Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6203–6212
Marseille, 11–16 May 2020

A Multi-Platform Arabic News Comment Dataset for Offensive Language Detection

Shammur A Chowdhury, Hamdy Mubarak, Ahmed Abdelali,
Soon-gyo Jung, Bernard J Jansen, Joni Salminen
Qatar Computing Research Institute
Hamad Bin Khalifa University, Doha, Qatar
{shchowdhury, hmubarak, aabdelali, sjung, bjansen, jsalminen}@hbku.edu.qa

Abstract

Access to social media often enables users to engage in conversation with limited accountability. This allows a user to share their opinions and ideology, especially regarding public content, occasionally adopting offensive language. This may encourage hate crimes or cause mental harm to targeted individuals or groups. Hence, it is important to detect offensive comments in social media platforms. Typically, most studies focus on offensive commenting in one platform only, even though the problem of offensive language is observed across multiple platforms. Therefore, in this paper, we introduce and make publicly available a new dialectal Arabic news comment dataset, collected from multiple social media platforms, including Twitter, Facebook, and YouTube. We follow two-step crowd-annotator selection criteria for low-representative language annotation task in a crowdsourcing platform. Furthermore, we analyze the distinctive lexical content along with the use of emojis in offensive comments. We train and evaluate the classifiers using the annotated multi-platform dataset along with other publicly available data. Our results highlight the importance of multiple platform dataset for (a) cross-platform, (b) cross-domain, and (c) cross-dialect generalization of classifier performance.

Keywords: Offensive Language, Arabic, Text Classification, Social Media Platforms, Twitter, Facebook, YouTube

1. Introduction

Social media platforms provide access for users from all over the world to connect with each other, share their knowledge, and express their opinions (Intapong et al., 2017). Within these platforms, people not only voice their opinions and concerns but also look for social connection, belongingness, and assurance of being accepted by society. Thus, any negative perceived image of a group or an individual, by a small fraction of the community, can impact users' psychological well-being (Gülaçtı, 2010; Waldron, 2012), e.g., by creating propaganda and giving rise to hate crimes.

Often, this type of abusive propaganda spreads through either by the content posted in the platforms or by the toxic behavior of users in the social media. Users may use offensive post/comments containing vulgar, pornographic and/or hateful language, to spread such verbal hostility. Hence, the increased risk and effect of such hostility using offensive language, in social media, has attracted many multidisciplinary researchers and provoked the need of automatically detecting the offensiveness of post/comments.

Various classification techniques have been studied and used to detect offensive language (Davidson et al., 2017; Mubarak and Darwish, 2019), along with related categories of hate speech (Badjatiya et al., 2017; Salminen et al., 2018), cyberbullying (Dadvar et al., 2013; Hosseinmardi et al., 2015), aggression (Kumar et al., 2018) among others. Modeling techniques such as keyword-based search, traditional machine learning to deep learning have been explored.

However, most of the previous studies are limited to Indo-European languages due to the availability of resources in these languages. Unlike these resource-rich languages, studies and resources for detecting offensive language in dialectal or Modern Standard Arabic (MSA) are still very limited. Similar to the Indo-European languages, most of the publicly available datasets for Arabic (Mubarak and Darwish, 2019; Mulki et al., 2019; Alakrot et al., 2018; Al-Ajlan and Ykhlef, 2018; Mubarak et al., 2017) originate mainly from one social media platform: either Twitter (TW) or YouTube (YT).

However, the challenges of detecting offensive language are not constrained to one or two platforms but have a cross-platform nature (Salminen et al., 2020). Therefore, research efforts are needed to develop rich resources that can be used to design and evaluate cross-platform offensive language classifiers.

In this study, we introduce one of the first Dialectal Arabic (DA) offensive language datasets, extracted from three different social media platforms: TW, YT, and Facebook (FB). Our annotated dataset comprises a collection of 4000 comments posted in social media platforms. The domain of the dataset is focused on the news comments from an international news organization targeting audiences from all over the world. The rationale for this choice of domain is that news content posted in the social media platforms attracts active interaction between the platform users and the content itself is directly not offensive – thus any offensiveness in their respective comments are the product of users' opinion and belief.

In addition, the study also addresses the difficulties and possible approaches taken to build such a versatile resource for DA dataset. Building a reliable dataset with high-quality annotation is necessary for designing accurate classifiers. Specifically, in this study, we highlight the need and quantify the criteria of annotator and the annotation selection while using crowdsourcing platforms for a less represented language (Difallah et al., 2018; Ross et al., 2009). We thoroughly evaluated the annotated data using (a) inter-annotator agreement between the accepted annotators, and

(b) measuring accuracy between the crowdsourced annotation with expert annotation.

To understand and to explore the generalizability of the introduced dataset, we present a series of classification experiments using Support Vector Machines (SVM).

These experiments include studying the performance of the trained classifier on (a) a large number of comments (both in- and out-of-domain data) (b) on cross-platform data, and (c) on particular DA (Egyptian and Levantine) data. We also investigate lexical cues and use of emojis in the offensive instances present in the dataset.

Overall, the key contributions of the paper can be summarised as follows:

- Introduction of a new Dialectal Arabic (DA) offensive language dataset from the comments of a news post, extracted from multiple social media platforms – TW, YT, and FB – of an international news agency's social media accounts. To best of our knowledge, this is one of the first multi-platform datasets for Arabic social media comments.
- Designing and quantifying a reliable two-step crowd-annotator selection criteria for low-representative language (such as DA).
- Analyzing lexical content and emoji usage in offensive language.
- Designing and evaluating classification models for:
    - in-domain and cross-domain social media data
    - cross-platform performance
    - across-dialect – Egyptian and Levantine – dataset
- Publicly releasing the dataset and listing all the resources of the study[1].

The rest of the paper is structured as follows. We provide a brief overview of the existing work and datasets for Arabic and other languages in Section 2. We then discuss the data and the annotation collection procedures in Section 3. We present a detailed analysis of the lexical and emojis present in Section 4 and Section 5 provides the results of experiments. Finally, we conclude our work in Section 6.

2. Related Studies

Recent years have witnessed a rising use of offensive language in the social media platform. This has forced many websites and organizations to remove such use of offensive content, using either manual or automatic filtering process. The current trend of using offensive language and its effect on particular individual or groups has also attracted many multi-disciplinary studies.

Current conceptualization categorizes offensive language usage, for social media, mostly as hate speech, including remarks attacking particular race, religion, nationality among others; vulgar or obscene and pornographic comments, including explicit and rude sexual references (Jay and Janschewitz, 2008). Most studies are conducted in English (Davidson et al., 2017; Silva et al., 2016; Mondal et al., 2017; Badjatiya et al., 2017; Chatzakou et al., 2017a; Chatzakou et al., 2017b; Chatzakou et al., 2017c; ElSherief et al., 2018; Unsvåg and Gambäck, 2018; Agarwal and Sureka, 2014) and German (Wiegand et al., 2018), using supervised deep learning architectures, like variants of Recurrent Neural Networks (RNN) (Pitsilis et al., 2018; Pavlopoulos et al., 2017), Convolutional Neural Networks (CNN) (Zhang et al., 2018), as well as Naive Bayes (NB) (Rish and others, 2001), SVM (Platt, 1998), and others.

Unlike Indo-European languages, a handful of research has been conducted for Arabic languages. Even though Arabic as a language is spoken by a large portion of the world's population, in practice the language is mutually unintelligible based on the lack of knowledge on dialects. These make modelling offensive language for Arabic a notable challenge, mostly due to lack of resource availability.

The authors in (Alakrot et al., 2018) used a YT comment dataset of $16k$ containing comments from Egypt, Iraq and Libya and annotated it with native speakers using ternary scheme (offensive, inoffensive and neutral) with the inter-annotator agreement of 71%. Using SVM, the authors obtained a F-measure of 82%. Unfortunately, the aforementioned dataset was not available to us while carrying out this study.

In (Mubarak et al., 2017), the authors first presented a TW collection of $(1.1k)$ contents including 100 Egyptian tweets from 10 controversial user accounts and their corresponding comments. The data was annotated by 3 annotators from Egypt with an agreement of 85%. This dataset was also used in (Mubarak and Darwish, 2019) to test a method for generating large scale training data using a seed list of offensive words.

In addition to the tweet data, the authors, in (Mubarak et al., 2017), also publicly released a large dataset of $\approx 32k$ deleted comments from Aljazeera.net, an online news portal. The comments include varieties of Arabic dialects and MSA and were annotated by 3 annotators with an agreement of 87%. For both datasets, the scheme for annotation included obscene, offensive (but not obscene), and clean.

In (Albadi et al., 2018), the authors explored hate speech detection for different religious groups such as Muslims, Jews, and others. For the study, the authors introduced a multi-dialectal dataset of $6.6k$ tweets. In addition, the authors created lexicon used commonly in religious discussion and with scores representing polarity and strength for (non)hate, by using techniques like Pointwise Mutual Information (PMI). The author also suggested that the annotator agreement varies based on which religious group is being targeted. Moreover, the study presented classification results using lexicon-based, SVM and GRU classification with pre-trained word embedding techniques and showed a performance of 77% of F-measure.

A Levantine (Syrian, Lebanese, Palestinian and Jordanian) DA dataset was presented in a recent study by the authors in (Mulki et al., 2019). In this work, the authors created a manually annotated political Twitter dataset, of size $5.8k$, for classifying hate speech, abusive and normal content with the help of native speakers. The authors also pre-

[1] Data download link: https://github.com/shammur/Arabic-Offensive-Multi-Platform-SocialMedia-Comment-Dataset

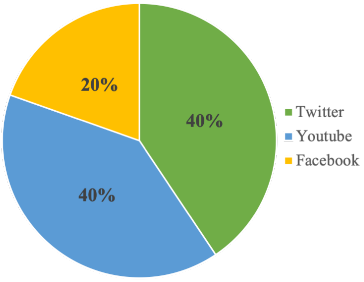

Figure 1: Platform wise distribution of the comments in the dataset.

sented lexical analysis in addition to classification results using SVM and NB algorithms.

Unlike most of the previous work, for Arabic offensive language detection, we introduce a multi-dialectal Arabic comment dataset for offensive comment detection from multiple social media platforms, including TW, YT, and FB. To the best of our knowledge, this is one of the first studies for Arabic and also one of the few datasets for offensive content detection research to include data from more than one platform. In addition to our dataset, we utilize three other datasets (Mubarak et al., 2017; Mulki et al., 2019) to study how such multi-platform data could help generalize offensive detection models. We also study the use of emojis and their discriminating characteristics for offensive language.

3. Data

3.1. Data Collection

To study how offensive language is used in online interaction for news posts and to design a classifier, we collected over $100k$ comments from different social media platforms - including Facebook, Twitter, and YouTube, for a well-reputed international news agency. We first collected all news content posted, from 2011 to 2019, by the news agency in their social media accounts. We collected the contents from each platform through their own API (YouTube, Facebook, and Twitter). Then, using each content ID, the comments for the content are collected. As Twitter does not provide the API for retrieving comments directly, we used the Standard Search API of Twitter which only provides comments for past 7 days only. To overcome this challenge, we periodically collected the comments in every 6 hours for a certain period to extract the complete comment history of the news post.

3.2. Data Preparation and Selection

For the data selection, we retain comments that includes 5 or more Arabic tokens apart from emojis, while anonymizing them by replacing any user mentions with a 'USER.IDX' tag and urls with 'URL' tags. In addition, we removed duplicate comments based on textual content. We then selected a random subset of 4000 comments, in total, from the three major social media platforms (comment distribution shown in Figure 1) for manual annotation.

3.3. Annotation Guideline

The annotation task, presented in this paper, is straightforward – *"Given a comment, categorize if the comment is offensive or not"*. Even though the task seems benign, deciding whether the comment is offensive or not is not simple. The annotators were instructed to rely on their instinct and suggested to ignore their personal biases such as political, religious belief or their cultural background. To make the decision-making process easier, we provided elaborated and detailed instructions to the annotators with examples.

For the task, we asked the annotators to consider instances as offensive, if comment contain (a) *abusive words*; (b) explicit or implicit *hostile intention* that *threats or project violence*; (c) contempt, humiliate or to underestimate groups or individuals using – *animal analogy*, *name calling*, attacking their *political ideologies* or their *disabilities*, *cursing*, *insulting religious beliefs*, *incitement racial and ethnic hatred*, or other forms of *insulting/profanity*. Details of the annotation guideline and examples presented to the annotator, in each cases, are made publicly available [2].

In addition to the detailed examples, we apologize for the presence of any offensive words that might hurt the annotators' personal beliefs and thanked them for taking into account the accuracy of the task and for their participation.

3.4. Annotations via Crowdsourcing

Given the specified task, we used Amazon Mechanical Turk (AMT)[3], a well-known crowdsourcing platform, to obtain manual annotations of these 4000 news comments (mentioned in Section 3.2). For each comment, we collected 3 judgments with a cost of 1 cent per judgment. Our pilot study for the task shows that 1 cent per comment is a reasonable amount to pay for such a small (binary) task and with higher rewards, no further performance improvement with respect to time or quality is found. Details of parameters used for the annotation task is shown in Table 1.

AMT is an efficient and cost-effective way to acquire manual annotations, however, it can possess several limitations for a task. Some of the main limitations faced for the annotation is to judge annotators language proficiency and the task understanding capabilities. To ensure the quality of the annotation and language proficiency, we utilized two different evaluation criteria of the annotator.

The first criterion, for the annotators, to qualify for attempting the task includes answering a series of multiple-choice questions (Example 1) - designed by experts - that reflects the annotators' language proficiency and understanding of the questions.

Example 1 ماذا نسمي والد الأب؟

What we call the father of the father?
Options are:

- العم *(The uncle)*
- الخال *(The maternal-uncle)*
- الوالد *(The father)*
- الجد *(The grandfather)*

[2] https://github.com/shammur/Arabic-Offensive-Multi-Platform-SocialMedia-Comment-Dataset/annotation_guideline/annotation_guideline.pdf

[3] `http://mturk.com`

No. of Comments	4000
Cost per Comment	1 cent
Comments/HIT	25
Gold/HIT	5
Gold Review Policy	dynamic
Assignments/HIT	3
Task duration	7 days
Assignment Duration	1 hour
Avg. Annotator/Comment	3.23
Max. Annotator/Comment	5*
No. Comment with more than 3 annotation	787

Table 1: Crowdsourcing annotation task details. *In this case one or two annotators could not pass the quality evaluation. Thus their annotations were rejected and the task was extended to a new annotator.

The annotator must provide correct answers to 80% of the questions (i.e., 8 out of 10) to pass the qualification test. Once the annotators pass the test, they can attempt the main classification task. A total of 26 annotators qualified this test; among them, only ≈ 54% scored full and the rest of the annotators scored the minimum, 80%, to pass the test.

In order to evaluate an annotator's assignment for task accuracy, we used gold standard instances hidden in the designed Human Intelligence Tasks (HITs). For each HIT, we assigned 25 comments for the annotation and out of them, 5 were randomly assigned as the test questions. A threshold of 80% is again set for the selection criteria implying that 4 out of 5 gold instances must be correct in order for the assignment by an annotator to be accepted. These test questions (comments) are selected randomly from a pool of 60 comments and are annotated by the domain experts. The selected gold comments included 55% (out of 60) of offensive comments and have an average agreement of ≈ 95.63%.

A closer look into the agreement per class indicates, that the annotators were more confident in case of non-offensive comments with an average agreement of 97.91% (minimum agreement of 84.76%) in compare to offensive comments (average = 95.63% with minimum % of agreement is 61.7%).

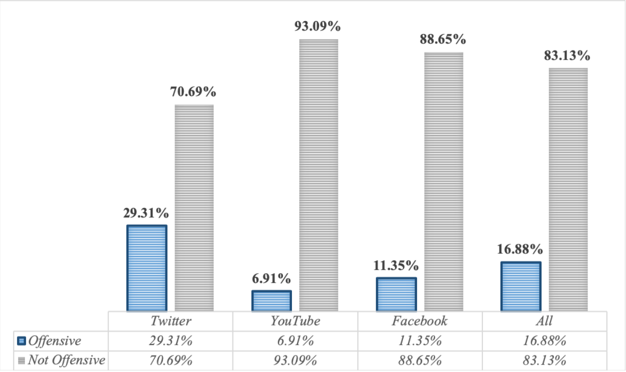

Figure 2: Platform wise annotation label distribution. "All" represents the total % of instances for offensive and not-offensive comments in the annotated dataset.

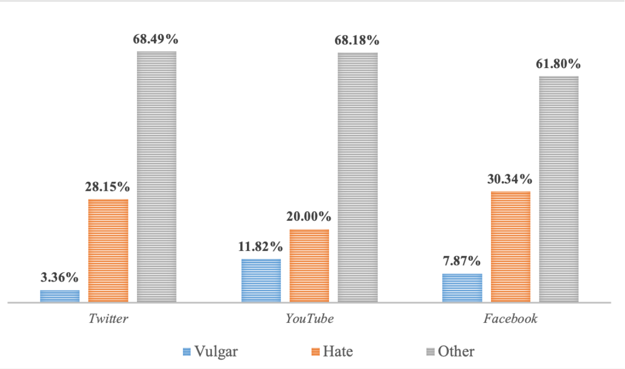

Figure 3: Platform wise annotation label distribution for types of offensive comments present in the annotated dataset.

3.5. Annotation Results

Annotation Agreement and Evaluation: As each comment has 3 annotations (judgments), we assigned the label agreed by the majority (i.e. $\frac{2}{3}$) of the annotator. We have observed that for offensive comments, only 54.2% of comments have unanimous agreement, whereas for non-offensive comments – 83.7% of the annotators are in complete agreement. This shows the complexity of identifying the offensive task, especially in an Arabic language context, due to the ambiguity of lexical variations in Arabic dialects, compared to other languages such as English and German. For the crowd-annotated dataset we obtain, a Fleiss's kappa (Falotico and Quatto, 2015), $\kappa = 0.72$. Although, κ theoretically applies to a fixed set of annotators, but for our case, we randomly assigned 3 judgments to the hypothetical annotator A1-3.

To assess the reliability of the crowd-annotations, we randomly selected 500 comments to be annotated by a domain expert. Using the expert annotation as the reference, we observed that the accuracy of our crowdsourced dataset is 94%. From further error analysis, we observed that the mislabeled comments by the crowd annotators include offensive contents written using sarcasm, irony along with some long comments mixed with statements rather than opinion.

Annotation Distribution: Based on the majority voting, we present the distribution of the finally obtained annotated dataset in Figure 2. From the dataset, we observed that the percentage of offensive comments varies in each platform, as shown in Figure 2, with TW comments, which has more offensive instances followed by FB and then YT.

Types of Offensive Comments: To further understand the annotated data and to discriminate between the different types of offensive comments, we manually annotated the obtained 675 (16.88%, see in Figure 2 – 'All:Offensive') offensive comments, in our dataset, for hate speech and vulgar (but not hate) categories. The obtained distribution of each class (vulgar and hate speech) are presented in Figure 3. From the class distribution, we can observe that in all the three platforms "hate" categories are more prominent with respect to only "vulgar". Our results indicate that FB, followed by TW, to have significantly more hate comments than YT channel comments.

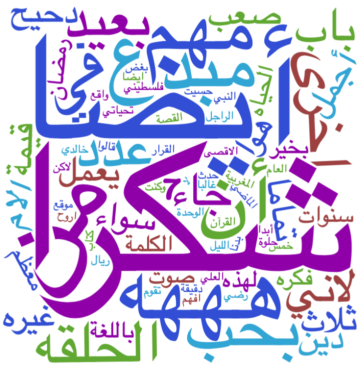

Figure 4: Word-cloud for highly non offensive (NOT) tokens, with valence score, $\vartheta(.) = 1$, in comments. Some of the most frequent words include 'thank you', 'important' along with other conversation fillers and reaction like 'hahah'.

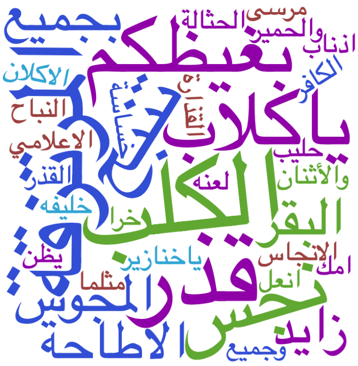

Figure 5: Word-cloud for highly offensive tokens, with valence score, $\vartheta(.) = 1$, in comments. Corresponding English examples can be found in Table 3.

4. Data Analysis

4.1. Lexical Analysis

In order to understand how distinctive the comments are with respect to the offensive and non-offensive classes, we analyzed the lexical content of the dataset.

We compared the vocabularies of the two classes using the valence score (Conover et al., 2011; Mubarak and Darwish, 2019) ϑ for every token, x, following the Equation 1:

$$\vartheta(x) = 2 * \frac{\frac{C(x|OFF)}{T_{OFF}}}{\frac{C(x|OFF)}{T_{OFF}} + \frac{C(x|NOT)}{T_{NOT}}} - 1 \tag{1}$$

where $C(.)$ is the frequency of the token x for a given class (OFF *or* NOT). T_{OFF} and T_{NOT} are the total number of tokens present in each class. The $\vartheta(x) \in [-1, +1]$, with $+1$ indicating that the use of the token is significantly highly in offensive content than non-offensive and vice-versa.

We identified the most distinctive unigrams in each classes, using valence score and presented using word-clouds in Figure 4 and Figure 5. We also presented top 10 highly offensive tokens in Table 3 with its frequency. Furthermore, using the same score, we also presented top highly distinctive bi- and tri-grams, in Table 2.

From the lexical analysis, we observed that words like 'dogs', 'bark', 'dirty', 'envy' are more common in offensive instances than not-offensive instances. Similarly when looking into the bi- and tri-grams, we observed that use of words like 'rage', 'curse' are very frequent to offensiveness.

4.2. Use of Emojis in Offensive Language

Analyzing the comment reveals treats that are specific to each of the types: offensive and non-offensive. Further analysis of the usage of emojis shows that there are some specifics for each type. Their usage of emojis is different.

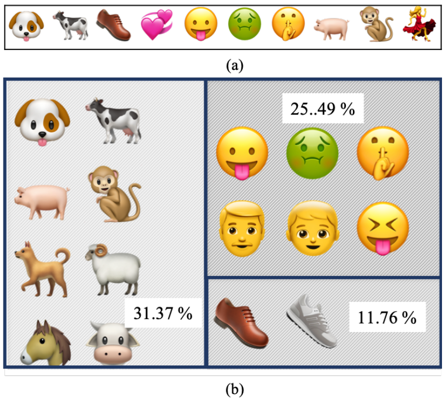

Figure 6: Reported analysis on use of emojis in offensive comments. Figure (a) – lists the 10 most highly offensive emojis with valence score, $\vartheta = 1.0$. Figure (b) – Top 3 groups of emojis with $\vartheta = 1.0$. For each group, ω is reported in percentage.

To understand the distinctive usage of emoji, we also calculated the valence $\vartheta(.)$ score for all the emojis. Figure 6, shows top 10 emojis present in the offensive comments dataset with valence score, $\vartheta(.) = 1$.

Furthermore, we grouped all the emojis based on their categories – including "people face and smilieys", "animal and natures", among others – as defined by Unicode Consortium[4]. We then calculated the weight of each group, ω_g using the following Equation 2 :

$$\omega_g = \frac{\sum_{e_i \in g} C(e_i|\vartheta = 1.0)}{\sum_{e \in L} C(e|\vartheta = 1.0))} \tag{2}$$

where for e_i represent the i^{th} emoji present in group g, given the valence score $\vartheta = 1.0$. L is the list of emojis present in the offensive dataset and $C(.)$ is the usage frequency of the given emoji. The ω for top 3 frequent groups are shown in Figure 6.

Our analysis of offensive emoji also contained studying the most frequent bi-grams with high valence score $\vartheta = 1.0$. We observed that the top 2 most frequent bi-grams are also from the *animals group* – " 🐶 🐶 " (frequency = 3) and " 🐄 🐄 " (3).

[4] https://unicode.org/emoji/charts-12.1/emoji-ordering.html

Highly OFF		Highly NOT	
Ngrams	*Freq.*	*Ngrams*	*Freq.*
قناة الخنزيره (piglet channel)	5	رسول الله (Messenger of God)	27
موتوا بغيظكم (die with your rage)	4	جدا جدا (many many)	17
تم الاطاحة (was downed)	4	حول قوة (will power)	17
الاطاحة بجميع (downing all)	4	الدول العربية (Arab countries)	15
بجميع مخططات (with all plots)	4	ماشاء الله (God's willing)	14
لعنة الله عليكم (God' curse on you all)	4	قوة الا بالله (power except with God)	17
حليب الحمير (Donkey's milk)	4	دول الخليج (Gulf countries)	10
و مين (and who)	2	انا كنت (I was)	9
مين اللي (who is the one)	2	الا الله (except God)	9

Table 2: List of top most frequent n-grams (n=2,3) for both offensive (**OFF**) and non-offensive (**NOT**) comments, using valence score, are as follows. (.) are the translation in English. Freq. represents frequency of the ngram in the corresponding class.

Tokens	**Frequency**
الخنزيره (Piglet)	21
الكلب (Dog)	8
قذر (Dirty)	8
المرتزقة (Mercenaries)	7
تنبح (Bark)	6
بغيظكم (With your envy/rage)	5
نجس (filthy)	5
ياكلاب (O Dogs)	5
أنف (Nose)	4
مخططات (Plots)	4

Table 3: List of top 10 most highly offensive unigrams, with $\vartheta(word) = 1.0$. (.) are the translation in English.

Consequently, our findings suggest that emojis representing animals appear uniquely in offensive comments, indicating their use by analogy for offensive purpose. This observation is in line with our findings for distinctive lexical tokens (such as الكلب – *"dog"*) usage in offensive comments.

5. Experiments and Results

This section describes the details of feature extraction and machine learning algorithm, used to test the performance of the dataset. The motivation for such experimental exploration is to test the performance of the released dataset and to have baseline results in order to facilitate future studies.
In addition, we also utilize the presence of comments from a different platform, to test *generalizability* of a platform-dependent model (trained on a specific platform data) across (a) other social media/news platforms and (b) out-of-domain short content and comment datasets.

5.1. Datasets

In addition to our dataset, refereed as **M**ulti Platforms **O**ffensive **L**anguage **D**ataset (MPOLD), we also utilized three previously published dataset, details in Table 4.
For evaluating the performance on out-of-domain[5] datasets, we used Egyptian and Levantine Arabic datasets.
The Egyptian Arabic dialect data includes 100 TW posts and their 10 corresponding comments[6] per tweet. The tweets are extracted from 10 most controversial Egyptian TW users (Mubarak et al., 2017). This dataset is refereed in this paper as **E**gyptian **T**weets and corresponding **C**omment **D**ataset (ETCD) for further mentions. The Levantine dataset (referred to as L-HSAB)(Mulki et al., 2019), are collected from the timelines of politicians, social/political activists, and TV anchors and does not contain any comments.
Similarly, we also used the **D**eleted **C**omments[7] **D**ataset (further refereed as DCD) from `Aljazeera.net` (Mubarak et al., 2017), to study how a offensive detection model perform in a different platform/s.

Preprocessing: To prepare the data for classification, we first tokenized the input sentences removing punctuation, URLs, stop words, diacritics present in the text. We inserted whitespace in cases where there was no separation between the adjacent words and emojis in the text. For our study, we kept the emojis and hashtags due to the contextual information represented using them.

5.2. Classification Design

As a baseline for the MPOLD dataset, we conducted experiments using a classical machine learning algorithm – SVM. The SVM algorithm (Platt, 1998) is based on the Structural Risk Minimization principle from computational learning theory. The algorithm is proven as universal learners and well-known for its ability to learn independently of the dimensionality of the feature space and the class distribution of the training dataset. These properties of the algorithm make it one of the most popular supervised classification

[5] Here representing that the contents are not news posts.

[6] `http://alt.qcri.org/~hmubarak/offensive/TweetClassification-Summary.xlsx`

[7] `http://alt.qcri.org/~hmubarak/offensive/AJCommentsClassification-CF.xlsx`

Abbr.	Dataset Description	Labels	#I (#OFF%)	avg. len	Used for
MPOLD*	News comments from TW, FB, and YT	OFF, NOT	4000 (16.9%)	22.8	train/ test
		V, HS, Oth		18.0/18.5/41.7	
ETCD*	100 Egyptian tweets and corresponding comments	OBs, OFF-OBs, NOT	1100 (≈ 59.0%)	13.5	test
L-HSAB*	Levantine tweets	Abs, HS, NOT	5846 (≈ 37.6%)	12.0	test
DCD**	Deleted comments from a news website	OBs, OFF-OBs, NOT	31692 (≈ 82%)	17.4	train/ test

Table 4: Details of the datasets used for the study. * represent the out-of-domain dataset and ** represent the different platforms. #I represent the number of instances, whereas #OFF% represents the number of offensive labels used in this study. In the case of ETCD and DCD, we merged the offensive comments – obscene (OBs) and other offensive comments (OFF-OBs), as OFF for matching with our labels. Similarly, we merged HS and Abs (abusive) for the L-HSAB dataset to OFF. avg. len represents the average length of the corresponding content in terms of the number of words. V - Vulgar, HS - Hatespeech Oth-Other offensive categories.

methods and thus our choice for the baseline and all the experiments conducted in this study.

To train the classifier, we transform the online comments into bag-of-character-n-grams vector weighted with logarithmic term frequencies (tf) multiplied with inverse document frequencies (idf). For this, we created character n-grams using character inside word boundaries only and padded the edge character with whitespace. To utilize the contextual information, i.e n-grams, we extracted features using $n = 1$ to $n = 5$.

5.3. Performance Evaluation

To measure the performance of the classification, we reported macro-averaged F_1-measure (F_m), calculated using macro-averaged precision and recall – an average of P and R for both classes, respectively.

The motivation behind such choice is due to the imbalanced class distribution which makes the well-known measures such as accuracy and micro-average F_1-measure not well-representative of the performance. Since the performance of both classes is of interest, we also report the F-measure of the individual classes.

	F$_m$	NOT	OFF
Chance	0.49	0.83	0.14
Lex	0.53	0.84	0.21
SVM	**0.74**	**0.92**	**0.56**

Table 5: Reported performance on MPOLD dataset using 5-fold cross validation (with SVM and Chance baseline)

5.4. Results

MPOLD performance: To evaluate the classification performance, we performed five-fold[8] cross-validation on the dataset, maintaining the natural distribution of the classes in each fold. The performance on all the instances is shown in Table 5.

Testset	Trained on	F$_m$	NOT	OFF
FB	DCD	0.34	0.43	0.25
	TW	0.62	0.90	0.34
	YT	0.62	0.92	0.31
	TW + YT	**0.68**	***0.93***	***0.44***
	TW + YT + DCD	**0.78**	***0.94***	***0.63***
TW	DCD	0.52	0.51	***0.53***
	FB	0.51	0.83	0.19
	YT	0.54	0.82	0.25
	FB + YT	**0.59**	***0.84***	0.35
	FB + YT + DCD	**0.84**	***0.90***	***0.78***
YT	DCD	0.44	0.67	0.22
	FB	0.53	***0.96***	0.10
	TW	0.60	0.94	0.27
	TW + FB	**0.63**	0.95	***0.31***
	TW + FB + DCD	**0.82**	***0.97***	***0.67***
DCD	TW	0.37	***0.39***	0.34
	FB	0.21	0.31	0.10
	YT	0.29	0.26	0.32
	FB + YT + TW	**0.40**	0.35	***0.45***

Table 6: F-measure performances based on platform wise dataset. The test sets presented here are a subset of MPOLD data. For instance TW ∈ MPOLD, is the dataset extracted from Twitter. Similarly, YT represents YouTube data, FB represents comments from Facebook. DCD - Deleted comments dataset. NOT = Not-offensive, OFF = offensive instances. The blue rows presents the results when the platform-wise data present in our dataset is modelled along with online-news-platform data, DCD.

To compare the obtained results, we reported *chance level* baseline based on the prior distribution of the labels in each fold. In addition to this, we also present a simple lexical-based baseline (*Lex*) using a list of possible offensive words and phrases. This word list is the combination of obscene words presented in (Mubarak et al., 2017) and filtered entries from Hatebase (`Hatebase.org`) lexicon. From Table 5, we observed that the SVM performs significantly better than the lexicon and chance level baselines.

Platform specific performance: To access the generalization capabilities of the designed model, we evaluated

[8] We choose 5 instead of 10, as the number of folds, to ensure a good sample size and distribution of both the labels in each fold for test purpose.

Testset	Model	F$_m$	NOT	OFF
ETCD	Lex	0.51	0.58	0.44
	DCD	0.64	**0.76**	0.53
	MPOLD	0.61	0.66	0.56
	MPOLD$^+$	**0.68**	0.61	**0.75**
L-HSAB	Lex	0.46	0.74	0.18
	DCD	0.59	0.60	0.58
	MPOLD	**0.64**	**0.77**	0.51
	MPOLD$^+$	0.62	0.64	**0.59**

Table 7: Reported F-measures of the models trained using DCD, MPOLD (MPOLD) and MPOLD + DCD (MPOLD$^+$) and tested on out-of-domain test sets. The test sets includes: (a) ETCD - controversial Egyptian tweets and comments dataset; and (b) L-HSAB Levantine tweets datasets. *Lex* is lexicon-based baseline for the test sets.

the models using leave-one-platform-dataset-out (LOPO) cross-validation. The performance of the models on each social media platform data is presented in Table 6. Besides the three platforms in our dataset, for the experiment, we also included DCD – deleted comments from online news portal, as a fourth platform.

We observed that in most cases models trained on one particular platform performs poorly compared to the models trained using multiple-platform datasets. This multiple platform model includes both the combination of (a) 2 platforms from MPOLD data; or (b) 2 platforms from MPOLD + DCD datasets.

For instance, the model trained on $TW + YT$ performs significantly better, on FB test set, than the individual YT and TW models. The same pattern is also observed when the performance of DCD in combination ($YT + TW$) model are compared. In this case, a significant jump in OFF class is noticed thus increasing the overall macro-averaged f-measure (F_m). This indicates the *importance of multiple platform datasets* for generalization of the offensive detection model.

In addition to a different platform data of a similar domain, DCD dataset also brings two more added advantages for training the model. These advantages include (a) addition of comments written using MSA along with DA, and (b) increase the number of offensive examples for the model training. The MPOLD comments are mostly dialectal Arabic, therefore adding DCD includes examples of comments in MSA for the model to recognise and learn. At the same time, increases the number of (not)offensive examples for the training. Thus the results from the combined platform models (for eg. $TW + YT + DCD$) indicates the impact of varieties in examples and size of the training data.

However, by comparing the results of individual models – trained on only DCD, TW, YT or FB – we observed that even though DCD has a large number of training instances, this model cannot outperform TW/YT/FB alone in most cases. This again highlights the inadequacy of model generalization when trained on a particular platform and tested on another platform.

Cross domain performance: The performance of the models trained on MPOLD and MPOLD+DCD (MPOLD$^+$) across domain is presented in Table 7. The results indicate that for ETCD dataset, both our models (MPOLD and MPOLD$^+$) outperform the baseline and the DCD models for OFF class f-measure. The overall best performance is obtained using MPOLD$^+$, with an improvement of 7% and 4% over MPOLD and DCD models.
As for L-HASB data, MPOLD model outperforms all other models in terms of overall f-measure. However, for the offensive class only (OFF) the best performance is obtained by MPOLD$^+$.
The performance of the MPOLD and MPOLD$^+$ models on both ETCD and L-HSAB dataset indicates *the positive impact of multiple-platform* training dataset and implicates the cross-domain generalizability.

6. Conclusion

In this paper, we introduced one of the first dialectal Arabic comment dataset extracted from multiple social media platform. Utilizing this dataset, we designed classification experiments to study the impact of multi-platform data for platform-wise model generalization. For this, we evaluated the model using platform-wise cross-validation. In addition to our dataset, we also exploit existing publicly available dataset. We also utilized deleted comment dataset from a news website. We observed that combined platform models perform significantly higher than most individual models. Thus indicating the importance of such multi-platform dataset.
Using our designed models, we also evaluated other out-of-domain (ETCD and L-HSAB) data extracted from TW. Besides being out-of-domain dataset, another advantage of evaluating these datasets is to see how our models perform in particular for DA content; i.e., ETCD (Egyptian dialects) and L-HSAB (Levantine dialects). Our findings indicate that the multi-platform news comment dataset model has the potential to capture diversity in different dialects and domains. In addition to evaluation of the generalization power of the models, we also presented an in-depth analysis of emoji usage in offensive comments. Our findings suggest emojis in the animal category are exploited in the offensive comments, similar to our lexical observation.
To the best of our knowledge, this is one of the first studies to explore the effect of multiple platform datasets, for offensive language detection, on cross-platform, cross-domain and for dialectal varieties present in Arabic. In future, we plan to extend the study by introducing a larger dataset with further fine-grained classification and content analysis.

7. Bibliographical References

Agarwal, S. and Sureka, A. (2014). A focused crawler for mining hate and extremism promoting videos on youtube. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 294–296. ACM.

Al-Ajlan, M. A. and Ykhlef, M. (2018). Optimized twitter cyberbullying detection based on deep learning. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–5. IEEE.

Alakrot, A., Murray, L., and Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181.

Albadi, N., Kurdi, M., and Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017a). Hate is not binary: Studying abusive behavior of# gamergate on twitter. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 65–74. ACM.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017b). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22. ACM.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017c). Measuring# gamergate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th international conference on world wide web companion*, pages 1285–1290. International World Wide Web Conferences Steering Committee.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.

Dadvar, M., Trieschnigg, D., Ordelman, R., and de Jong, F. (2013). Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Difallah, D., Filatova, E., and Ipeirotis, P. (2018). Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 135–143. ACM.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In *Twelfth International AAAI Conference on Web and Social Media*.

Falotico, R. and Quatto, P. (2015). Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Gülaçtı, F. (2010). The effect of perceived social support on subjective well-being. *Procedia-Social and Behavioral Sciences*, 2(2):3844–3849.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer.

Intapong, P., Charoenpit, S., Achalakul, T., and Ohkura, M. (2017). Assessing symptoms of excessive sns usage based on user behavior and emotion: analysis of data obtained by sns apis. In *9th International Conference on Social Computing and Social Media, SCSM 2017 held as part of the 19th International Conference on Human-Computer Interaction, HCI International 2017*, pages 71–83. Springer Verlag.

Jay, T. and Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288.

Kumar, S., Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, pages 933–943. International World Wide Web Conferences Steering Committee.

Mondal, M., Silva, L. A., and Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94. ACM.

Mubarak, H. and Darwish, K. (2019). Arabic offensive language classification on twitter. In *International Conference on Social Informatics*, pages 269–276. Springer.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Mulki, H., Haddad, H., Ali, C. B., and Alshabani, H. (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.

Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. (2017). Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135.

Pitsilis, G. K., Ramampiaro, H., and Langseth, H. (2018). Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Ross, J., Zaldivar, A., Irani, L., and Tomlinson, B. (2009). Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep*.

Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S.-g., An, J., Kwak, H., and Jansen, B. J. (2018). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In *Proceedings of The International AAAI Conference on Web and Social Media (ICWSM 2018)*, San Francisco, California, USA, June.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-g., Almerekhi, H., and Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1).

Silva, L., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2016). Analyzing the targets of hate in online social media. In *Tenth International AAAI Conference on Web and Social Media*.

Unsvåg, E. F. and Gambäck, B. (2018). The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85.

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language.

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