Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic

Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic

Ibrahim Abu Farha[1]**, Wajdi Zaghouani**[2] **and Walid Magdy**[1,3]
[1]School of Informatics, The University of Edinburgh, Edinburgh, United Kingdom
[2]College of Humanities and Social Sciences, Hamad Bin Khalifa University, Doha, Qatar
[3]The Alan Turing Institute, London, United Kingdom
i.abufarha@ed.ac.uk
wzaghouani@hbku.edu.qa
wmagdy@inf.ed.ac.uk

Abstract

This paper provides an overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. The shared task has two subtasks: sarcasm detection (subtask 1) and sentiment analysis (subtask 2). This shared task aims to promote and bring attention to Arabic sarcasm detection, which is crucial to improve the performance in other tasks such as sentiment analysis. The dataset used in this shared task, namely ArSarcasm-v2, consists of 15,548 tweets labelled for sarcasm, sentiment and dialect. We received 27 and 22 submissions for subtasks 1 and 2 respectively. Most of the approaches relied on using and fine-tuning pre-trained language models such as AraBERT and MARBERT. The top achieved results for the sarcasm detection and sentiment analysis tasks were 0.6225 F1-score and 0.748 F_1^{PN} respectively.

1 Introduction

Work on opinion mining and subjective language analysis has been prominent in the natural language processing (NLP) field during the last two decades. One of the main tasks in this area is sentiment analysis (SA). One of the early works on SA is (Pang et al., 2002), where the authors analysed the sentiment in movie reviews. Following that, and embarked with the popularity of social media, SA became one of the popular topics in NLP. Most of the work on SA targeted English, while other languages, including Arabic, lagged behind. In the last decade, researchers on Arabic NLP started targeting SA such as the work of Abdul-Mageed et al. (2011). Since then, there have been numerous works on Arabic SA such as the works of (Abdulla et al., 2013; Alayba et al., 2018; Abdul-Mageed, 2019; Al-Smadi et al., 2019; Abu Farha and Magdy, 2021). Work on Arabic SA has been hindered by many challenges such as the large variation in dialects (Habash, 2010; Darwish et al., 2014) and the complex morphology of the language (Abdul-Mageed et al., 2011). With the advancement of work on SA, researchers started tackling the challenges affecting this task such as sarcasm (Hussein, 2018). Sarcasm can be defined as a form of verbal irony that is intended to express contempt or ridicule (Joshi et al., 2017). Sarcasm is considered one of the main challenges for SA systems since it implies expressing the opinion in an indirect way, where the intended meaning is different from the literal one (Wilson, 2006).

There have been several related works on English sarcasm detection including datasets such as the works reported in (Abercrombie and Hovy, 2016; Barbieri et al., 2014a,b; Filatova, 2012; Ghosh et al., 2015; Joshi et al., 2016; Oprea and Magdy, 2020) and detection systems such as (Rajadesingan et al., 2015; Joshi et al., 2015; Amir et al., 2016). Currently, there are few attempts to work on Arabic sarcasm. Those include the work by soukhria2017, a shared task on irony detection Ghanem et al. (2019) along with the participants' submissions and dialectal sarcasm datasets by Abbes et al. (2020); Abu Farha and Magdy (2020).

In this shared task, we offer our sarcasm and sentiment detection in Arabic task that is co-organised with the WANLP 2021 workshop on Arabic NLP. The goal of the shared task is to provide resources and encourage researchers to work on Arabic sarcasm detection. The shared task has two subtasks, sarcasm detection (subtask 1) and sentiment analysis (subtask 2). We provided the participant with a new dataset (ArSarcasm-v2), which is publicly available[1]. The dataset is annotated for sarcasm, sentiment and dialect. We received 27 submissions for subtask 1 and 22 submissions for subtask 2. This paper provides an overview of the shared task

[1] ArSarcasm-v2 is available at: http://github.com/iabufarha/ArSarcasm-v2

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 296–305
Kyiv, Ukraine (Virtual), April 19, 2021.

and the achieved results by the participants along with their approaches.

Most of the approaches used by participants were based on fine-tuning pre-trained language models. A small number of participants utilised other deep learning and conventional machine learning algorithms. The top team in the sarcasm detection task was BhamNLP (Alharbi and Lee, 2021), who achieved an F1-score of 0.6225 over the sarcastic class. While the top team in the sentiment analysis task was CS-UM6P (El Mahdaouy et al., 2021), who achieved F_1^{PN} of 0.748.

2 Related Work

Our shared task offers two subtasks on Arabic Sarcasm and sentiment detection. In the following, we discuss the literature in both tasks within the Arabic NLP community.

2.1 Arabic Sarcasm Classification

Arabic sarcasm did not receive the same degree of attention as English. The work on Arabic sarcasm detection is limited to a few works. soukhria2017 were the first to work on Arabic sarcasm/irony detection. In their work, they created a corpus of sarcastic Arabic tweets, which they collected using a set of political keywords. They filtered the sarcastic tweets using distant supervision, where they relied on some markers such as the Arabic equivalent of #sarcasm such as #سخرية, #مسخرة, #تهكم and #استهزاء. The final dataset consists of 5,479, 1,733 of which are sarcastic/ironic. They experimented with various classifiers such as SVM, Naive Bayes, Logistic Regression, Linear Regression. Random Forest was the best model, where it achieved an F1-score of 0.73. ghanem2019idat organised a shared task on Arabic sarcasm/irony detection. They prepared their dataset through collecting tweets related to different topics such as the US elections. Then they filtered out tweets that contain sarcastic hashtags, where they used the same hashtags used by soukhria2017. To prepare the final dataset, the authors sampled tweets from both the sarcastic and non-sarcastic portions, then they manually annotated them. The final dataset consists of 5,030 tweets, 2,614 of which are sarcastic. The shared task saw the participation of 18 teams. The first place was obtained by Khalifa and Hussein (2019), where they achieved an F1-score of 0.85. In their work, they relied on a set of features that include word n-grams, topic modelling features, sentiment features, statistical features and word embeddings. They experimented with multiple classifiers such as BiLSTM, Random Forest, XGBoost. Their best model was an ensemble of XGBoost, neural network and Random Forest. In a recent work by Abu Farha and Magdy (2020), the authors proposed ArSarcasm dataset for sarcasm detection, which contains around 10K tweets out of which around 1,600 are sarcastic. They presented a basic baseline that uses BiLSTM and achieved a F-score of 0.46 over the sarcastic class. Another recent study, Abbes et al. (2020) created a corpus of ironic tweets, namely DAICT. To prepare the corpus, the authors followed the same approach used by Ghanem et al. (2019) the corpus consists of 5,358 tweets distributed as follows: 4,809 sarcastic, 435 non-sarcastic and 114 labelled as ambiguous.

2.2 Arabic Sentiment Analysis

Unlike Arabic sarcasm detection, Arabic sentiment analysis (SA) has been under the researchers' radar for a while. Early work on Arabic SA such as in Abdul-Mageed et al. (2011); Abbasi et al. (2008), focused on modern standard Arabic (MSA). Since then, researchers started targeting dialectal Arabic (DA) such as the work of Mourad and Darwish (2013), where the authors introduced an expandable Arabic sentiment lexicon along with a corpus of tweets. Other datasets include the works of Kiritchenko et al. (2016); Rosenthal et al. (2017); Elmadany et al. (2018). Other works focused on proposing and comparing various approaches for Arabic SA (El-Beltagy et al., 2017; Al-Smadi et al., 2019; Abdulla et al., 2013; Alayba et al., 2018; Abu Farha and Magdy, 2019).

A recent comprehensive study by Abu Farha and Magdy (2021) provides a thorough comparative analysis of the available approaches on SA. In their work, they compared a large variety of models on three benchmark datasets. Their analysis shows that deep learning models combined with word embeddings achieve much better performance compared to classical machine learning models, such as SVMs. However, their experiments show that the utilisation of transformer-based language model achieves better results the best deep learning model architecture that uses word-embeddings. They show that using a fine-tuned AraBERT(Antoun et al., 2020) outperforms all existing classical and deep learning models on all the three benchmark datasets they examined.

Set	Sarcasm		Sentiment			Total
	Sarcastic	Non-sarcastic	Positive	Negative	Neutral	
Training	2,168	10,380	2,180	4,621	5,747	12,548
Testing	821	2,179	575	1,677	748	3,000
Total	2,989	12,559	2,577	6,298	6,495	15,548

Table 1: Statistics of training and testing sets, showing the number of examples for both sarcasm detection and sentiment analysis tasks.

Dialect	Sarcastic	Non-Sarcastic	Negative	Positive	Neutral	Total
MSA	1,523	9,362	3,986	1,890	5,009	10,885
Egypt	1,085	1,896	1,564	524	893	2,981
Gulf	214	752	411	192	363	966
Levant	152	519	312	143	216	671
Maghreb	15	30	25	6	14	45
Total	2,989	12,559	6,298	2,755	6,495	15,548

Table 2: Statistics of ArSarcasm-v2 dataset showing the distribution of sarcasm and sentiment over the dialects.

3 Dataset

The shared task provides the ArSarcasm-v2, which is a new dataset for Arabic sarcasm detection. The dataset is an extension of the original ArSarcasm dataset (Abu Farha and Magdy, 2020).

3.1 Resources

ArSarcasm-v2 uses the whole original ArSarcasm dataset (Abu Farha and Magdy, 2020) as part of its training data. The original ArSarcasm consists of 10,547 tweets, 1,682 of which are sarcastic. Additional sarcastic tweets are added to the ArSarcasm-v2 dataset from the DAICT dataset (Abbes et al., 2020), which represents a corpus of ironic/sarcastic tweets. DAICT contains 5,358 tweets, 4,809 of which are ironic/sarcastic.

Since the goal is to extend the larger ArSarcasm, and because DAICT is mostly sarcastic, a new set of random tweets were collected over the period November-December 2020. The tweets where collected using the Twitter streaming API with the language filter set to Arabic ("lang:ar"). Since sarcasm is usually present in percentage, the new tweets were used to balance out DAICT.

3.2 Annotation

For the annotation process, we used appen[2] crowdsourcing platform. ArSarcasm represents the majority portion of ArSarcasm-v2. Thus, the goal was to annotate the new portions to have similar labels to ArSarcasm. Additionally, DAICT was only annotated for sarcasm/irony, thus a new annotation was needed. To ensure consistency with ArSarcasm, we followed the same procedure and used the same guidelines to annotated the new portions. The original ArSarcasm paper defined sarcasm as *an utterance that is used to express ridicule, where the intended meaning is different from the apparent one*. Appendix A shows the guidelines (in Arabic) that have been shown to annotators.

Since DAICT is only annotated for sarcasm/irony, it was used as a pool of sarcastic examples which were balanced with the set of random Arabic tweets. A new set of 5,000 tweets, 2,500 of which are from DAICT, were annotated. The annotators were asked to provide three labels for each tweet as the following:

- **Sarcasm:** sarcastic or non-sarcastic.
- **Sentiment:** positive, negative or neutral.
- **Dialect:** Egyptian, Gulf, Levantine, Maghrebi or Modern Standard Arabic (MSA).

Only annotators with an Arab origin were allowed to participate. This was verified through their profile (usage of the Arabic language). Each tweet was annotated by at least three different annotators. The quality of annotation was monitored using a set of 100 hidden test questions that appear randomly during the task, each of those questions has the correct label for sentiment, sarcasm and

[2] https://www.appen.com/

dialect. If the performance of an annotator in these test questions dropped below 80%, this annotator is eliminated and all the labels he/she provided are also ignored. Agreement among annotators was 78.9% for sentiment, 87.3% for sarcasm and 77.0% for dialects.

3.3 Dataset Statistics

The new ArSarcasm-v2 dataset consists of 15,548 tweets, 10,547 of them were taken from the original ArSarcasm dataset while the rest (5,001 tweets) from DAICT and the new collection of tweets. These additional 5,001 tweets were split into two parts: 2,001 tweets added to the original ArSarcasm to form the set of training data of 12,548 tweets, and the remaining 3000 were used as the test set, as shown in Table 1. Each of the tweets has three labels for sarcasm, sentiment and dialect. Tables 1 and 2 show the statistics of the new dataset, where we can see that 19.2% of the data is sarcastic (2,989 tweets). Also, the annotation shows that most of the data is either in MSA or Egyptian dialect while the Maghrebi dialect is under-represented with only 45 tweets.

4 Shared Task

This section provides an overview of the shared task, the description of the subtasks and the evaluation metrics.

4.1 Tasks Description

The shared task on sarcasm detection and sentiment analysis in Arabic contains two subtasks as follows:

- **Sarcasm Detection (subtask 1)**: the goal is to identify whether a tweet is sarcastic or not.
- **Sentiment Analysis (subtask 2)**: the goal is to classify the tweet to one of the sentiment classes: positive, negative or neutral.

The data for both subtasks was provided as train/test split without a specific development set. Table 1 shows the statistics of the two sets. The training set consists of 12,548 tweets, while the testing set consists of 3,000 tweets. The participants had access to the tweets' text and the dialect label during the testing phase.

4.2 Evaluation Metrics

The main evaluation metric for subtask 1 (sarcasm detection) is the F1-score of the sarcastic class only (F1-sarcastic), since it is the main class to be detected. Sarcasm is usually present in small percentages in the data, thus the task is an imbalanced classification task. F1-sarcastic is calculated using the following equation:

$$F_1^{sarcastic} = 2 \cdot \frac{P^{sarcastic} \cdot R^{sarcastic}}{P^{sarcastic} + R^{sarcastic}}, \qquad (1)$$

Where $P^{sarcastic}$, $R^{sarcastic}$ are the precision and recall with respect to the sarcastic class.

For the sentiment analysis, the macro F1-score over the positive and negative classes was used (F_1^{PN}). It is worth noting that the neutral class is excluded from the metric calculation and not the whole task. Thus miss-classified neutral tweets will lead to the increase of false positives for the positive or negative class, and thus should lead to the reduction of the F_1^{PN} value. This metric is the main adapted measure in multiple sentiment analysis shared tasks in different languages (Kiritchenko et al., 2016; Rosenthal et al., 2017).

F_1^{PN} is calculated using the following equation:

$$F^{PN} = \frac{1}{2}(F_1^P + F_1^N), \qquad (2)$$

Where F_1^P, F_1^N are the F_1 with respect to the positive and negative classes respectively, while the neutral class is ignored.

4.3 Participating Teams

The shared task saw the participation of 30 unique teams. The sarcasm detection task (subtask 1) received 27 submissions, while the sentiment analysis task (subtask 2) received 22 submissions. Table 3 shows the list of the participating teams whose papers were accepted[3].

4.4 Shared Task Results

Tables 4 and 5 show the results of both subtask 1 and subtask 2 respectively. The results are sorted in descending order based on the official metric of the corresponding subtask, where F1-sarcastic and F_1^{PN} are the official metrics for subtask 1 and subtask 2 respectively. For each team, only the last submission was considered for the leaderboard. For subtask 1 (sarcasm detection), BhamNLP Alharbi and Lee (2021) achieved first place with an F1-sarcastic of 0.6225. For subtask 2 (sentiment analysis), CS-UM6P El Mahdaouy et al. (2021) team achieved first place with an F_1^{PN} of 0.748.

[3] We received system description papers from only 17 of the participating teams.

Team	Affiliation of the first author	Subtask(s)
AIMTechnologies	A.I.M Technologies	1, 2
ALI-B2B-AI	Alibaba Group, China	1, 2
ArabicProcessors (Gaanoun and Benelallam, 2021)	INSEA, Morocco	1, 2
BhamNLP (Alharbi and Lee, 2021)	University of Birmingham, King Abdulaziz University	1, 2
CS-UM6P (El Mahdaouy et al., 2021)	Mohammed VI Polytechnic University, Morocco	1, 2
DeepBlueAI (Song et al., 2021)	DeepBlue Technology (Shanghai) Co., Ltd, China	1, 2
DM-JUST(dalya) (Faraj and Abdullah, 2021)	Jordan University of Science and Technology, Jordan	1
Fatemah (Husain and Uzuner, 2021)	Kuwait University, Kuwait	1, 2
iCompass (Naski et al., 2021)	iCompass, Tunisia	1, 2
IDC (Israeli et al., 2021)	The Data Science Institute, Interdisciplinary Center, Israel	1, 2
ITAM	University Mohamed First, Oujda, Morocco	1, 2
Juha (Abuzayed and Al-Khalifa, 2021)	iWAN research group, Saudi Arabia	1, 2
Laila & Daliyah (Laila) (Bashmal and Alzeer, 2021)	King Saud University, Saudi Arabia	1
Naglaa Abdelhade (Naglaa)	Assiut university, Egypt	2
NAYEL (Nayel et al., 2021)	Benha University, Egypt	1, 2
Phonemer (Wadhawan, 2021)	Flipkart Private Limited	1, 2
rematchka (Abdel-Salam, 2021)	Computer Engineering, Cairo University, Egypt	1, 2
SalamBERT (Husain and Uzuner, 2021)	Kuwait University, Kuwait	1, 2
Serpente (Ghoul and Lejeune, 2021)	Sorbonne University, France	1, 2
SpeechTrans (Lichouri et al., 2021)	CRSTDLA Research Center, Algeria	1, 2
SPPU_AASM (Hengle et al., 2021)	Pune University, India	1, 2
ZTeam (Elagbry et al., 2021)	Helwan University, Egypt	1, 2

Table 3: The list of participating teams who provided their affiliation details along with the citation for those who submitted a system description paper. Runs that did not provide any details on their affiliation are not listed, but their results are listed in Tables 4 and 5.

4.5 Approaches by Top Submissions

The participating teams used a variety of approaches for both subtasks. Most of the teams used pre-trained language models such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2020). Abu Farha and Magdy (2021), provide an extensive comparison of pre-trained language models on ArSarcasm-v2 dataset. A few of the participants used deep learning and conventional machine learning approaches. All the teams, that are participating in the two subtasks, used the same architecture for both tasks.

For the sarcasm detection task, BhamNLP (Alharbi and Lee, 2021) team was ranked first with an F1-sarcastic of 0.6225. In their approach, they used a multi-task learning architecture that is trained for sarcasm and sentiment classification. The model is based on both MARBERT and a CNN-LSTM model, where the output of each of these models is concatenated and fed to the final output layer. The CNN-LSTM used both word and character embeddings. The second place (SPPU-AASM) (Hengle et al., 2021) used an ensemble of AraBERT and CNN-BiLSTM model, which achieved an F1-sarcastic of 0.614. The third place (DeepBlueAI) (Song et al., 2021) used an ensemble of AraBERT and XLM-R, which achieved an F1-sarcastic of 0.6127.

For the sentiment analysis task, CS-UM6P team El Mahdaouy et al. (2021) achieved first place with an F_1^{PN} of 0.748. Their model is based on MARBERT-based Multi-Task Learning with task attention interaction layer for sarcasm and sentiment detection. The model used MARBERT as an encoder to produce sentence embeddings. Those embeddings are fed into separate attention and output layers specific to each task. The second place was obtained by DeepBlueAI (Song et al., 2021) with an F_1^{PN} of 0.7392. They used a similar architecture to the ones used for sarcasm detection. The third place was obtained by (rematchka) Abdel-Salam (2021) fine-tuned MARBERT for sentiment classification and achieved an F_1^{PN} of 0.7321.

4.6 Other Interesting Approaches

salambert-arsarcasm built their on the hypothesis that tweets with negative sentiment and tweets with sarcasm content are more likely to have offensive content. Thus, they pre-trained AraBERT (Antoun et al., 2020) on offensive language data then fine-tuned it for the target task. In (Israeli et al., 2021), the authors filtered the data through down sampling the non-sarcastic class. Their hypothesis is that the test set would be similar to the extra portions added to the original ArSarcasm. Thus, for both ArSarcasm and the added tweets, they built a

Rank	Team	F1-sarcastic	Accuracy	Macro-F1	Precision	Recall
1	BhamNLP	0.6225	0.7700	0.7286	0.7193	0.7460
2	SPPU-AASM	0.6140	0.7410	0.7096	0.7031	0.7447
3	DeepBlueAI	0.6127	0.7830	0.7310	0.7279	0.7345
4	CS-UM6P	0.6000	0.7680	0.7183	0.7122	0.7268
5	dalya	0.5989	0.7830	0.7251	0.7268	0.7235
6	Laila	0.5968	0.7063	0.6829	0.6874	0.7337
7	Phonemer	0.5872	0.7830	0.7200	0.7264	0.7147
8	AIMTechnolgies	0.5852	0.7467	0.7014	0.6934	0.7174
9	IDC	0.5677	0.7670	0.7041	0.7062	0.7022
10	rematchka	0.5662	0.7803	0.7095	0.7231	0.7004
11	UBC	0.5468	0.7723	0.6974	0.7119	0.6880
12	SalamBERT	0.5348	0.7727	0.6922	0.7128	0.6807
13	Juha	0.5191	0.6980	0.6495	0.6443	0.6661
14	ZTeam	0.5189	0.7533	0.6765	0.6858	0.6700
15	ALI-B2B-AI	0.5139	0.7617	0.6780	0.6965	0.6678
16	ArabicProcessors	0.5086	0.7797	0.6833	0.7296	0.6665
17	MMFOUAD	0.5056	0.6917	0.6408	0.6360	0.6557
18	Fatemah	0.5041	0.7607	0.6732	0.6950	0.6622
19	Kalawy	0.4870	0.7247	0.6494	0.6514	0.6476
20	rehab88	0.4870	0.7247	0.6494	0.6514	0.6476
21	iCompass	0.4860	0.7730	0.6702	0.7195	0.6543
22	Serpente	0.4109	0.7630	0.6313	0.7116	0.6194
23	SpeechTrans	0.3371	0.7287	0.5833	0.6359	0.5802
24	AhmedAbdou	0.2542	0.7340	0.5462	0.6486	0.5569
25	ITAM	0.2509	0.7253	0.5414	0.6218	0.5517
26	NAYEL	0.2440	0.7460	0.5457	0.7048	0.5602
27	rematchka	0.1657	0.7047	0.4932	0.5497	0.5185

Table 4: Results achieved by participants in subtask 1 (sarcasm detection). The main metric is the F1-score of the sarcastic class (F1-sarcastic).

topic model for each dialect and removed irrelevant topics from ArSarcasm. Additionally, they utilised a language model to augment the data with new sarcastic examples. The augmentation was done through replacing and adding new words. Finally, they fine-tuned MARBERT model (Abdul-Mageed et al., 2020) for each dialect. Other participants used BERT models in different ways. The majority of the participants used ensemble methods, where the combined BERT-based models with other models. While most participants used the same architecture for both tasks, some participants relied on multi-task learning to train the model on both tasks simultaneously such as in Alharbi and Lee (2021); El Mahdaouy et al. (2021).

5 Conclusion and Future Directions

This paper provides an overview of the shared task on sarcasm detection and sentiment analysis in Arabic. We provide an overview of the current state of research on Arabic sarcasm. The paper provides an overview of the new ArSarcasm-v2 dataset which was used for the shared task. We also provide a high-level description of the top participating teams in the shared task. The aim of this shared task is to encourage researchers to work on Arabic sarcasm, which was reflected by the popularity of the task and having 27 run submissions and 17 system description papers discussing different approaches applied on this task.

We hope that this task would not be the last on Arabic Sarcasm detection. More datasets on Arabic Sarcasm would further help the development of better detection models. In addition, much work is still required for this challenging task, since as it is noticed, the state-of-the-art performance is 0.62 F-score, which shows that there is large room of improvement to be achieved.

Rank	Team	F_1^{PN}	Accuracy	Macro-F1	Precision	Recall
1	CS-UM6P	0.7480	0.7107	0.6625	0.6660	0.6713
2	DeepBlueAI	0.7392	0.7037	0.6570	0.6591	0.6714
3	rematchka	0.7321	0.6957	0.6587	0.6498	0.6748
4	Phonemer	0.7255	0.6983	0.6531	0.6515	0.6623
5	IDC	0.7190	0.6923	0.6446	0.6429	0.6582
6	ArabicProcessors	0.7145	0.6817	0.6439	0.6362	0.6693
7	Juha	0.7139	0.6853	0.6297	0.6362	0.6513
8	iCompass	0.7085	0.6743	0.6423	0.6393	0.6488
9	UBC	0.7081	0.6760	0.6346	0.6274	0.6452
10	SPPU-AASM	0.7073	0.6840	0.6232	0.6421	0.6388
11	BhamNLP	0.7014	0.6753	0.6296	0.6287	0.6570
12	Fatemah	0.6877	0.6630	0.6210	0.6136	0.6318
13	AIMTechnolgies	0.6850	0.6677	0.6236	0.6213	0.6263
14	ALI-B2B-AI	0.6556	0.6333	0.5955	0.5873	0.6159
15	Serpente	0.6506	0.6473	0.5784	0.5899	0.5710
16	SalamBERT	0.6259	0.6073	0.5635	0.5580	0.5813
17	ZTeam	0.6241	0.6053	0.5545	0.5578	0.5786
18	NAYEL	0.5936	0.5980	0.5291	0.5434	0.5207
19	SpeechTrans	0.5787	0.5923	0.5222	0.5321	0.5161
20	Naglaa	0.5638	0.5793	0.5158	0.5646	0.5068
21	GOF	0.4288	0.5147	0.4275	0.5764	0.4546
22	ITAM	0.3845	0.5293	0.3768	0.4054	0.3983

Table 5: Results achieved by participants in subtask 2 (sentiment analysis). The main metric is the macro average of the F1-scores of the positive and negative classes (F_1^{PN}).

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):1–34.

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. DAICT: A dialectal Arabic irony corpus extracted from Twitter. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6265–6271, Marseille, France. European Language Resources Association.

Reem Abdel-Salam. 2021. Wanlp 2021 shared-task: Towards irony and sentiment detection in arabic tweets using multi-headed-lstm-cnn-gru and marbert. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Muhammad Abdul-Mageed. 2019. Modeling arabic subjectivity and sentiment in lexical space. *Information Processing & Management*, 56(2):291–307.

Muhammad Abdul-Mageed, Mona T. Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 587–591, Stroudsburg, PA, USA. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–6.

Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Florence, Italy. Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2020. 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, pages 32–39, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for arabicsentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Ibrahim Abu Farha and Walid Magdy. 2021. A comparative study of effective approaches for arabic sentiment analysis. *Information Processing Management*, 58(2):102438.

Abeer Abuzayed and Hend Al-Khalifa. 2021. Sarcasm and sentiment detection in arabic tweets using bert-based models and data augmentation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Mohammad Al-Smadi, Bashar Talafha, Mahmoud Al-Ayyoub, and Yaser Jararweh. 2019. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *International Journal of Machine Learning and Cybernetics*, 10(8):2163–2175.

Abdulaziz M. Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. 2018. A combined cnn and lstm model for arabic sentiment analysis. In *Machine Learning and Knowledge Extraction*, pages 179–191, Cham. Springer International Publishing.

Abdullah I. Alharbi and Mark Lee. 2021. Multi-task learning using a combination of contextualised and static word embeddings for arabic sarcasm detection and sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177, Berlin, Germany. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2014a. Italian irony detection in twitter: a first approach. In *The First Italian Conference on Computational Linguistics CLiC-it*, page 28.

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014b. Modelling sarcasm in twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58.

Laila Bashmal and Daliyah H. Alzeer. 2021. Arsarcasm shared task: An ensemble bert model for sarcasm detection in arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Kareem Darwish, Walid Magdy, et al. 2014. Arabic information retrieval. *Foundations and Trends® in Information Retrieval*, 7(4):239–342.

Samhaa R. El-Beltagy, Mona El Kalamawy, and Abu Bakr Soliman. 2017. NileTMRG at SemEval-2017 task 4: Arabic sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 790–795, Vancouver, Canada. Association for Computational Linguistics.

Abdelkader El Mahdaouy, Abdellah El Mekki, Nabil El Mamoun, Kabil Essefar, Ismail Berrada, and Ahmed Khoumsi. 2021. Deep multi-task model for sarcasm detection and sentiment analysis in arabic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Hazem E. Elagbry, Shimaa Attia, Ahmed Abdel-Rahman, Ahmed Abdel-Ate, and Sandra Girgis. 2021. A contextual word embedding for arabic sarcasm detection with random forests. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

AbdelRahim A Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. Arsas: An arabic speech-act and sentiment corpus of tweets. In *OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*, page 20.

Dalya Faraj and Malak Abdullah. 2021. Sarcasmdet at sarcasm detection task 2021 in arabic using arabert pretrained model. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Lrec*, pages 392–398. Citeseer.

Kamel Gaanoun and Imade Benelallam. 2021. Sarcasm and sentiment detection in arabic language: A hybrid approach combining embeddings and rule-based features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. Idat at fire2019: Overview of the track on irony detection in arabic tweets. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 10–13.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1003–1012.

Dhaou Ghoul and Gaël Lejeune. 2021. Sarcasm and sentiment detection in arabic: investigating the interest of character-level features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Nizar Y. Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Amey Hengle, Atharva Kshirsagar, Shaily Desai, and Manisha Marathe. 2021. Combining context-free and contextualized word representations for arabic sarcasm detection and sentiment identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Fatmeh Husain and Ozlem Uzuner. 2021. Leveraging offensive language for sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330 – 338.

Abraham Israeli, Yotam Nahum, Shai Fine, and Kfir Bar. 2021. The idc system for the sentiment classification and sarcasm detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):73.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762.

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark Carman. 2016. Harnessing sequence labeling for sarcasm detection in dialogue from tv series 'friends'. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155.

M. Khalifa and Noura Hussein. 2019. Ensemble learning for irony detection in arabic tweets. In *FIRE*.

Svetlana Kiritchenko, Saif Mohammad, and Mohammad Salameh. 2016. Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016)*, pages 42–51.

Mohamed Lichouri, Mourad Abbas, Besma Benaziz, Aicha Zitouni, and Khaled Lounnas. 2021. Preprocessing solutions for detection of sarcasm and sentiment for arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 55–64.

Malek Naski, Abir Messaoudi, Hatem Haddad, Moez Ben Haj Hmida, Chayma Fourati, and Aymen Ben Elhaj Mabrouk. 2021. icompass at shared task on sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Hamada Nayel, Eslam Amer, Aya Allam, and Hanya Abdallah. 2021. Machine learning-based model for sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Silviu Oprea and Walid Magdy. 2020. isarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 97–106. ACM.

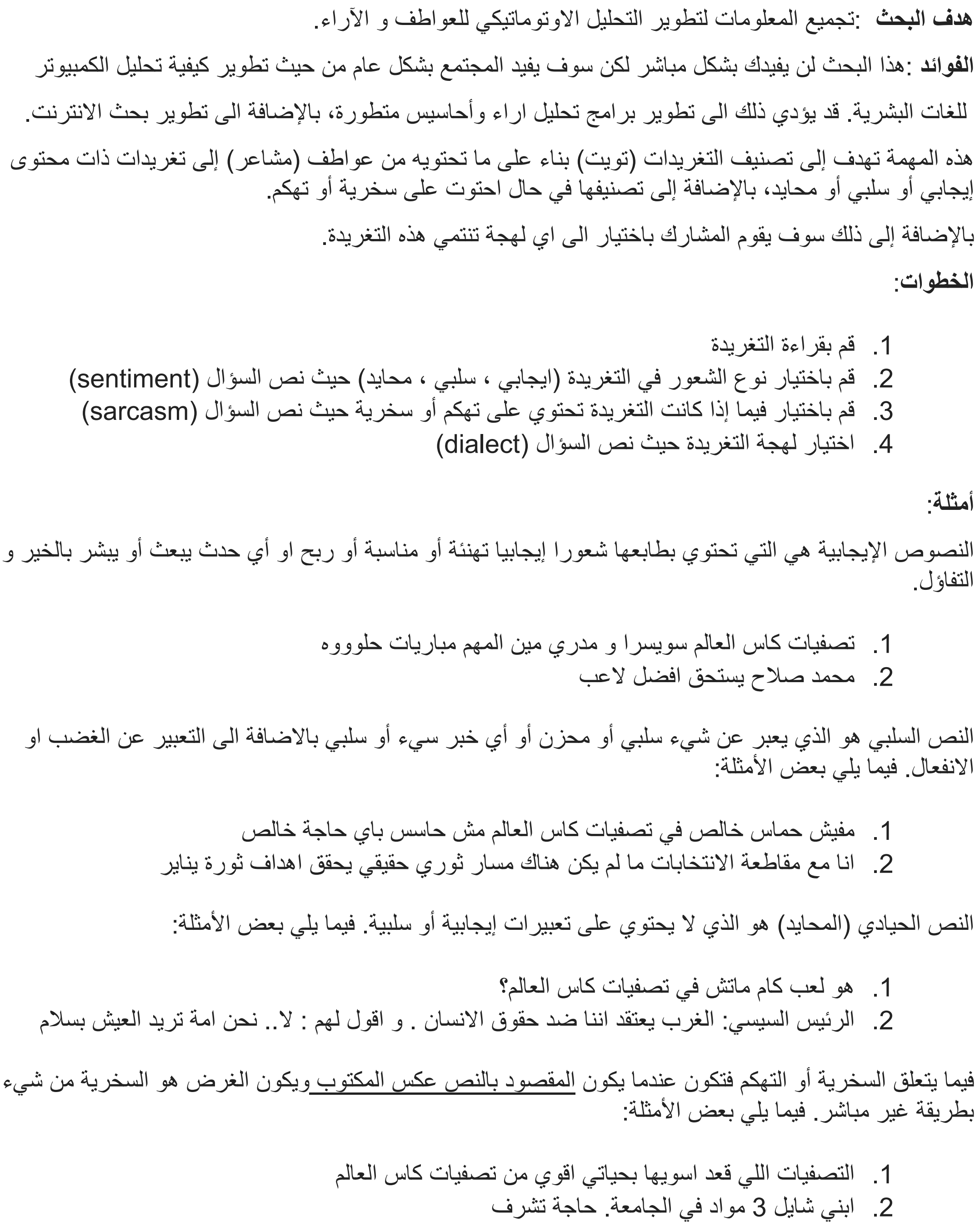
Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada. Association for Computational Linguistics.

Bingyan Song, Chunguang Pan, Wang Shengguang, and Luo Zhipeng. 2021. Deepblueai at wanlp-eacl2021 task 2: A deep ensemble-based method for sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Anshul Wadhawan. 2021. Arabert and farasa segmentation based approach for sarcasm and sentiment detection in arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722.

A Annotation Guidelines



هدف البحث :تجميع المعلومات لتطوير التحليل الاوتوماتيكي للعواطف و الآراء.

الفوائد :هذا البحث لن يفيدك بشكل مباشر لكن سوف يفيد المجتمع بشكل عام من حيث تطوير كيفية تحليل الكمبيوتر للغات البشرية. قد يؤدي ذلك الى تطوير برامج تحليل اراء وأحاسيس متطورة، بالإضافة الى تطوير بحث الانترنت.

هذه المهمة تهدف إلى تصنيف التغريدات (تويت) بناء على ما تحتويه من عواطف (مشاعر) إلى تغريدات ذات محتوى إيجابي أو سلبي أو محايد، بالإضافة إلى تصنيفها في حال احتوت على سخرية أو تهكم.

بالإضافة إلى ذلك سوف يقوم المشارك باختيار الى اي لهجة تنتمي هذه التغريدة.

الخطوات:

1. قم بقراءة التغريدة
2. قم باختيار نوع الشعور في التغريدة (ايجابي ، سلبي ، محايد) حيث نص السؤال (sentiment)
3. قم باختيار فيما إذا كانت التغريدة تحتوي على تهكم أو سخرية حيث نص السؤال (sarcasm)
4. اختيار لهجة التغريدة حيث نص السؤال (dialect)

أمثلة:

النصوص الإيجابية هي التي تحتوي بطابعها شعورا إيجابيا تهنئة أو مناسبة أو ربح او أي حدث يبعث أو يبشر بالخير و التفاؤل.

1. تصفيات كاس العالم سويسرا و مدري مين المهم مباريات حلوووه
2. محمد صلاح يستحق افضل لاعب

النص السلبي هو الذي يعبر عن شيء سلبي أو محزن أو أي خبر سيء أو سلبي بالاضافة الى التعبير عن الغضب او الانفعال. فيما يلي بعض الأمثلة:

1. مفيش حماس خالص في تصفيات كاس العالم مش حاسس باي حاجة خالص
2. انا مع مقاطعة الانتخابات ما لم يكن هناك مسار ثوري حقيقي يحقق اهداف ثورة يناير

النص الحيادي (المحايد) هو الذي لا يحتوي على تعبيرات إيجابية أو سلبية. فيما يلي بعض الأمثلة:

1. هو لعب كام ماتش في تصفيات كاس العالم؟
2. الرئيس السيسي: الغرب يعتقد اننا ضد حقوق الانسان . و اقول لهم : لا.. نحن امة تريد العيش بسلام

فيما يتعلق السخرية أو التهكم فتكون عندما يكون المقصود بالنص عكس المكتوب ويكون الغرض هو السخرية من شيء بطريقة غير مباشر. فيما يلي بعض الأمثلة:

1. التصفيات اللي قعد اسويها بحياتي اقوي من تصفيات كاس العالم
2. ابني شايل 3 مواد في الجامعة. حاجة تشرف

Figure 1: A sample of the guidelines provided to the annotators.