Available online at https://www.asjp.cerist.dz/en/PresentationRevue/134

RIST

Information Processing at the Digital Age Journal

Volume 27. No 1 (2023)

AraCovid19-SSD: Arabic Covid-19 Sentiment And Sarcasm Detection Dataset

Mohamed Seghir Hadj Ameur[a], Hassina Aliane[a]

[a] *Centre de recherche sur l'information scientifique et technique (CERIST), Alger, Algérie*

Abstract

Coronavirus disease (COVID-19) is an infectious respiratory disease that was first discovered in late December 2019, in Wuhan, China, and then spread worldwide causing a lot of panic and death. Users of social networking sites such as Facebook and Twitter have been focused on reading, publishing, and sharing novelties, tweets, and articles regarding the newly emerging pandemic. A lot of these users often employ sarcasm to convey their intended meaning in a humoristic, funny, and indirect way making it hard for computer-based applications to automatically understand and identify their goal and the harm level that they can convey. Motivated by the emerging need for annotated datasets that tackle these kinds of problems in the context of COVID-19, this paper builds and releases AraCOVID19-SSD, a manually annotated Arabic COVID-19 sarcasm and sentiment detection dataset containing 5,162 tweets. To confirm the practical utility of the built dataset, it has been carefully analyzed and tested using several classification models.

Keywords: Arabic COVID-19 Dataset; Annotated Dataset; Sarcasm Detection; Sentiment Analysis; Social Media; Arabic Language;

1. Introduction

Covid-19 is a highly infectious respiratory disease, Lai et al., 2020 that was first identified in Wuhan, China, in late December 2019, and then declared as a global pandemic on March 2020 by the World Health Organization (WHO), Gennaro et al., 2020. Since its outbreak, governments around the world have adopted several protection measures such as closing borders, travel restrictions, quarantine, social distance and containment. As of late July 2021, COVID-19 has caused more than 170 million confirmed cases and 3 million deaths worldwide[*].

The severity of the mentioned measures along with the increased number of cases and deaths have significantly impacted people's morals causing a lot of uncertainty, grief, fear, stress, mood disturbances, and

[*] https://www.worldometers.info/coronavirus/

mental health issues, Bäuerle et al., 2020. Many people relied on social networking sites such as Facebook and Twitter to express their feelings, thoughts, and opinions by publishing and sharing content related to this new emerging pandemic. The content that they shared often employed sarcasm[†] to convey their intended meaning in a humorous, funny, and indirect way making it hard for computer based applications to automatically understand and identify their intent and the harm level that they can cause. The presence of sarcastic phrases makes the task of sentiment analysis more difficult as the intended meaning is conveyed via indirect often humorous ways. This led the research community to devote a lot of interest and attention to the task of automatic sarcasm and sentiment detection. As part of the efforts that are being made to create and share COVID-19 related datasets and tools, Shahi and Nandini, 2020, Elhadad et al., 2021, Alqurashi et al., 2021, Ameur et Aliane,2021, this paper builds and releases a manually annotated Arabic COVID-19 sarcasm and sentiment detection dataset containing 5,162 tweets. The built dataset is carefully analyzed and tested using several classification models. The main contributions of this paper can be summarized as follows:

• We collected, treated, and made available a large annotated Arabic COVID-19 Twitter sentiment and sarcasm detection dataset which can be very helpful to the research community.

• To the best of our knowledge, this is the first paper that shares an annotated dataset for both Arabic sentiment and sarcasm detection in the context of the COVID-19 pandemic.

• We compared the results of multiple bag-of-words and pre-trained transformer baselines for the two considered tasks (sentiment analysis and sarcasm detection) and reported the obtained results.

The remainder of this paper is organized as follows: Section 2 presents the sentiment and sarcasm detection research studies that have been published in the context of the COVID-19 pandemic. The details of our dataset collection, construction, and statistics are then provided in Section 3. Then, in Section 4, we present and discuss the experiments we have conducted and the results we have obtained. Finally, In Section 5, we conclude our work and highlight some possible future works.

2. Related work

In the last decade, the research studies that have been made in regards to Arabic sentiment analysis and sarcasm detection have increased significantly. As such, a large number of datasets have been built and shared to be used by the research community. Rushdi et al.,2011 presented an Arabic opinion mining dataset containing 500 movie reviews gathered from several blogs and web pages. This dataset contains the same number of positive and negative instances, 250 each. The authors used several machine learning algorithms so as to provide baseline results for their annotated dataset. Nabil et al., 2015 desribed the Arabic Social Sentiment Analysis Dataset (ASTD). It contains 10,000 Arabic tweets manually annotated with four labels: “objective”, “subjective positive”, “subjective negative”, “subjective mixed”. Their paper also presented the statistics of their constructed dataset as well as its baseline results. Al-Twairesh et al., 2016 created two Arabic sentiment lexicons using a large tweets dataset containing 2.2 million tweets. Their lexicons were generated using two methods and evaluated by using internal and external datasets. Aly et al., 2013 created an Arabic sentiment analysis dataset containing over 63,000 book reviews, each review is rated on a scale of 1 to 5 stars. They provided baseline results for their dataset by testing it on the tasks of sentiment polarity and rating classification. Abu Kwaik et al., 2020 presented an Arabic sentiment analysis dataset containing 36,000 annotated tweets. The authors employed distant supervision and self-training approaches to annotate the collected tweets. They also released 8,000 tweets that have been manually annotated as a gold standard. For the task of sarcasm detection, several datasets have been published. Karoui et al., 2017 created a sarcasm and irony dataset using political tweets from Twitter. They gather the tweets using politician names as keywords and classify them into ironic and non-ironic tweets. Their created dataset contains a total of 5,479 tweets,

[†] Sarcasm can be defined as ‘a cutting, often ironic remark intended to express contempt or ridicule’ www. thefreedictionary.com

1,733 of which are ironic and the remaining are non-ironic. Ghanem et al., 2019 created a shared task for Arabic irony detection consisting of binary classification of tweets as ironic or non-ironic. They released a dataset composed of 5,030 tweets regarding the Middle East and Maghreb regions' political events. Their tweets were composed of Modern Standard Arabic (MSA) as well as different Arabic regional dialects. Abbes et al., 2020 created an irony-detection corpus (DAICT) that includes a total of 5,358 annotated MSA and dialectal Arabic tweets. The tweets were collected on the basis of different hashtags regarding irony and sarcasm. Their classification included 3 labels: "Ironic", "Not Ironic", and "Ambiguous". Farha et al. 2020 presented "ArSarcasm", an Arabic sarcasm detection dataset built by re-annotating an existing sentiment analysis dataset. Their dataset contains 10,547 tweets, 16% of which are sarcastic. They used different baselines to test the utility of their dataset and reported the obtained results. To the best of our knowledge there are no research studies that have attempted to build a sentiment analysis and sarcasm detection dataset that is devoted to the COVID-19 pandemic, thus, we believe that our dataset will be an important addition to the efforts that are being held to make more COVID-19 related datasets available for the research community.

3. Dataset

This section first presents the "AraCOVID19-SSD"[‡] dataset, its design goals, and the different classes that it contains. Then, it explains the process of tweets collection and annotation that has been adopted and provides the dataset's statistics.

3.1. "AraCOVID19-SSD" Dataset Description

The "AraCOVID19-SSD" considers two tasks: sarcasm detection and sentiment analysis. The tasks' descriptions and their annotation details are provided in Table 1.

Table 1. "AraCOVID19-SSD" tasks, values, and their signification

Tasks	**Value**	**Explanation**
Sarcastic	Yes, No	Whether the tweet is sarcastic or not.
Sentiment	Positive, Negative, Neutral	Whether the tweet's Sentiment is Positive, Negative, or Neutral.

All the 5,162 Arabic tweets of the "AraCOVID19-SSD" dataset are annotated for the two aforementioned tasks (Table 1). A small portion of the "AraCOVID19-SSD" annotated tweets are illustrated in Table 2.

Table 2. Example of some Arabic tweets along with their respective annotations

N°	Tweet ID	Tweet Text	Tweet Annotations
1	1245735322589282309	#الله يلعن الجامعة و الله يلعن الكورونا خلص عاد	Sarcastic: **No** Sentiment: **Negative**
2	1222092494600646659	أحسن شي هالأيام تدخل مصعد مفلل ناس وتعطس فيه😭😭😭 وتقول الله يلعن الساعة اللي رحت فيها الصين😂😂	Sarcastic: **Yes** Sentiment: **Positive**
3	1330781324915970048	خايف يجيبو لقاح كورونا وما يكفي يقومو يزيدو ماي حكومتنا وبعرفا😂	Sarcastic: **Yes** Sentiment: **Positive**
4	1263493680440147975	#وزير_الصحه عندي وصفه تخلص البلاد من الكورونا خلال ساعه باذن الله ، صلوني بمبار علماء الطب اشرخ لهم ما عندي .0522878900. وصفه ناجحه ١٠٠/١٠٠.	Sarcastic: **No** Sentiment: **Neutral**

[‡] https://github.com/MohamedHadjAmeur/AraCovid19-SSD

3.2. Data collection

The first step that we followed to build the dataset was to prepare a set of keywords, then we retrieved the tweets based on those keywords. The keywords that we used were made to retrieve the largest possible number of tweets related to COVID-19, a portion of the keywords that we used are الكوفيد، كوفيد، الكورونا، كورونا، الوباء، وباء، الفيروس، فيروس.

The retrieved tweets were filtered in the following way:

• All the retweets of a given tweet were removed.

• Identical tweets that share the same textual content (when ignoring the tweets' links) were removed. This is done to ensure that the text of each considered tweet is unique.

• Very short tweets that contain less than 5 Arabic words were filtered.

• Tweets were gathered within the period spanning from December 15, 2019, and December 15, 2020.

After the filtering step, we have ended up with a total of 300k unique Arabic tweets related to the COVID-19 pandemic.

3.3. Data annotation

Due to the high cost of the annotation task, we only required each tweet to be annotated by one expert annotator. This allowed us to annotate a total of 5, 162 Arabic tweets from the 300k gathered tweets. We plan to continue annotating the remaining gathered tweets gradually according to our financial capacities. The manual annotation task was carried out by providing the annotator with the full text of the tweet, including the links, and ask him/her to read the tweet, check the tweet's links if necessary, and annotate it for each one of the 2 labels (tasks). This results in a dataset in which each tweet is labeled for each one of the 2 tasks (as shown in Table 2).

3.4. "AraCOVID19-SSD" Statistics

The statistics of our "AraCOVID19-SSD" dataset are provided in Table 3.

Table 3. Statistics about the number of annotated tweets for each task

Tasks	**Annotation Statistics**	**Total tweets**
Sarcastic	Yes:1802 , No: 3360	5162
Sentiment	Positive:1964 , Negative: 1197, Neutral: 2001	5162

As shown in Table 3, the two considered tasks contain more than 1000 instances for each one of their values, which helps train robust classification models.

4. experiments and Results

Our experiments aim at evaluating the quality of our annotated dataset and provide baseline results for the two tasks that it includes. To this end, for both sarcasm detection and sentiment analysis tasks, several deep learning and bag-of-words models were trained and tested. In the following, first, we will present the Arabic preprocessing that we performed, and the different models that we considered. Then, we will report and discuss the results obtainded.

4.1. Preprocessing

We applied a basic preprocessing to all the collected Arabic tweets, which includes:
• The removal of diacritical marks.
• The removal of elongated and repeated characters.
• Arabic characters normalization.
• The removal of links and users' references (users' notifications).
• Tweets tokenization in which punctuation, words, and numbers are separated. We note that this preprocessing has been used only when training the baseline models (Section 4.2); it has not been used for the annotation task nor in the final dataset.

4.2. Considered models

Aside from the classical bag-of-words models, pre-trained transformer models have been recently used in many NLP tasks and have continuously achieved new state-of-the-art results, Devlin et al., 2017. In the following, we will highlight both the bag-of-words and the transformer models that we considered in our experiments.

4.2.1. Transformer models

In our experiments we used three pre-trained transformer models:
• AraBERT[§] : A BERT (Bidirectional Encoder Representations from Transformers) model, Devlin et al.,2017 pretrained on 200 million Arabic MSA sentences gathered from different sources, Antoun et al., 2020.
• Multilingual BERT (mBERT)[**]: A BERT-based model pre-trained on the first 104 major Wikipedia languages[††] .
• XLM-Roberta[‡‡] : A large multi-lingual language model, trained on 2.5TB of filtered Common Crawl data, Conneau et al., 2020.

4.2.2. Bag-of-Words models

In our experiments we considered three bag-of-words models:
• Support Vector Machines (SVMs), Hearst et al., 2020 are discriminative classifiers that use maximum-margin hyperplanes (support vectors) to classify high-dimensional data into a set of predefined categories.
• Random Forests model, Breiman, 2001 is an extension to the standard decision tree, Myles et al., 2004 introduced to tackle the overfitting problem that usually occurs when a decision tree learns highly irregular patterns as a consequence of growing too deep. It constructs multiple trees from random sub-samples of the same training data. Then, the final prediction is made by averaging the predictions of all the trained trees.
• Logistic Regression Wright, 1995 is a process of modeling the probability of an outcome given an input variable. It is useful for classification problems, where the goal is to determine if a new instance fits best into a given category.

[§] https://huggingface.co/aubmindlab/bert-base-arabertv02
[**] https://huggingface.co/bert-base-multilingual-cased
[††] https://meta.wikimedia.org/wiki/List_of_Wikipedias
[‡‡] https://huggingface.co/transformers/model_doc/xlmroberta.html

4.3. Software and tools

The implementation of the different models has been done using the following libraries:

• Scikit-learn, Pedregosa et al., 2011 [§§] is a python-based machine learning library. We used it to train the bag-of-words baselines and to evaluate the performance of all the considered models.

• Flair, Akbik et al., 2018[***] is a framework for building state-of-the-art NLP models. We used it to train our classification models.

• Hugging-face-transformers, Wolf et al., 2020[†††] is a framework for building and pre-training different state-of-the-art NLP models. We used it to test our pre-trained models.

• PyTorch, Paszk et al., 2019[‡‡‡] is an open-source library designed for implementing deep neural networks. We used it as a backend for both the Hugging-face-transformers and the Flair frameworks.

4.4. Evaluation methodology

To evaluate the performances of our considered classification models, we have used a stratified 5-fold cross-validation method. This is done by randomly partitioning the instances of each one of our dataset's tasks into 5 disjoint sets of equal size. In this five-fold cross-validation, five experiments are performed, in each one, one of the five sets is selected for testing, and the remaining four are used for training. For each experiment, the weighted F-score is calculated, and finally, the average F-score for all the five experiments (the 5-folds) is reported.

4.5. Results and discussion

The results of the experiments that we have performed in regards to the tasks of sarcasm detection and sentiment analysis are provided in Table 4.

Table 1. An example of a table

		Sarcasm			**Sentiment**		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Bag Of Words models	Logistic Regression	0.9538	0.9540	0.9538	0.9069	0.91	0.9072
	Random Forest	0.9511	0.9512	0.9511	0.8994	0.9019	0.8997
	SVM	0.9597	0.9599	0.9597	0.9223	0.9243	0.9224
Transformer models	XML-Roberta	0.9361	0.9369	0.9360	0.9186	0.9211	0.9187
	BERT-Multi	0.9416	0.9431	0.94162	0.8928	0.89592	0.8930
	Arabert	0.9527	0.9535	0.9527	0.9225	0.9234	0.9226

The experiments that we have performed show that high-level classification results are achieved for both the sentiment and sarcasm detection tasks. Indeed, all the tested models surpassed 0.89 f-score, we believe that the high f-scores that have been achieved are mainly due to the richness of the dataset (the high number

[§§] https://scikit-learn.org/stable/

[***] https://github.com/flairNLP/flair

[†††] https://github.com/huggingface/transformers

[‡‡‡] https://github.com/pytorch/pytorch

of instances in each class of the considered tasks). We can also observe that the SVM model and the Arabert transformer model gave the best performance by reaching an f-score of more than 0.95 on the sarcasm detection task and more than 0.92 on the sentiment analysis task. The quality of the obtained results reflects the importance of having a large annotated dataset and confirms our adopted annotation schema's practical utility.

5. Conclusion

In this paper, we have presented "AraCOVID19-SSD" an Arabic COVID-19 sentiment analysis and sarcasm detection dataset. The dataset contains 5,162 Arabic tweets; each tweet is annotated for two tasks: sentiment analysis and sarcasm detection. All the dataset's tweets have been manually annotated and validated by human annotators. The quality of the final annotated dataset has been examined via several bag-of-words and transformer models. The considered models were trained and tested using the developed dataset and the obtained results were reported. As future work, we plan to improve the annotation method and continue enriching the annotated dataset with new tweets to keep it up-to-date with the latest events and discussions that are being shared on Twitter in regards to the COVID-19 pandemic.

References

Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J., & Hsueh, P. R. 2020. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *International journal of antimicrobial agents*, *55*(3), 105924.

Di Gennaro, F., Pizzol, D., Marotta, C., Antunes, M., Racalbuto, V., Veronese, N., & Smith, L. 2020. Coronavirus diseases (COVID-19) current status and future perspectives: a narrative review. *International journal of environmental research and public health*, *17*(8), 2690.

Bäuerle, A., Teufel, M., Musche, V., Weismüller, B., Kohler, H., Hetkamp, M., ... & Skoda, E. M. (2020). Increased generalized anxiety, depression and distress during the COVID-19 pandemic: a cross-sectional study in Germany. *Journal of Public Health*, *42*(4), 672-678.

Shahi, G. K., & Nandini, D. 2020. FakeCovid--A multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343*.

Elhadad, M. K., Li, K. F., & Gebali, F. 2021. COVID-19-FAKES: a Twitter (Arabic/English) dataset for detecting misleading information on COVID-19. In *Advances in Intelligent Networking and Collaborative Systems: The 12th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2020) 12* (pp. 256-268). Springer International Publishing.

Alqurashi, S., Alhindi, A., & Alanazi, E. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*.

Ameur, M. S. H., & Aliane, H. (2021). Aracovid19-mfh: Arabic covid-19 multi-label fake news & hate speech detection dataset. *Procedia Computer Science*, *189*, 232-241.

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. 2011. OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, *62*(10), 2045-2054.

Al-Twairesh, N., Al-Khalifa, H., & Al-Salman, A. 2016, August. Arasenti: large-scale twitter-specific Arabic sentiment lexicons. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 697-705).

Aly, M., & Atiya, A. 2013, August. Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 494-498).

Kwaik, K. A., Chatzikyriakidis, S., Dobnik, S., Saad, M., & Johansson, R. 2020, May. An Arabic tweets sentiment analysis dataset (ATSAD) using distant supervision and self-training. In *Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection* (pp. 1-8).

Karoui, J., Zitoune, F. B., & Moriceau, V. 2017. Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, *117*, 161-168.

Ghanem, B., Karoui, J., Benamara, F., Moriceau, V., & Rosso, P. 2019, December. Idat at fire2019: Overview of the track on irony detection in arabic tweets. In *Proceedings of the 11th Forum for Information Retrieval Evaluation* (pp. 10-13).

Abbes, I., Zaghouani, W., El-Hardlo, O., & Ashour, F. (2020, May). Daict: A dialectal arabic irony corpus extracted from twitter. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6265-6271).

Farha, I. A., & Magdy, W. 2020, May. From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection* (pp.

32-39).
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Antoun, W., Baly, F., & Hajj, H. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.
Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.
Breiman, L. 2001. Random forests. *Machine learning*, *45*, 5-32.
Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. 2004. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(6), 275-285.
Wright, R. E. (1995). Logistic regression.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.
Akbik, A., Blythe, D., & Vollgraf, R. 2018, August. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638-1649).
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.