© https://fti-tn.net/publications Future Technologies and Innovations (FTI) Proceedings: 4th international conference on computer applications and information security (iccais'2021) / March 19 / 2021/ Tunisia: https://fti-tn.net/iccais-2021-list-of-papers

SalamREPO: an Arabic Offensive Language Knowledge Repository

Fatemah Husain
Department of Information Science, Kuwait University
Kuwait City, State of Kuwait
f.husain@ku.edu.kw

Ozlem Uzuner
Department of Information Sciences and Technology
George Mason University
Fairfax, USA
ouzuner@gmu.edu

***Abstract*— The problem of online offensive language has become universal. There has been very few research in developing Arabic resources to support the detection of offensive language from user-generated content. Previous work on this topic did not consider the variation of Arabic dialects and cultures. In contrast, this research aims at extracting knowledge from several different offensive language datasets with the purpose of building a comprehensive dialectal and cultural knowledge repository for Arabic offensive language, called SalamREPO. SalamREPO contains multiple offensive and not offensive lists of bigrams and trigrams that we generate using collocation extraction techniques following a corpus-driven approach.**

Keywords-Natural Language Processing; Offensive Language; Collocations; Arabic Language; User-Generated Content.

I. INTRODUCTION

Arabic is a challenging language for Natural Language Processing (NLP) because of its variation in forms and dialects, especially for user-generated content. Online offensive language is further complicated by the culturally specific nature of offensive content. Moreover, offensive terms often span the edges of one word, in a way that creates completely non-compositional terms, whose meaning needs to be preserved for effective offensive language detection.

A wide range of countries and communities across the globe speak Arabic; each has a different culture, which creates complex offensive terms that might not be understood by others from other Arabic speaking communities. For example, the word "عافية/Afiah" means "health" in Gulf, Egyptian, Iraqi, and Levantine and is often used within not offensive context, while it means "fire" in Moroccan and is often used within offensive context.

Having domain experts covering all Arabic cultures to create an accurate repository of offensive terms is costly. In addition, offensive terms are emerging, which means new offensive terms always need to be added to the offensive terms repository. The currently available resources for Arabic offensive terms are very limited in terms of their size, scope, and the algorithms used in extracting these terms.

This work aims to apply collocation extraction techniques using a corpus-driven approach to develop an Arabic offensive knowledge repository called SalamREPO. It combines knowledge from several Arabic dialects and platforms. The use of collocation extraction techniques supports identifying non-compositional phrases, phrases whose meaning differs from the composition of their parts. Having a repository that combines cultural and dialectal knowledge about offensive terms can reduce the gap between human-level understanding and system-level understanding of Arabic offensive language. To our knowledge, there is no previous Arabic repository that includes collocation sets of offensive and not offensive phrases as SalamREPO.

II. BACKGROWND

A. *Collocations*

Manning and Schütze (1999) describe collocations as "an expression consisting of two or more words that correspond to some conventional way of saying a thing" [5: P.141]. They characterize collocations by limited compositionality. They define some principal approaches to finding collocations, including selection of collocations by frequency, selection based on mean and variance of the distance between the main word and the collocating words, hypothesis testing, and mutual information.

The simplest method of finding collocations uses co-occurrence frequencies. While this frequency-based method works sufficiently well with large corpora, it suffers from some limitations, such as free word combinations that are not necessary collocations, and it cannot detect infrequent collocations. Alternative methods find collocations using word-based distance methods that use mean and variance. The use of mean and variance supports flexibility in the relationship between collocated words by defining a collocational window to account for the co-occurrence of the words, rather than having fixed phrase distance. The collocational window sets the number of words on each side of the main word to be considered in generating collocations. The offset of the collocated words in the corpus is used to calculate the mean and the variance.

Both frequency-based and mean and variance-based methods require testing to check if the co-occurrence happened by chance or if they indicate actual collocations. Manning and Schütze [5] emphasize the importance of integrating hypothesis testing into the process of finding collocations. Among the statistical tests that are commonly applied are the t-test,

Pearson's chi-square test (x^2), and the likelihood ratio. The t-test takes a sample from the data and assumes that the sample is drawn from a normal distribution with the same mean (μ).

Unlike the t-test, the chi-square test (x^2) assumes that the sample is not normally distributed. It compares observed frequencies to expected frequencies. The larger the difference, the larger the confidence that the co-occurred words are actual collocation and not co-occur by chance. For sparse data, the likelihood ratio is a better choice for hypothesis testing, especially for finding rare collocations.

The main concept of Pointwise Mutual Information (PMI) is borrowed from the information theory. PMI measures how much more the collocated words co-occur than if they were independent. Manning and Schütze [5] pinpoint that PMI is not widely used for collocation extraction because other methods show more accurate results.

Manning and Schütze [5] propose some heuristics to improve collocations and to filter for phrases, such as applying part-of-speech pattern filtering defined by Justeson and Katz (1995). These patterns are provided for the English language, and consider Adjectives (A), Nouns (N), and Prepositions (P) such as A/N, N/N, A/A/N, A/N/N, N/A/N, N/N/N, and N/P/N. Al-Mustansiriya [6] defines 12 patterns for Arabic bigram collocations that consider Verbs (V), Nouns (N), Noun Phrases (NP), Prepositional Noun Phrases (PNP), Conjunctions (C), Adjectives (A), and Adverbial Phrases (AP). Table I shows some examples of Al-Mustansiriya's [6] 12 patterns.

TABLE I. PATTERNS OF ARABIC COLLOCATIONS (based on Al-Mustansiriya [6])

Tag Pattern	Example in Arabic	Example in English (Translation)
V[a] /N[b]	ضرب الخيمة	he pitched the tent
V/PNP[c]	استقال من العمل	he resigned from work
V/PNP (adverb)	نفذ بشدة	he precisely implemented
V/NP[d] (adverbial-condition)	اتصل هاتفيا	he made a phone call
V/C[e]/V	طار وحلق	he flew and soured
N/N	مسرح الاحداث	scene of events
V/C/N	عزم وإصرار	intention and insistence
N/A[f]	قوة عظمى	ultimate power
N/PNP	في غاية الادب	extremely polite
N/P[g]	مقارنة ب	in comparison with
A/N	حسن الاخلاق	having high morals
A/AP	مستنكر بشدة	strongly condemns

a. verb, b. noun, c. prepositional noun phrase, d. noun phrase, e. conjunction, f. adjective, g. preposition.

B. *Related Work*

Current available offensive terms repositories are not inclusive to wealth of Arabic cultures and dialects [7]. One of the largest hate speech resources is Hatebase[1], which covers over 90 languages, but has limited number of Arabic terms, most of which are not correctly written. A large number of the Arabic words in Hatebase are written using English alphabet, which creates inconsistency among the terms in the same list. This inconsistency of Hatebase's content is due to the fact that most of these terms are added through crowdsourcing without following clear pre-defined rules for data entry. The PeaceTech Lab[2] avoids the problem of crowdsourcing offensive terms by providing a series of hate speech lexicons based on countries covering areas of political conflicts that have been collected through surveying people from the same area. However, their lexicons cover very limited Arabic-speaking countries; Iraq, Yamen, Libya, and Sudan.

Some Arabic offensive language researchers develop offensive word lists. Conversely, these lists are very limited to specific datasets in scope. For example, Mubarak, Darwish, and Magdy [8] construct an Arabic offensive dataset from Aljazeera.net news comments, and apply the Log Odds Ratio (LOR) to extract a lexicon list for obscene and offensive terms. The list contains unigrams, bigrams, and phrases that appear more than nine times in the dataset. Then, some manual assessment is conducted to ensure correctness of the lexicons. The final list includes a total of 288 unigrams, bigrams, and phrases in addition to 127 hashtags.

A domain specific lexicon has been provided by Albadi, Kurdi, and Mishra [9], in their religious hate speech study. They create a Twitter dataset for religious hate speech in Arabic, then, they use the dataset to extract multiple hate unigram lexicons by using different algorithms; Chi-square statistical test for AraHate-Chi, PMI for AraHate-PMI, and Bi-Normal Separation (BNS) sentiment scoring method for AraHate-BNS. The final list includes 1,523 words and their corresponding hate scores.

A dialect specific dataset has been constructed for Levantine from Twitter by Mulki et al. [10]. The authors create two unigram lexicon lists for hate speech and abusive words that consists of ten words each, based on the tweets from their dataset. Each word in the lexicon lists has a score to indicate the degree of the relatedness of each word to the class; hate or abusive, which are calculated using words distribution for each class and word-class correlation.

Haddad et al. provide a dialect specific lexicon for Tunisian. They start by creating a Tunisian hate and abusive dataset. This dataset is used as the source to develop multiple lexicons. The first lexicon list includes the most frequent ten unigrams for hate speech and abusive language along with a percentage of the distribution under the specific class. Another lexicon is created for discriminatory terms within each class, which is constructed by calculating the word's correlation with the normal class; then, assigning a hate score (HtS) and an abusive score (AbS) for the words that either mostly or rarely appear in hate speech and abusive language classes.

Our approach differs from previous studies by applying bigrams and trigrams collocation extraction methods using various offensive language datasets. It provides researchers in the field with dialectal and cultural collocations that could be

[1] https://hatebase.org/

[2] https://www.peacetechlab.org/

used in multiple NLP applications for offensive and not offensive domains.

III. METHODOLOGY

Figure 1 presents an overview of the primary approach of constructing SalamREPO. The first phase consists of collecting Arabic offensive datasets. Second, we filter and preprocess the datasets. Third, we apply previously discussed Manning and Schütze [5] principles to extract collocations. Then, we rank the results, conduct manual validation, and filter the collocations again to ensure that all of them are actual collocations. At the end, we organize the final lists of collocations into bigrams and trigrams lists.

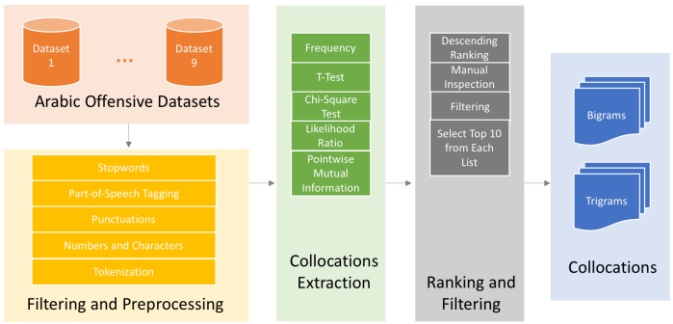

Figure 1. SalamREPO construction methodology

A. Datasets

We construct the repository based on the content from nine Arabic Offensive language datasets. These datasets cover several types of offensive content as follows:

1. The Aljazeera.net Deleted Comments dataset includes three classes: offensive (25,506 comments), obscene (533 comments), and clean (5,653 comments) [8].

2. The Egyptian Tweets dataset includes three classes: offensive (444 tweets), obscene (203 tweets), and clean (453 tweets) [8].

3. The YouTube dataset includes two classes: offensive (5,813 comments) and not offensive (9,237 comments) [12].

4. The religious hate speech dataset has two hierarchies of labeling but we focus only on the first labels, which has binary classes: hate (2,762 tweets) or not hate (3,375 tweets) [9].

5. The Levantine Twitter Dataset for Hate Speech and Abusive Language (L-HSAB) includes three classes: hate (468 tweets), abusive (1,728 tweets), and normal (3,650 tweets) [10].

6. The Tunisian Hate and Abusive Speech (T-HSAB) includes three classes: hate (1,078 comments), abusive (1,126 comments), and normal (3,820 comments) [11].

7. The Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) includes two hierarchies for labeling. The first label has two classes: offensive (1,900 tweets), and not offensive (8,100 tweets). The second label has two classes: hate (500 tweets), and not hate (9,500 tweets) [13].

8. The multi-platform hate speech dataset includes binary classes: hate (10,000 comments) or not hate (10,000 comments) [14].

9. The Multi-Platform Offensive Language Dataset (MPOLD) includes two classes: offensive (675 comments) and not offensive (3,325 comments) [15].

B. Filtering and Preprocessing

We drop all duplicated samples from the datasets. We split the datasets based on the labels. We use stop words provided by the NLTK library for the Arabic language to filter out Arabic stop words. Text was further cleaned to remove all numbers, special characters (e.g., emoji, RT, @), and punctuations as defined by the NLTK library for Arabic and English.

We prepare text for two-word collocations and three-word collocations by using nltk.collocations to find bigrams and trigrams. To extract collocations of specific grammatical patterns, we create a language model for part-of-speech tagging using nltk.tag.stanford and test it on a sample of our data. We find that the results were not accurate because Stanford Arabic Tagger is trained on Modern Standard Arabic (MSA) and our datasets are in dialectal Arabic. Thus, we omit grammatical patterns during this step, but utilize them for manual inspection using the Arabic collocation patterns provided by Al-Mustansiriya [6] to see if the patterns are still applicable for dialectal Arabic.

C. Collocations Extraction

The NLTK library provides functions for collocation extraction. We use the student_t for t-test, chi_sq for Chi-Square test, likelihood_ratio for likelihood ratio, and pmi for PMI from nltk.collocations.BigramAssocMeasures for bigram collocations and nltk.collocations.TrigramAssocMeasures for trigram collocations. Additionally, we calculate collocations based on simple frequencies of co-occurrence.

D. Ranking and Filtering

The resulting collocations were ranked based on their scores from highest to the lowest. Initially, we create class-based lists of the top 20 collocations from each measurement per dataset. We manually investigate all lists to only include actual collocations. We investigate the grammatical patterns exhibited

by the Arabic collocations [6]. Table II shows examples of patterns from the top-ranked collocations.

At the same time, during manual inspection, we filter the collocations' lists to remove compositional phrases, such as "هزه ارضيه بالكويت"/ "earthquake in Kuwait" and " السيسي يصدر قرارا"/ "Al-Sisi issues a decision". Another example is " الى جهنم وبئس"/ "go to hell", which was removed; however, " جهنم وبئس المصير"/ "go to hell" that has the same meaning was kept because it is the minimal form of the collocation. The minimal form in this context is the form that preserves the complete meaning as a unit (collocation) without additional connecting particles. We adjust some collocations that include the conjunction "و", which means "and" without a space to separate it from the first word to remove it from the collocation if it is the leading word in the phrase. For example, "ودين الاسلام اكمل"/ "and Islam is the complete religion" is converted to " دين الاسلام اكمل"/ "Islam is the complete religion". Moreover, we find several names of famous personalities; singers, ministers, actors…etc.; thus, we translate their names into English with short explanation of their main role between parentheses. For instance, "راغب علامة" was translated to "Ragheb Alama (Lebanese singer)".

TABLE II. SAMPLE PATTERNS OF COLLOCATIONS FROM SALAMREPO

#	Tag Pattern	Example in Arabic	Example in English (Translation)
1	A[a]\N[b]	صغير الرياض	Small of Al-Riyath
2	N\N	ابن الكلب	Dog's son
3	NP[c]N	الاطاحة بجميع مخططات	The failing of all plans
4	N\A\N	الطائفة 'اليهودية صنعاء	The 'Sanaa' Jewish community
5	V\N\P	رضي الله عنها	God bless her
6	N\N\N	الإلحاد النفاق البدع	Atheism hypocrisy new
7	N\C[d]\A\N	الله ونعم الوكيل	God and the best agent
8	V\C\N\V	تسير والكلاب تنبح	She is walking and the dogs are barking

a. adjective, b. noun, c. prepositional noun phrase, d. conjunction.

E. *Final Collocations*

We organize the final lists of collocations based on multiple factors. The main criterion is the number of collocated words. Thus, bigrams are listed separately from the trigrams. We assign each dataset a numeric code to be attached to each collocation, and to allow for further analysis. Moreover, we add the class label; such as hate, not hate, abusive, offensive; to each collocation based on the class label of the source that it was extracted from.

IV. RESULTS AND DISCUSSION

A. *Results*

The structure of SalamREPO depends on the Arabic offensive datasets that were used in extracting the collocations. Some datasets have two hierarchies of labeling, such as the OSACT dataset, which classifies text into offensive or not offensive and then further classifies the resulted offensive samples into hate speech or not hate speech. While other datasets have one hierarchy of labeling, such as Aljazeera, L-HSAB, or MPLOT, with three or two classes. Figure 2 shows the overall structure of SalamREPO. as can be seen, it covers both high-level labels of offensive or not offensive, and the more detailed lower-level hierarchies

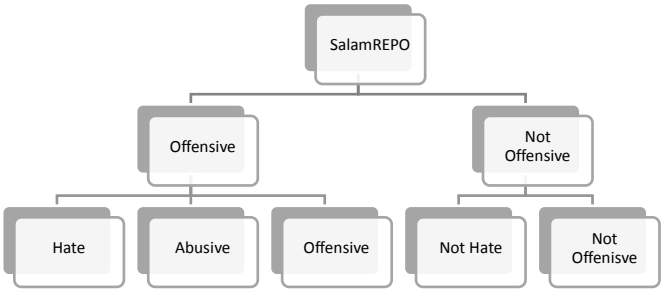

Figure 2. Structure of Collocations Categories

Table III provides some examples for collocations from different offensive and not offensive categories as they appear in SalamREPO. Examples 1 and 2 are generated by simple frequency measurement, 3 and 4 are generated by t-test, 5 and 6 are generated by Chi-Square, 7 and 8 are generated by likelihood ratio, and the last two examples are generated by the PMI.

TABLE III. SAMPLES OF COLLOCATIONS FROM SALAMREPO

#	Collocation	Collocation in English (Translation)	Source[a]	Category	Sub-Category
1	روحي عالمطبخ	not for ladies	5	offensive	hate
2	حبيبي طير انت	get away	2	offensive	offensive
3	ابن وسخة	son of a bitch	8	offensive	hate
4	جهنم وبئس المصير	go to hell	1	offensive	offensive
5	عربت خربت	to become an Arabic is to get ruined	1	offensive	offensive
6	يلبسكم لباس الصحة	become healthy	4	Not offensive	Not hate
7	ترحون فدوه	sacrifice yourself for	3	offensive	-
8	نعم الوكيل فيك	God suffices	7	offensive	-
9	حسبي الله	God suffices	7, 9	Not offensive	Not hate
10	شمس الدين باشا	Shams El-Din Basha (Tunisian singer)	6	Not offensive	Normal

a. Refers to the dataset used in extracting the collocation based on the same order mentioned in the methodology section

The final repository includes a total of 1,084 (450 unique) collocations from all lists. The table below shows the total number of collocations per measurement for bigrams and trigrams separately. Some collocations appear both as offensive and not offensive.

TABLE IV. SAMPLES OF COLLOCATIONS FROM SALAMREPO

	Frequ-ency	t-test	Chi-Squ are	PMI	Like-lihood ratio	Totals	Unique
Bigrams	197	107	135	141	111	691	274
Trigrams	171	54	57	56	55	393	176
Totals	368	161	192	197	166	1084	450

The resulting collocations show some patterns that could be valuable for further consideration and analysis to find their relationship with offensive Arabic content. Some collocations consist of repetitions of the same word, such as "جدا جدا"/ "a lot a lot", "لا لا لا "/ "no no no", and "قهر قهر"/ "broken heart broken heart".

Names for famous figures appear very frequently among the collocations. Some names are for politicians, examples include "جبران باسيل"/ "Gebran Bassil (a Lebanese politician and president of the Free Patriotic Movement)", "صدام حسين"/ "Saddam Hussein (the fifth President of Iraq)", and " محمد بن سلمان"/ "Mohammed Bin Salman (refers to Mohammed bin Salman bin Abdulaziz Al Saud, the Crown Prince of Saudi Arabia)". Other names are for television personalities. For example, "فيصل القاسم"/ "Faisal Al-Qasim (also written as Faisal Al-Kasim is a British-Syrian television personality based in Qatar)" and "احمد منصور"/ "Ahmed Mansour (an Egyptian journalist, television presenter, television host, and interviewer on Al Jazeera Channel)". Some musicians' names are also among the collocations, such as "كاظم الساهر"/ "Kadim Al Sahir (an Iraqi singer)", "نبيل شعيل"/ "Nabil Shuail (Kuwaiti singer)", and "ولد عواطف"/ "Weld Awatef (Tunisian singer)".

Additionally, several prayer phrases are observed among the collocations, including "رضي الله عنها"/ "God be pleased with her", "أطال الله أعماركم"/ "may God give you a long life", and "بارك الله فيكم"/ "may God bless you".

We also find some transliterated phrases from English to Arabic, such as "سناب شات"/ "Snapchat", "ون بلس"/ "one plus", and "البيدوفيل ناكح الأطفال"/ "pedophilia".

SalamREPO will be available upon publication of the paper on GitHub.

B. System Implications

Information about offensive and not offensive collocations can advance automatic offensive language detection in Arabic. Collocations can also be integrated into disambiguation systems to support the identification of different senses. Collocations are also commonly used to support translation systems by identifying cultural stereotypes, idioms, and translating them appropriately.

I. LIMITATIONS

SalamREPO depends on the available Arabic offensive language datasets. Thus, it does not cover all Arabic dialects. The missing dialects need to be addressed. Another limitation is related to the ability to self-learn new offensive terms. Improving the architecture of SalamREPO to include real-time data or periodic data as the source to update the lexicon lists needs to be addressed. Moreover, due to the limited available part-of-speech tagging tools for dialectal Arabic, we expect some error rate from the tools we used in conducting our experiments. Besides, we borrow some of the phrase patterns defined by Justeson and Katz [16] and Manning and Schütze [5], which might need to be adjusted for Arabic by linguistic experts from the Arabic language field to have more accurate results.

II. CONCLUSIONS

In this paper, we discuss the process of developing SalamREPO, a knowledge repository for Arabic offensive language that consists of multi-dialectal offensive and not offensive collocations. We describe five statistical analysis approaches applied to extracting instances of Arabic offensive language collocations from nine datasets. SalamREPO differs from previous Arabic offensive language lexicons by including collocations, which are special non-compositional phrases. The collocations in SalamREPO are multi-dialectal Arabic, bigrams, and trigrams. In the future, we will further extend SalamREPO to include more collocations both in terms of length and patterns.

REFERENCES

[1] J. R. Firth, "A synopsis of linguistic theory, 1930-1955," Studiesin linguistic analysis, 1957.

[2] S. Evert, "The statistics of word co-occurrences: word pairs andcollocations," Ph.D. dissertation, University of Stuttgart, 2005.

[3] R. Krishnamurthy, Collocations, 12 2006, pp. 596–600.

[4] M. Halliday, "Lexis as a linguistic level," vol. 2(1), pp. 57–67,1966.

[5] C. D. Manning and H. Schütze, Foundations of StatisticalNatural Language Processing. Cambridge, MA, USA: MITPress, 1999.

[6] Al-Mustansiriya, "Collocation in english and arabic : A linguis-tic and cultural analysis," 2012.

[7] F. Husain and O. Uzuner, "A Survey of Offensive Language Detection for the Arabic Language." ACM Trans. Asian Low-Resour. Lang. Inf. Process. 20, 1, Article 12 (March 2021), 44 pages.

[8] H. Mubarak, K. Darwish, and W. Magdy, "Abusivelanguage detection on Arabic social media," in Proceedingsof the First Workshop on Abusive Language Online.Vancouver, BC, Canada: Association for ComputationalLinguistics, August 2017, pp. 52–56. [Online]. Available:https://www.aclweb.org/anthology/W17-3008

[9] N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers?analysis and detection of religious hate speech in the arabictwittersphere," in 2018 IEEE/ACM International Conference onAdvances in Social Networks Analysis and Mining (ASONAM),August 2018, pp. 69–76.

[10] H. Mulki, H. Haddad, C. B. Ali, and H. Alshabani, "L-HSAB: A Levantine twitter dataset for hate speech andabusive language," in Proceedings of the Third Workshop onAbusive Language Online. Florence, Italy: Association forComputational Linguistics, August 2019, pp. 111–118. [Online].Available: https://www.aclweb.org/anthology/W19-3512

[11] H. Haddad, H. Mulki, and A. Oueslati, "T-hsab: A tunisian hatespeech and abusive dataset," in Arabic Language Processing:From Theory to Practice, K. Smaïli, Ed. Cham: SpringerInternational Publishing, 2019, pp. 251–263.

[12] A. Alakrot, L. Murray, and N. S. Nikolov, "Datasetconstruction for the detection of anti-social behaviourin online communication in arabic," Procedia ComputerScience, vol. 142, pp. 174–181, 2018a. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S1877050918321756

[13] H. Mubarak, K. Darwish, W. Magdy, T. Elsayed, and H. Al-Khalifa, "Overview of osact4 arabic offensive language detectionshared task," vol. 4, 2020.

[14] A. Omar, T. M. Mahmoud, and T. Abd-El-Hafeez, "Compar-ative performance of machine learning and deep learning algo-rithms for arabic hate speech detection in osns," in Proceedingsof the International Conference on Artificial Intelligence andComputer Vision (AICV2020), A.-E. Hassanien, A. T. Azar,T. Gaber, D. Oliva, and F. M. Tolba, Eds. Cham: SpringerInternational Publishing, 2020, pp. 247–257.

[15] S. A. Chowdhury, H. Mubarak, A. Abdelali, S.-g. Jung,B. J. Jansen, and J. Salminen, "A multi-platform Arabicnews comment dataset for offensive language detection," inProceedings of The 12th Language Resources and EvaluationConference. Marseille, France: European Language ResourcesAssociation, May 2020, pp. 6203–6212. [Online]. Available:https://www.aclweb.org/anthology/2020.lrec-1.761

[16] Justeson, J.S. and Slava M. Katz (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1):9-27.