA (Antoun *et al.*, 2020). It has been trained on approximately 24 GB of Arabic text and applied on three NLP tasks: sentiment analysis, named entity recognition and question answering. AraBERT achieves state-of-the-art performance compared to several baselines including mBERT from Google and other single-language approaches.

(4) *DziriBERT*: It is a new bidirectional transformer for the Algerian Arabic dialect (Abdaoui *et al.*, 2021). It has been pre-trained on over 1 million tweets, which represents 20 million tokens and 150 MB of data size. The monodialectal language model DziriBERT has been evaluated on two text classification tasks achieving new state-of-the-art results in sentiment analysis and emotion recognition in the Algerian dialectal Arabic.

5. Results and discussion

The results obtained from experimenting with the pre-trained language models are summarized in Tables 8–11. We employed six evaluation metrics: precision, recall, F-measure, Cohen's kappa, specificity and accuracy. Underlined values indicate the best performance of the specified model across the five data combinations. Italic values indicate that the specified model outperforms the other models on the specified data combination.

For monolingual data combinations (RED_Arabic_TransArabizi and RED_Arabic_only), it is apparent from Table 10 that the monolingual transformer AraBERT noticeably beats the multilingual models mBERT and XLM-R in terms of all measures. Correspondingly,

Tables 10 and 11 report comparable performance of AraBERT and DziriBERT on the two data combinations of dialectal Arabic texts in terms of accuracy, specificity and Cohen-kappa. However, excluding the transliterated Arabizi from the data seems slightly boosting the performance of AraBERT and DziriBERT in terms of all measures except precision. These findings corroborate the earlier study of Antoun *et al.* (2020) that highlighted the good performance of the Arabic monolingual BERT compared to mBERT even when dealing with the dialectal Arabic. XLM-R is involved with this observation since its architecture is mainly based on BERT architecture. The authors of AraBERT (Antoun *et al.*, 2020) suggested that the good performance of their monolingual model is due to the high size of the pre-training data in addition to the vocabulary size (64 k of vocab size) exploited for its development. We also suppose that the segmentation (tokenization) preprocessing function of AraBERT that takes into consideration structural aspects of the Arabic language, played a role in avoiding word redundancy and thus improving the quality of the dialectal Arabic data. Likewise, this might be the reason for the superiority of AraBERT over DziriBERT in classifying Algerian

Model	Data	Prec	Rec	F1-score	Acc	Spec	Kappa
mBERT (bert-base-multilingual-cased[a])	RED_All_Data	*0.49*	*0.47*	*0.47*	*0.59*	*0.89*	*0.41*
	RED_All_Trans_Arabizi	*0.51*	*0.46*	*0.46*	*0.60*	*0.89*	*0.43*
	RED_All_No_Arabizi	*0.53*	*0.47*	*0.46*	*0.60*	*0.89*	*0.42*
	RED_Arabic_TransArabizi	0.48	0.47	0.47	0.61	0.89	0.44
	RED_Arabic_only	0.51	0.49	0.49	0.63	0.90	0.47

Note(s): [a]https://huggingface.co/bert-base-multilingual-cased

Table 8. Classification results for the *mBERT* multilingual transformer

Model	Data	Prec	Rec	F1-score	Acc	Spec	Kappa
XLM-R (xlm-roberta-base[a])	RED_All_Data	0.37	0.38	0.36	0.55	0.87	0.34
	RED_All_Trans_Arabizi	0.38	0.39	0.38	0.55	0.87	0.35
	RED_All_No_Arabizi	0.36	0.39	0.37	0.58	0.88	0.37
	RED_Arabic_TransArabizi	0.38	0.40	0.38	0.56	0.87	0.36
	RED_Arabic_only	0.45	0.46	0.45	0.63	0.90	0.47

Note(s): [a]https://huggingface.co/xlm-roberta-base

Table 9. Classification results for the *XLM-R* multilingual transformer

Model	Data	Prec	Rec	F1-score	Acc	Spec	Kappa
AraBERT (aubmindlab/ bert-base-arabertv02[a])	RED_Arabic_TransArabizi	*0.62*	*0.55*	*0.56*	*0.67*	*0.91*	*0.53*
	RED_Arabic_only	*0.59*	*0.56*	*0.57*	*0.68*	*0.91*	*0.54*

Note(s): [a]https://huggingface.co/aubmindlab/bert-base-arabertv02

Table 10. Classification results for the *AraBERT* monolingual transformer

Model	Data	Prec	Rec	F1-score	Acc	Spec	Kappa
DziriBERT (alger–ia[a])	RED_Arabic_TransArabizi	0.59	0.52	0.53	*0.67*	*0.91*	0.52
	RED_Arabic_only	0.56	0.53	0.53	*0.68*	*0.91*	*0.54*

Note(s): [a]https://huggingface.co/alger-ia/dziribert

Table 11. Classification results for the *DziriBERT* monodialectal transformer

dialectal Arabic texts, although the latter model was specifically developed for this dialect. Note that the vocabulary of the Algerian dialectal Arabic, which has several structural specificities, is mostly inspired by standard Arabic.

Comparing the multilingual models, Tables 8 and 9 reveal that mBERT achieves a statistically significant improvement compared to XLM-R on almost all data combinations. Nevertheless, the performance of XLM-R has been boosted when XLM-R was applied solely on the dialectal Arabic texts (RED_Arabic_only). The reason for this requires further investigation, yet, we speculate that the volume and the quality of data have an effect on it. That is, mBERT and XLM-R models can be affected positively by the good quality and the small quantity of data. These findings however need to be treated with caution as recent studies showed that the performance of the pre-trained models like BERT slightly decreases when the number of instances in the training data is less than approximately 5 K (Edwards *et al.*, 2020).

What is interesting in the results is that the performance of the multilingual models mBERT and XLM-R has been improved when the Arabizi texts were excluded from the multilingual data combinations or transliterated with Arabic script: mBERT and XLM-R achieve their best performances when the data have consisted exclusively of Arabic texts.

Though transliterating Arabizi has enhanced the performance of the transformers mBERT and XLM-R, we should sound a note of caution with regard to such observation as we noted a decrease in the performance of the monolingual AraBERT by adding the transliterated Arabizi to the dialectal Arabic texts. In the same way, the transliterated Arabizi appears to not significantly improve the performance of the monodialectal model DziriBERT. Actually, the Buckwalter transliteration method usually fails in providing understandable texts owing to the phonetic richness of Arabic and the difficulty of automatic transliteration without understanding the meaning of the words.

It is worth mentioning that the performance of the four models in these experiments is not ideal. This discrepancy could be attributed to the use of recommended parameter settings, which may not be appropriate for the datasets we experimented with.

Together, these results provide important insights into the limitation of the multilingual pre-trained models in dealing with low-resource languages such as Arabic in comparison with Latin-based languages in which they note good performance (Pires *et al.*, 2019). Similarly, while the performance of the AraBERT monolingual pre-trained model can be affected by the dialectal forms of the Arabic language, its results remain better than the monodialectal model DziriBERT. Meanwhile, the findings support our hypothesis on the negative impact of the Arabizi variant on the effectiveness of the multilingual and monolingual transformers for Arabic. A further study with more focus on these drawbacks to improve Arabic hate speech detection is therefore suggested.

6. Conclusion

The expansion of hate and racial online content that targets people on the basis of their sociocultural differences and identities have steered attention to the power of social media such as YouTube to shape the public opinion towards such unacceptable phenomena and to exacerbate its visibility.

Our aim in this paper was to characterize the properties and the pervasiveness of the online content that targets the African refugees and illegal migrants in Algeria as well as to evaluate the transfer learning language models which recently present the state-of-the-art approach in NLP tasks like text classification.

The results of this investigation show that hate speech against refugees and migrants has existed modestly since 2014. However, it only arose in 2017 when African illegal migrants

were targeted with an anonymous rejectionist hashtag on social media platforms. According to the analysis, the seriousness of this speech is weak and ineffective. This is evidenced by the sharp decline that ended with this discourse to moderate levels after 2017, when it was faced with an online counter-campaign and a wave of public criticism and sympathy that was mainly based on religious background.

On the other hand, classifying hate speech following the transfer learning approach has revealed that the monolingual transformer AraBERT outperforms both the monodialectal transformer DziriBERT and the multilingual transformers mBERT and XLM-R. Additionally, the Arabizi variant emerges as a noise that obviously affects the efficiency of the pre-trained language models. In general, therefore, it appears that the transfer learning models still need further ameliorations to reach optimal performance in classifying texts of low-resource languages such as Arabic and mainly its dialectal variants as the Algerian dialectal Arabic.

This study is subject to certain limitations. For instance, investigating the source channels that promote hatred against African refugees and illegal migrants on YouTube was not addressed. Likewise, more work is required to analyze the correlation between the published content on YouTube and the distribution of hateful comments in the corresponding comments sections. On the other hand, it is important to bear in mind that these findings have not yet undertaken statistical significance tests due to the lack of reliable standards for statistical significance analysis and generally meta-analysis in NLP. Furthermore, considerably more work will need to be done to enhance the quality of the data for the existing pre-trained language models or to retrain them so that they take into account the shortcomings related to the complexities of the Arabic language and its different dialectal variants.

References

Abdaoui, A., Berrimi, M., Oussalah, M. and Moussaoui, A. (2021), "DziriBERT: a pre-trained language model for the Algerian dialect", arXiv preprint, arXiv:2109.12346.

Agarwal, S. and Sureka, A. (2017), "Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr micro-blogging website", arXiv preprint, arXiv: 1701.04931.

Al-Hassan, A. and Al-Dossari, H. (2021), "Detection of hate speech in Arabic tweets using deep learning", *Multimedia Systems*, pp. 1-12, doi: 10.1007/s00530-020-00742-w.

Alatawi, H.S., Alhothali, A.M. and Moria, K.M. (2021), "Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT", *IEEE Access*, Vol. 9, pp. 106363-106374, doi: 10.1109/ACCESS.2021.3100435.

Alorainy, W., Burnap, P., Liu, H. and Williams, M. (2018), "Cyber hate classification: 'othering' language and paragraph embedding", arXiv preprint, arXiv:1801.07495.

Alshalan, R. and Al-Khalifa, H. (2020), "A deep learning approach for automatic hate speech detection in the Saudi Twittersphere", *Applied Sciences*, Vol. 10 No. 23, p. 8614.

Antoun, W., Baly, F. and Hajj, H. (2020), "Arabert: transformer-based model for Arabic language understanding", arXiv preprint, arXiv:2003.00104.

Attia, S. (2017), *Algérie : pourquoi la situation des migrants subsahariens est-elle si problématique*, Jeune Afrique, 20 July, available at: https://www.jeuneafrique.com/459154/societe/situation-migrants-subsahariens-algerie-problematique/ (accessed 01 October 2021).

Bigoulaeva, I., Hangya, V. and Fraser, A. (2021), "Cross-lingual transfer learning for hate speech detection", *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 15-25.

Burnap, P. and Williams, M.L. (2015), "Cyber hate speech on Twitter: an application of machine classification and statistical modeling for policy and decision making", *Policy and Internet*, Vol. 7 No. 2, pp. 223-242.

Caselli, T., Basile, V., Mitrović, J. and Granitzer, M. (2020), "Hatebert: retraining BERT for abusive language detection in English", arXiv preprint, arXiv:2010.12472.

Chikhi, L. (2017), "Algeria to grant legal status to African migrants amid worker shortages, racism", Reuters, 03 July, available at: https://www.reuters.com/article/us-algeria-economy-migrants-idUSKBN19O23G (accessed 01 October 2021).

Clayton, J. (2021), "Trump sues Twitter, Google and Facebook alleging 'censorship'", BBC News, 07 July, available at: https://www.bbc.com/news/world-us-canada-57754435 (accessed 01 October 2021).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2019), "Unsupervised cross-lingual representation learning at scale", arXiv preprint, arXiv:1911.02116.

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017), "Automated hate speech detection and the problem of offensive language", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11, No. 1 May 2017.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018), "Bert: pre-training of deep bidirectional transformers for language understanding", arXiv preprint, arXiv:1810.04805.

Dinakar, K., Reichart, R. and Lieberman, H. (2011), "Modeling the detection of textual cyberbullying", in *Fifth International AAAI Conference on Weblogs and Social Media*.

Dwoskin, E. and De Vynck, G. (2021), "Facebook's AI treats palestinian activists like it treats American Black activists. It blocks them", The Washington Post, 28 May, available at: https://www.washingtonpost.com/technology/2021/05/28/facebook-palestinian-censorship/ (accessed 01 October 2021).

Edwards, A., Camacho-Collados, J., De Ribaupierre, H. and Preece, A. (2020), "Go simple and pre-train on domain-specific corpora: on the role of training data for text classification", *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5522-5529.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W.Y. and Belding, E. (2018), "Hate lingo: a target-based linguistic analysis of hate speech in social media", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12 No. 1.

Fortuna, P. and Nunes, S. (2018), "A survey on automatic detection of hate speech in text", *ACM Computing Surveys (CSUR)*, Vol. 51 No. 4, pp. 1-30.

Gagliardone, I., Gal, D., Alves, T. and Martinez, G. (2015), *Countering Online Hate Speech"*, Unesco Publishing.

Gitari, N.D., Zuping, Z., Damien, H. and Long, J. (2015), "A lexicon-based approach for hate speech detection", *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 10 No. 4, pp. 215-230.

Greevy, E. and Smeaton, A.F. (2004), "Classifying racist texts using a support vector machine", *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 468-469.

Haddad, B., Orabe, Z., Al-Abood, A. and Ghneim, N. (2020), "Arabic offensive language detection with attention-based deep neural networks", *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pp. 76-81.

Hafid, T. (2019), "L'Algérie reçoit quotidiennement plus de migrants que toute l'Europe", Le Soir d'Algérie, 18 November, available at: https://www.lesoirdalgerie.com/entretien/lalgerie-recoit-quotidiennement-plus-de-migrants-que-toute-leurope-33790 (accessed 01 October 2021).

Hewitt, S., Tiropanis, T. and Bokhove, C. (2016), "The problem of identifying misogynist language on twitter (and other online social spaces)", *Proceedings of the 8th ACM Conference on Web Science*, pp. 333-335.

Himmel, R. and Baptista, M.M. (2020), "Migrants, refugees and othering: constructing Europeanness, an exploration of Portuguese and German media", *Comunicação e sociedade*, Vol. 38, pp. 179-200.

Köffer, S., Riehle, D.M., Höhenberger, S. and Becker, J. (2018), *Discussing the Value of Automatic Hate Speech Detection in Online Debates", Multikonferenz Wirtschaftsinformatik (MKWI 2018): Data Driven X-Turning Data in Value, Leuphana 2018*, Germany.

Liu, S. and Forss, T. (2015), "New classification models for detecting hate and violence web content", *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), IEEE*, Vol. 1, pp. 487-495.

Mishra, P., Yannakoudakis, H. and Shutova, E. (2018), "Neural character-based composition models for abuse detection", arXiv preprint, arXiv:1809.00378.

Mulki, H. and Ghanem, B. (2021), "Let-Mi: an Arabic Levantine twitter dataset for misogynistic language", arXiv preprint, arXiv:2103.10195.

Musette, M.S. (2014), *Algérie: les facettes de la migration pour cause de crises*, Forced Migration Review (FMR), Refugee Studies Centre in the Oxford Department of International Development, University of Oxford.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y. and Yeung, D.Y. (2019), "Multilingual and multi-aspect hate speech analysis", arXiv preprint, arXiv:1908.11049.

Özerim, M.G. and Tolay, J. (2021), "Discussing the populist features of anti-refugee discourses on social media: an anti-Syrian Hashtag in Turkish Twitter", *Journal of Refugee Studies*, Vol. 34 No. 1, pp. 204-218.

Perifanos, K. and Goutsos, D. (2021), "Multimodal hate speech detection in Greek social media", *Multimodal Technologies and Interaction*, Vol. 5 No. 7, p. 34, doi: 10.3390/mti5070034.

Pires, T., Schlinger, E. and Garrette, D. (2019), "How multilingual is multilingual BERT?", arXiv preprint, arXiv:1906.01502.

Powell, J. (2017), "Us vs them: the sinister techniques of 'Othering' – and how to avoid them", *The Guardian*, 8 November, available at: https://www.theguardian.com/inequality/2017/nov/08/us-vs-them-the-sinister-techniques-of-othering-and-how-to-avoid-them (accessed 01 October 2021).

Reidpath, D.D. and Allotey, P. (2018), in *The Health Of Refugees: Public Health Perspectives from Crisis to Settlement*, *Social Exclusion, Othering, and Refugee Health Policy*, p. 39.

Rizoiu, M.A., Wang, T., Ferraro, G. and Suominen, H. (2019), "Transfer learning for hate speech detection in social media", arXiv preprint, arXiv:1906.03829.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N. and Wojatzki, M. (2017), "Measuring the reliability of hate speech annotations: the case of the European refugee crisis", arXiv preprint, arXiv:1701.08118.

Saha, K., Chandrasekharan, E. and De Choudhury, M. (2019), "Prevalence and psychological effects of hateful speech in online college communities", *Proceedings of the 10th ACM Conference on Web Science*, pp. 255-264.

Siegel, A.A. (2020), "Online hate speech. social media and democracy: the state of the field", *Prospects for Reform*, pp. 56-88.

Siegel, A., Tucker, J., Nagler, J. and Bonneau, R. (2017), "Socially mediated sectarianism", Unpublished Manuscript, available at: http://alexandra-siegel.com/wp-content/uploads/2017/08/Siegel_Sectarianism_January2017.pdf.

UNHCR (2016), "UNHCR viewpoint: 'Refugee' or 'migrant' – which is right", available at: https://www.unhcr.org/news/latest/2016/7/55df0e556/unhcr-viewpoint-refugee-migrant-right.html (accessed 01 October 2021).

Warner, W. and Hirschberg, J. (2012), "Detecting hate speech on the world wide web", *Proceedings of the Second Workshop on Language in Social Media*, pp. 19-26.

Waseem, Z. and Hovy, D. (2016), "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter", *Proceedings of the NAACL Student Research Workshop*, pp. 88-93.

Zahiri, S.M. and Ahmadvand, A. (2020), "CRAB: class representation attentive BERT for hate speech identification in social media", arXiv preprint, arXiv:2010.13028.

Zhang, Z., Robinson, D. and Tepper, J. (2018), "Detecting hate speech on twitter using a convolution-GRU based deep neural network", *European Semantic Web Conference*, Cham, Springer, pp. 745-760.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q. (2020), "A comprehensive survey on transfer learning", *Proceedings of the IEEE*, Vol. 109 No. 1, pp. 43-76.

Corresponding author
Djamila Mohdeb can be contacted at: djamila.mohdeb@univ-bba.dz