Resources Building for Arabic Harmful Online Content: Survey

Fatiha Charef, Abdelhafid Zitouni, Mahieddine Djoudi, Hichem Rahab

Resources Building for Arabic Harmful Online Content: Survey

Fatiha Charef[1,*], Abdelhafid Zitouni[2], Mahieddine Djoudi[3], and Hichem Rahab[4]

[1] University Center of Aflou, Algeria.
[2] LIRE Laboratory, University of Constantine 2, Algeria.
[3] TechNE Laboratory, University of Poitiers, France.
[4] ICOSI Laboratory, University of Khenchela, Algeria.* Corresponding author: Fatiha Charef f.charef@cu-aflou.edu.dz

Abstract. Users of social networks and Internet sites face numerous challenges. Problems such as fake news, satire, rumors, misinformation, misleading information, cyberbullying, spam content, offensive language, hate, and offensive speech fall under the category of harmful online content (HOC). This danger has taken advantage of social media's popularity and the abundance of news that spreads quickly, causing problems for individuals and society. Moreover, to combat this danger, researchers in the AI domain have persistently advanced and proposed novel approaches across various domains. Given the progress made in this work, choosing data to evaluate their approaches was always a challenge. Our contribution aims to identify the process and criteria for creating a high-quality dataset for HOC detection, primarily in the Arabic news domain. Therefore, we have collected a list of existing and available Arabic datasets, identified their characteristics, and determined the purpose of their creation. Researchers can use our study's results as a reference to choose an appropriate dataset for their future research.

Keywords: Harmful online content, Fake news, Offensive language, NLP, Arabic datasets

1 Introduction

The digital environment, particularly social media and online platforms, has a profound influence on individuals' everyday existence and society as a whole. It has given us great chances to communicate with individuals all over the world, have access to a wealth of knowledge, and debate a wide variety of issues. However, cyberspace suffers from several malicious activities, such as toxic language, cyberbullying, online harassment, and the dissemination of rumors and fake news. Many people spread harmful content throughout society due to the ease and speed of news and information dissemination, which can have devastating economic, social, and political consequences [45]. Fact-checking sites aim to combat the dissemination of harmful and misleading information online by

C. A. Kerrache et al. (eds.), *Proceedings of the International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)*, Advances in Intelligent Systems Research 184,
https://doi.org/10.2991/978-94-6463-496-9_29

verifying the accuracy of claims made in online content. They gather information from reliable sources and seek out primary sources whenever possible. It takes a lot of effort and time, so it is essential to have early verification systems in place. Researchers use machine learning models, deep learning approaches, and natural language processing methods to build early detection systems for HOC. To enhance the effectiveness of these models, selecting good datasets is an essential step in the detection process. Researchers have created different datasets for HOC detection, which vary in language, dimension, domain, news source, HOC types, news content, and application uses. Analysing and categorising such datasets according to these characteristics can offer practical advantages to researchers and professionals involved in HOC detection.

To select works on HOC detection, we used the Scopus database and numerous keywords, such as "harmful online content detection", "fake news detection", "hate speech detection", and "offensive language detection". We can see in the figure 1 that the number of works is constantly increasing. The Arab world, too, is not exempt from this danger. When we compare work in English and Arabic, we observe a significant discrepancy that suggests a shortage of workers in Arabic. Therefore, our work aims to collect and study datasets used for the automatic detection of harmful Arabic online content (HAOC).

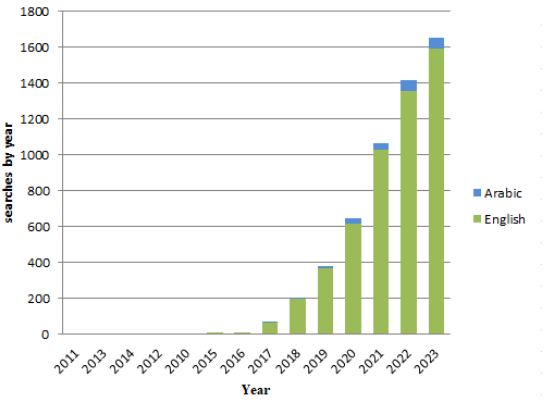

Fig. 1. Searches by year why starting at 2011

We organise the paper as follows: Section 2 delves into concepts that intersect with harmful content. Subsequently, in Section 3, we provide a discussion of the different stages that went into creating the dataset. Section 4 summarises the paper's conclusion.

2 Related Work

There are several literature reviews on HOC identification. Firstly, we start with studies conducted in English language. The authors [48,44] aim to expand the study on identifying and detecting false information on social media platforms. They achieve this by defining and characterising fake news, examining existing models that utilise various news features, compiling a list of relevant datasets, establishing performance criteria for these models, and proposing future research paths for analysing fake news on the internet. [19] compare 27 fake news detection datasets and identifies factors for selecting appropriate ones, including language, domain, source, content type, size, false information type, rating scale, application purpose, extraction period, and spontaneity. The review article from [23] highlights issues such as datasets, overfitting, and machine learning, and suggests future research to improve them to mitigate their impact on policy, economy, health, and societal stability. The study by [4] examines the effectiveness of hate speech detection systems by evaluating different datasets. The findings suggest these datasets are small and do not provide reliable data for identifying numerous forms of hate speech, including racism, sexism, and aggressive content. The authors [49] discuss recent advancements in hate speech detection and opinion mining, utilising machine learning, deep learning, and transformer models, and provide a comprehensive dataset collection. [47] The survey explores various research methods for identifying cyberbullying, dividing them into four main groups: machine learning models, lexicon-based methods, text rule-based methods, and mixed-initiative approaches. The challenges include the creation of datasets that stimulate additional tasks to identify indirect bullying, taking into account the role and type of bullying as determined by researchers. The author's survey, [33], delves into cyberbullying, its forms, detection methods, dataset analysis, preprocessing techniques, and textual analysis methodologies. It categorises approaches like machine learning and deep learning, identifying limitations and challenges. According to [11], the proliferation of HOC platforms is a significant societal issue, affecting various aspects such as spam, misinformation, hate speech, harassment, offensive language, bullying, violence, graphic content, sexual abuse, self-harm, and many others. Researchers have developed various methods to identify harmful content, yet the types of content and research efforts diverge. The study surveys existing methods and content moderation policies, suggesting future directions.

Secondly, we can mention works in Arabic language: The goal is to identify and halt the dissemination of misleading information. Particularly on Twitter, a popular platform in the Arab world, [35] reviews recent Arab efforts in truth detection, presenting steps for detecting truth in Arabic tweets and surveys on content credibility. However, research on the Arabic language in this field is now at an early stage and needs more development. [27] explores advanced NLP techniques for Arabic offensive language detection, focusing on machine learning and deep learning methods. It discusses challenges, linguistic resources, new developments, targeted offensive language types, data sources, and efficiency of classifi-

cation models. The study [10] compares seven optimizers for identifying rumors in the ArCOV19-Rumors dataset from Twitter. It found that oversampling data doesn't improve machine learning and deep learning models' efficiency. Ensemble learning, including stacking, enhances logical and realistic nature. Long Short Short-Term Memory and Bidirectional Memory LSTM with RMSprop optimizers yield best results. [21] The author reviews various techniques for identifying false information related to COVID-19 and examines datasets collected during this period in multiple languages, including Arabic. [9] The study reviews Arabic cyberbullying detection techniques, focusing on Twitter's accessibility and Arabic content. While classifiers showed good performance, imbalanced datasets were observed. Future work should explore pre-processing techniques and establish a standard Arabic CB dataset. [8] This research aims to study the applications of artificial intelligence, including machine learning, deep learning, and natural language processing, to texts to build models for detecting false information. The complexity of Arabic and the diversity of its dialects pose significant challenges when applying these models to Arabic text. [14] The study examines 49 Arabic datasets on toxic language, including their tasks, annotation methods, and potential reuse for future research.

Our study on related works for HOC detection highlights two key points: firstly, the limited number of works focusing on the Arabic language, and secondly, the absence of dedicated work for dataset building for the automatic detection of harmful Arabic online content (HAOC) and its various forms. Furthermore, there is a lack of dedicated datasets for researchers, which will serve as a reference for them to carry out their work.

3 Various Forms of HOC

Harmful Online Content (HOC) means content published on social media platforms or Internet pages that causes harm, distress, or adverse consequences to individuals or groups. It can include disseminating fake news, such as satire, hoaxes, rumors, disinformation, misinformation, spam content, and phishing, or using toxic language, such as hate speech, abusive speech, offensive language, cyberbullying, extremism, misogyny, and threats against individuals and groups. Researchers face difficulties in classifying harmful online content and identifying related concepts, which often overlap and can be understood differently [18].

- **Fake News:** False news has evolved in three key ways: volume, velocity, and variety. It has become more widespread due to the ease of posting news without verification procedures, with most fake news on social media focusing on trending events. Many terms overlap with the concept of fake news, such as satire, rumor, misinformation, cyberbullying, spam content, phishing, etc. [50]. They can be identified based on three main characteristics: authenticity, intent, and whether the information is news or not [48]. Some concepts can be defined in this paragraph.

- **Satire:** stories classified as false news aim to entertain the reader, not mislead him, and their tricks are exposed [48]. Example from the dataset [46]. الرئيس السيسي يضع صورته على العملة المصرية لرفع قيمتها, *President Sisi puts his picture on the Egyptian currency to raise its value.*

- **Hoaxes:** stories that are only meant to amuse people or deceive specific people [48].

- **Misinformation:** false information accidentally produced and disseminated without malicious intent [21]. Example from the dataset [24]. يساعد تناول الثوم على منع الإصابة بفيروس كورونا لتقوية المناعة في جسم الانسان, *Eating garlic helps prevent infection with the Coronavirus by strengthening immunity in the human body.*

- **Disinformation:** information intentionally created and disseminated, misleading information in order to harm people [21]. Example from the dataset [1]. زيارات اللبنانيين الشيعة إلى السيدة زينب تنشط بعد تقدم قوات النظام جنوب دمشق, *Lebanese Shiite visits to Ms. Zainab are active after regime forces advance south of Damascus.*

- **Rumor:** stories that do not originate from news events are characterised by uncertainty at the time of publication. Its credibility may be true, false, or unverified [48]. Example اللهم احفظ المدينة المنورة وأهلها من كل سوء ومكروه هزه ارضيه المدينه المنوره, *God save Medina, and its people from all evil. Earthquake in Medina.*

- **Propaganda:** news is crafted to sway a specific audience's emotions, viewpoints, and behaviors through dishonest methods, providing biased information for religious, political, or ideological reasons [20]. Example: قاذفات بعيدة المدى أرسلت إلى سوريا في مهمة خاصة, *Long-range bombers were sent to Syria on a special mission.*

- **Clickbait:** refers to online content that uses sensational or misleading headlines designed to attract attention and encourage visitors to click on a link to benefit from advertising revenue [44].

⋆ **Toxic Language:** under the umbrella term "toxic language" are several forms of harmful language, such as :

- **Hate Speech:** A speech targets individuals or groups based on their viewpoints, skin colour, religion, sex, race, or other distinguishing traits, which contain expressions that incite hostility or violence, lead to division among the people of the same society and threaten to destroy their lives [18]. Example from the dataset [22]. الاخوان كلاب اهل النار النعال يحاربون يرفعون شعار الاسلام الحل ظلما وعدوانا, *The Brotherhood are dogs, people of hell*

sandals they fight raising the slogan of Islam: the solution is injustice and aggression.

- **Offensive Language:** using vulgar terms and profanity without the intention of causing damage to others. Example from the dataset [13]. حرر بلدك يا خنزير, *Free your country, pig.*
- **Abusive Language:** intentional use of vulgar terms and profanity to cause damage to others. Example from the dataset [22] مخيب منضرها تفوه كلبة, *She looks bad, bitch.*
- **Cyberbullying:** An intentional act in which private information is shared, and offensive comments are posted via online media against a victim who is unable to protect herself. It is also more prevalent among teenagers. Example from the dataset [29] قليل اذب وغير محترم وطليقته ارتاحت منه ومن الحياة مع تافه نجس وليس برجل, *He was impolite and disrespectful and his ex-wife was relieved of him and of life with a petty impure person not a man.*
- **Misogyny:** using discriminatory language against women that violates them, marginalizes them, and places them at the base of society. Example from the dataset [38]. عجزت افهم بعض عقليات البنات المتخلفه, *I couldn't comprehend the mentalities of certain retarded girls.*
- **Extremism:** using political, religious, and/or social topics to segment society according to hateful ideologies.

4 Data Construction And Discussion

Our survey's goal is to analyse the steps involved in constructing a dataset and identify the challenges in obtaining a qualitative and quantitative dataset. We investigated a list of datasets on harmful Arabic content on the Internet and social networks, which will serve as a reference for researchers in this particular area of study.

4.1 Data Collection

Two primary types of digital environments can disseminate and circulate HAOC, such as public websites for news articles and social networks like Twitter, Instagram, Facebook, WhatsApp, and YouTube for posts and comments. Several techniques are available to collect data from the internet, including the use of search APIs and Python libraries for periodic scraping and the compilation of a list of relevant keywords or trending hashtags [1,15]. Additionally, the author collects real news from reliable websites such as AlJazeera, AlArabiya, CNN, BBC, and SkyNews. They can also use online fact-checking platforms like Misbar, Falsoo, Fatabayyano, and PolitiFact to verify the veracity of news [24,28]. The Yellow Press and Untrustworthy page websites are excellent places to get fake news [46,25,30,26]. Furthermore, we can construct the dataset using both automatic and manual methods to generate real and fake news. Native annotators can paraphrase the original content while maintaining the information's integrity by modifying its syntax [31,40]. An alternative method for gathering

HOC is manual collection, which is direct but time consuming and requires a great deal of effort. It also involves identifying and analysing the information that has been shared [12]. Most of the works cited in Table 1 use Twitter or public websites to collect data. Obtaining data from social networking platforms such as Facebook and Messenger has become increasingly challenging due to their policies and the various security measures they implement to protect users' privacy.

HOC impacts several domains, including politics, sports, health, business, and technology. Therefore, data collection that examines several topics is required. As we see in the table 1 the political and sport domains are the most used. Even in the case of various topics, the size of the political sub-topic is larger. This can be justified by the fact that political news is often highly polarised, which means that it can easily arouse strong emotions. This polarisation may encourage the production and dissemination of misinformation to manipulate people's opinions.

4.2 Data Language

Data language refers to the language of the news collected in the dataset, which may vary depending on the sources from which it was obtained and the authors' task targets [19]. The majority of dataset reviewers in this study use the modern Arabic standard (MSA) [1,31,40,46,24,32,30,25,28]. However, social media networks are characterised by the dissemination of news by everyone and in local dialects, which produces a large amount of information in different dialects, which poses a challenge to researchers in detecting harmful content. There are works that are interested in the local dialect [22,6], and others in the various dialects like the Middle East (Jordanian, Gulf, Syrian, Lebanese), North Africa (Egyptian, Algerian, Tunisian) [37,7,17,38,42,13,36,41,5,29,2,3]. Another challenge social platforms pose is "Arabizi" a language that transcribes Arabic phonetically using a combination of the Latin alphabet and numbers, widely used in Internet conversations and social networks. This particular obstacle has received little research or attention. [16].

4.3 Data Annotation

The annotation process plays a crucial role in the development of machine learning models. It directly influences their accuracy and performance. Accurate annotations can lead to good model performance and reliable predictions. Data annotation is a complex and costly task. It can be done manually by two to three expert annotators who are native speakers. Two annotators performed the annotation, while the third reviewed their output and resolved conflicts [2]. We can use statistical methods to evaluate the level of agreement between annotators such as Cohen's Kappa and Fleiss' Kappa [1,7,12]. On the other hand, automatic annotation is used to classify news into different categories using different machine learning algorithms [32]. Another way to annotate data is to use a single

annotator with the aim of reducing time and cost [24], The authors justified this by stating that the annotator was well-experienced with the task. However, this isn't always the case. Overall, most of the papers referenced in Table 1 choose manual annotation due to the complex specifications of the Arabic language and its variety of dialects.

4.4 News Content

Artificial intelligence (AI) algorithms that detect HOC require critical features derived from news stories, such as content, network, and user profile features. Content-based features concentrate on text, image, or video content, while social context features derive information from users, generated posts, and networks [48]. Textual features play a critical role in detecting HOC by analysing news articles' content, including their titles and main text. Natural Language Processing (NLP) is instrumental in extracting and analysing these features to identify patterns and characteristics indicative of HOC. Sentiment features are used to determine the emotional tone of a text. HOC articles often exhibit extreme sentiments to evoke strong reactions from readers. Models can identify potential HOC stories that attempt to manipulate readers' emotions by analysing the text's sentiment. Fabricated news, frequently generated for financial or political purposes, uses misleading language and sensational headlines. Evaluating emotional features to detect fake news improves the prediction model's performance [1,22,25,46]. With its rich and complex structure, text content provides a wealth of information for evaluating and training machine learning (ML) and deep learning (DL) algorithms. Which was used in most of the works mentioned in Table 1 like [39,22,31,40,6,7,17,15,32,16,34,13,2,3,29,36]. The social context of the news, such as posts, likes, shares, replies, followers, and their activities, can provide helpful information on fake news detection. Each social context is represented by a post (comment, review, reply) and the corresponding side information (metadata) [24,28].

4.5 Dataset Lengths

Typically, we use the number of news stories in the dataset to calculate its size. You can also express it in terms of the total archive's kilobytes or megabytes. The size of the data set in Table 1 varies from large to small. Content length poses various challenges in the context of machine learning (ML), deep learning (DL), and data analysis. Each model requires specific strategies for effective data handling and model training. Small datasets pose challenges in model evaluation due to underfitting and generalization, which necessitates other techniques to address this issue, such as data augmentation. On the other hand, large datasets can be more challenging due to overfitting and class imbalance. Large datasets are necessary for complex models to train effectively, while small datasets might not have sufficient details to allow for accurate training.

4.6 Balanced Dataset

A balanced dataset presents a uniform, or nearly uniform, range for the number of examples belonging to each class. Unbalanced datasets pose challenges for machine learning algorithms, particularly in terms of biassed predictions and evaluation metrics. However, various techniques exist to address these challenges and improve the performance of models trained on imbalanced data [23].

4.7 Dataset Homogeneity

The homogeneity of datasets can be determined at four levels [19]: first, homogeneity of new length: this refers to datasets containing news items of a similar length; Second, homogeneity in the news domain: this refers to having a corpus of texts that correspond to the subjects (politics, sport, economy, etc.); Third, homogeneity in the type of misinformation: this means that the texts in the corpus correspond to the type of disinformation (Satire, rumors, hate speech, etc.) Finally, application purpose homogeneity: the dataset's creation goal (such as satire detection, spam detection, rumor detection, cyberbullying detection, etc.) facilitates understanding of The circumstances.

4.8 Dataset effectiveness

The dataset's quality means improved data annotation, analysis of class distribution, sufficient dataset size, and other characteristics. The final step in constructing a dataset is to test its quality. Automatic HOC detection is a classification task that usually involves classifying textual content as harmful or no harm, depending on the harmful content type. We propose different models to assess the quality of the constructed dataset. Traditional machine learning: SVM, NB, LR, and DT; deep learning models: CNN, RNN, and LSTM; and transformer models: Bert, AraBERT and marBERT. We can conclude from this survey's work that traditional classifiers perform better with smaller datasets. Deep learning models also need large datasets. To address small-size problems, we can use augmentation or transfer learning. For unbalanced classes, we can use resampling techniques.

5 Conclusion

As social media grows in popularity, an increasing number of people are getting their news from these platforms rather than from traditional news sources. However, social media spreads various forms of harmful content, leading to grave consequences for both individual users and society at large. Building datasets for effective detection of misleading content involves different phases, and each has specific challenges. In this survey, We have included a range of available datasets for the automatic detection of harmful Arabic online content. (HAOC) and enumerated the distinctive attributes of each dataset, such as language, source,

features, annotation, size, creation date. There is a small data set dedicated to the Arabic language. This is due to the unique characteristics and challenges of the Arabic language, as well as the diversity of its dialects. Arabizi language was also ignored. We also plan to explore other datasets for another type of HOC and different Arabic dialects.

Table 1: Available datasets for harmful online content detection

N°	Dataset	Language	Source	Topics	Features	Label	Size	Date	Ref
1	***Credibility Dataset***	MSA	Twitter	Politics	26 Contents, 22 User Features,	1051 Credible, 810 Non-Credible	1862	2016	[1]
2	***ANS***	MSA	BBC, CNN, ALArabiya, SkyNews, France24	Politics, Technology, Economic, Sports, Culture	News Titles	3072 NotFake, 1475 Fake	4547	2020	[31]
3	***AraNews***	MSA	50 Newspapers	MultiTopic	17 categories (Name, Country, link) of Newspapers, Title, Text, Summary, Author, Date, Topic	Fake, NotFake	1M	2020	[40]
4	***Satirical News***	MSA	AlHudood, AlAhram AlMexici	Divers	Text	3185 fake, 3710 real	3185	2020	[46]
5	***ArCovid19 Rumors***	MSA	Twitter, Fatabyyano, Misbar, Twitter(WHO), PolitiFact, Snopes, FullFact	Covid19, Social, Politics, Sport, Entertainment, Religious	Tweets, Retweets, replies	1831 True, 1753 False, 5830 Other	9414	2020	[24]
6	***FakeNews Corpora***	MSA	Twitter	Covid19	Label, Tweet	19582 Fake, 14947 Genuine	34529	2021	[32]
7	***AFND***	MSA	134 News Websites	Divers	Title, Text, DatePub	207310 Credible, 167233 NoCredible, 232369 Undideced	606912	2022	[30]
8	***FactChecking/ Stance Corpus***	MSA	Verify, Reuters Websites	War in Syria	Text	1239 False, 1803 True	3042	2022	[25]
9	***SpamHarm***	MSA	Twitter	Various	Tweet, Label, NameUser, Location, IdTweetRep, IdUserRep, CoordinatesRepTweet,	1941 Spam, 11299 Ham	13240	2023	[28]
10	***THSAB***	Tunisian	Facebook, Youtube	Various	Comment, Class	1078 Hate, 1127 Abusive, 3834 Normal	6039	2019	[22]
11	***LHSAB***	Syrian, Lebanese	Twitter	Politic	Tweet, Class	468 Hate, 1728 Abusive,3650 Normal	5846	2019	[39]

Table 1: Available datasets for harmful online content detection

N°	Dataset	Language	Source	Topics	Features	Label	Size	Date	Ref
12	***MLMAHS***	Arabic, English, French	Twitter	Various	Directness (Direct, Indirect), Hostility, Target (Origin, Gender, Sexual Orientation, Religion, Disability, Other), Group (Individual, Other, Women, Special needs, African descent), Annotator sentiment	610 Abusive, 755 Hateful, 1151 Offensive, 615 Disrespectful, 41 Fearful, 1197 Normal	3353 Arabic	2019	[43]
13	***GHSD***	Saudi dialect	Twitter	Various	ID, Tweet, Class	2789 Hateful, 248 Abusive, 6775 Normal	9316	2020	[6]
14	***OSACT***	MSA, Various dialects	Twitter	Politic, Religion, Society, Sports, Others	Id, Tweet, Class	1915 Offensive (225 Vulgar, 506 Hate speech), 8085 Clean	10000	2020	[37]
15	***Dangerous Speech***	MSA, Various dialects	Twitter		Tweet, Label	1371 Dangerous, 3640 Safe	5011	2020	[7]
16	***MPOLD***	MSA, Various dialects	Twitter, Facebook, Youtube	Divers	Id, Platform, Comment, Majority Label (Off, Not Off), Agreement, NumOfJudgementUsed, TotalJudgement, Offensive (Vulgar, HateSpeech, None)	676 OFF, 3324 Not OFF	4000	2020	[17]
17	***ArMI***	MSA, Various dialects	Twitter	hashtags harming women	IDTweet, Tweets, Class	6006 Misogynistic, 3827 NotMisogynistic	9833	2021	[38]
18	***ArCOV19MCM, ArCOV19MLM***	MSA, Various dialects	Twitter	Covid19	IdTweet, Tweet, MultiLbel, MultiClass		6682	2022	[42]
19	***dataset, TunEL(THSAB, LHSAB, MLMAHS, OSACT)***	MSA, Various dialects	Twitter, Facebook, YouTube	Divers	Comments, Class	3850 Abusive, 6830 Hate, 12353 Normal	23033	2022	[13]
20	***Emojis Dataset***	MSA, Various dialect	Twitter	Divers	Tweets, Class	Clean 8235, Offensive 4463	12698	2023	[36]

Table 1: Available datasets for harmful online content detection

N°	Dataset	Language	Source	Topics	Features	Label	Size	Date	Ref
21	***DziriOFN***	MSA, Arabizi	Facebook	Politics, Sports	IDPost, Nucomment, Comment, OfnAbus, Arabic	3,227 offensive, 1334 abusive, 4188 Normal	8749	2021	[16]
22	***Harmful Tweet***	MSA, Various dialects	Twitter	Covid19	IDTweet, URLTweet, Tweet, Class	Harm, NotHarm	4957	2021	[41]
23	***ArMIS***	MSA, Various dialects	Twitter	hashtags harming women	Text, info annotation, Lang, Label	Misogyny, NotMisogyny	964	2022	[5]
24	***CyByDataset***	MSA, Various dialects	Twitter, Facebook		Text, Class	6000 CyBy, 6000 NoCyBy	12000	2023	[29]
25	***multilabelCyBy***	MSA, Various dialects	Instagram		IDuser, Username, URLProfile, Picture-Profile, IDComment, Text, timestamp, Label	sentiment (17376 POS, 18193 NEG, 11329 neutral), (5937 Toxic, 12256 Bullying, 17376 POS, 11329 neutral), Dialect	46898	2022	[2]
26	***AraByTweets***	MSA, Various dialects	Twitter, YouTube		Comment, Class	By, NotBy	30000	2023	[3]

References

1. Al Zaatari, A., El Ballouli, R., ELbassouni, S., El-Hajj, W., Hajj, H., Shaban, K., Habash, N., Yahya, E.: Arabic corpora for credibility analysis. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4396–4401 (2016)
2. ALBayari, R., Abdallah, S.: Instagram-based benchmark dataset for cyberbullying detection in arabic text. Data **7**(7), 83 (2022)
3. Alduailaj, A.M., Belghith, A.: Detecting arabic cyberbullying tweets using machine learning. Machine Learning and Knowledge Extraction **5**(1), 29–42 (2023)
4. Alkomah, F., Ma, X.: A literature review of textual hate speech detection methods and datasets. Information **13**(6), 273 (2022)
5. Almanea, D., Poesio, M.: Armis-the arabic misogyny and sexism corpus with annotator subjective disagreements. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 2282–2291 (2022)
6. Alshalan, R., Al-Khalifa, H.: A deep learning approach for automatic hate speech detection in the saudi twittersphere. Applied Sciences **10**(23), 8614 (2020)
7. Alshehri, A., Nagoudi, E.M.B., Abdul-Mageed, M.: Understanding and detecting dangerous speech in social media. arXiv preprint arXiv:2005.06608 (2020)
8. Alshehri, S., Alessa, N., Alhawiti, M., Majdoa, A., Aljohani, R., Aljehane, N., Alotaibi, M.: Natural language processing to detect fake news in arabic: A survey paper. In: 2023 3rd International Conference on Computing and Information Technology (ICCIT). pp. 647–651. IEEE (2023)
9. Alsunaidi, N., Aljbali, S., Yasin, Y., Aljamaan, H.: Arabic cyberbullying detection using machine learning: State of the art survey. In: Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering. pp. 499–504 (2023)
10. Amoudi, G., Albalawi, R., Baothman, F., Jamal, A., Alghamdi, H., Alhothali, A.: Arabic rumor detection: A comparative study. Alexandria Engineering Journal **61**(12), 12511–12523 (2022)
11. Arora, A., Nakov, P., Hardalov, M., Sarwar, S.M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., et al.: Detecting harmful content on online platforms: what platforms need vs. where research efforts go. ACM Computing Surveys **56**(3), 1–17 (2023)
12. Assaf, R., Saheb, M.: Dataset for arabic fake news. In: 2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT). pp. 1–4. IEEE (2021)
13. Badri, N., Kboubi, F., Habacha Chaibi, A.: Towards automatic detection of inappropriate content in multi-dialectic arabic text. In: International Conference on Computational Collective Intelligence. pp. 84–100. Springer (2022)
14. Bensalem, I., Rosso, P., Zitouni, H.: Toxic language detection: a systematic survey of arabic datasets. arXiv preprint arXiv:2312.07228 (2023)
15. Boucherit, O., Abainia, K.: Offensive language detection in under-resourced algerian dialectal arabic language. In: International Conference on Big Data, Machine Learning, and Applications. pp. 639–647. Springer (2021)
16. Boucherit, O., Abainia, K.: Offensive language detection in under-resourced algerian dialectal arabic language. In: International Conference on Big Data, Machine Learning, and Applications. pp. 639–647. Springer (2021)
17. Chowdhury, S.A., Mubarak, H., Abdelali, A., Jung, S.g., Jansen, B.J., Salminen, J.: A multi-platform arabic news comment dataset for offensive language detection.

In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 6203–6212 (2020)
18. Dehghani, M.: A comprehensive cross-language framework for harmful content detection with the aid of sentiment analysis. arXiv preprint arXiv:2403.01270 (2024)
19. D'Ulizia, A., Caschera, M.C., Ferri, F., Grifoni, P.: Fake news detection: a survey of evaluation datasets. PeerJ Computer Science **7**, e518 (2021)
20. Elhadad, M.K., Li, K.F., Gebali, F.: Fake news detection on social media: a systematic survey. In: 2019 IEEE Pacific Rim conference on communications, computers and signal processing (PACRIM). pp. 1–8. IEEE (2019)
21. Gharaibeh, M., Obeidat, R., Abdullah, M., Al-Harahsheh, Y.: Datasets and approaches of covid-19 misinformation detection: A survey. In: 2022 13th International Conference on Information and Communication Systems (ICICS). pp. 337–345. IEEE (2022)
22. Haddad, H., Mulki, H., Oueslati, A.: T-hsab: A tunisian hate speech and abusive dataset. In: International conference on Arabic language processing. pp. 251–263. Springer (2019)
23. Hamed, S.K., Ab Aziz, M.J., Yaakub, M.R.: A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. Heliyon (2023)
24. Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.: Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. arXiv preprint arXiv:2010.08768 (2020)
25. Harrag, F., Djahli, M.K.: Arabic fake news detection: A fact checking based deep learning approach. Transactions on Asian and Low-Resource Language Information Processing **21**(4), 1–34 (2022)
26. Himdi, H.T., Assiri, F.Y.: Development of classification model based on arabic textual analysis to detect fake news: Case studies. In: 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC). pp. 1–6. IEEE (2023)
27. Husain, F., Uzuner, O.: A survey of offensive language detection for the arabic language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) **20**(1), 1–44 (2021)
28. Kaddoura, S., Henno, S.: Dataset of arabic spam and ham tweets. Data in Brief **52**, 109904 (2024)
29. Khairy, M., Mahmoud, T.M., Omar, A., Abd El-Hafeez, T.: Comparative performance of ensemble machine learning for arabic cyberbullying and offensive language detection. Language Resources and Evaluation pp. 1–18 (2023)
30. Khalil, A., Jarrah, M., Aldwairi, M., Jaradat, M.: Afnd: Arabic fake news dataset for the detection and classification of articles credibility. Data in Brief **42**, 108141 (2022)
31. Khouja, J.: Stance prediction and claim verification: An arabic perspective. arXiv preprint arXiv:2005.10410 (2020)
32. Mahlous, A.R., Al-Laith, A.: Fake news detection in arabic tweets during the covid-19 pandemic. International Journal of Advanced Computer Science and Applications **12**(6), 778–788 (2021)
33. Mishra, A., Sinha, S., George, C.P.: Shielding against online harm: A survey on text analysis to prevent cyberbullying. Engineering Applications of Artificial Intelligence **133**, 108241 (2024)
34. Mohdeb, D., Laifa, M., Zerargui, F., Benzaoui, O.: Evaluating transfer learning approach for detecting arabic anti-refugee/migrant speech on social media. Aslib Journal of Information Management **74**(6), 1070–1088 (2022)

35. Mouty, R., Gazdar, A.: Survey on steps of truth detection on arabic tweets. In: 2018 21st Saudi Computer Society National Computer Conference (NCC). pp. 1–6. IEEE (2018)
36. Mubarak, H., Hassan, S., Chowdhury, S.A.: Emojis as anchors to detect arabic offensive language and hate speech. Natural Language Engineering **29**(6), 1436–1457 (2023)
37. Mubarak, H., Rashed, A., Darwish, K., Samih, Y., Abdelali, A.: Arabic offensive language on twitter: Analysis and experiments. arXiv preprint arXiv:2004.02192 (2020)
38. Mulki, H., Ghanem, B.: Armi at fire 2021: Overview of the first shared task on arabic misogyny identification. FIRE (Working Notes) pp. 820–830 (2021)
39. Mulki, H., Haddad, H., Ali, C.B., Alshabani, H.: L-hsab: A levantine twitter dataset for hate speech and abusive language. In: Proceedings of the third workshop on abusive language online. pp. 111–118 (2019)
40. Nagoudi, E.M.B., Elmadany, A., Abdul-Mageed, M., Alhindi, T., Cavusoglu, H.: Machine generation and detection of arabic manipulated and fake news. arXiv preprint arXiv:2011.03092 (2020)
41. Nakov, P., Alam, F., Shaar, S., Martino, G.D.S., Zhang, Y.: A second pandemic? analysis of fake news about covid-19 vaccines in qatar. arXiv preprint arXiv:2109.11372 (2021)
42. Obeidat, R., Gharaibeh, M., Abdullah, M., Alharahsheh, Y.: Multi-label multi-class covid-19 arabic twitter dataset with fine-grained misinformation and situational information annotations. PeerJ Computer Science **8**, e1151 (2022)
43. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., Yeung, D.Y.: Multilingual and multi-aspect hate speech analysis. arXiv preprint arXiv:1908.11049 (2019)
44. Pierri, F., Ceri, S.: False news on social media: a data-driven survey. ACM Sigmod Record **48**(2), 18–27 (2019)
45. Rahab, H., Zitouni, A., Djoudi, M.: Arabic fake news and spam handling: Methods, resources and opportunities. In: 2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP). pp. 1–7. IEEE (2021)
46. Saadany, H., Mohamed, E., Orasan, C.: Fake or real? a study of arabic satirical fake news. arXiv preprint arXiv:2011.00452 (2020)
47. Salawu, S., He, Y., Lumsden, J.: Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing **11**(1), 3–24 (2017)
48. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter **19**(1), 22–36 (2017)
49. Subramanian, M., Sathiskumar, V.E., Deepalakshmi, G., Cho, J., Manikandan, G.: A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Engineering Journal **80**, 110–121 (2023)
50. Zubiaga, A., Liakata, M., Procter, R.: Learning reporting dynamics during breaking news for rumour detection in social media. arXiv preprint arXiv:1610.07363 (2016)