From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset

Citation for published version:
Abu Farha, I & Magdy, W 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.* European Language Resources Association (ELRA), pp. 32-39, The 4th Workshop on Open-Source Arabic Corpora and Processing Tools, Marseille, France, 12/05/20. <https://www.aclweb.org/anthology/2020.osact-1.5/>

Published In:
Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools

From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset

Ibrahim Abu-Farha[1] and Walid Magdy[1,2]
[1] School of Informatics, The University of Edinburgh
Edinburgh, United Kingdom
[2] The Alan Turing Institute
London, United Kingdom
i.abufarha@ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

Sarcasm is one of the main challenges for sentiment analysis systems. Its complexity comes from the expression of opinion using implicit indirect phrasing. In this paper, we present ArSarcasm, an Arabic sarcasm detection dataset, which was created through the reannotation of available Arabic sentiment analysis datasets. The dataset contains 10,547 tweets, 16% of which are sarcastic. In addition to sarcasm the data was annotated for sentiment and dialects. Our analysis shows the highly subjective nature of these tasks, which is demonstrated by the shift in sentiment labels based on annotators' biases. Experiments show the degradation of state-of-the-art sentiment analysers when faced with sarcastic content. Finally, we train a deep learning model for sarcasm detection using BiLSTM. The model achieves an F1-score of 0.46, which shows the challenging nature of the task, and should act as a basic baseline for future research on our dataset.

Keywords: Arabic, sarcasm detection, sentiment analysis

1 Introduction

Work on subjective language analysis, has been prominent in the literature during the last two decades. A major theme that dominated the area is the work on sentiment analysis (SA). According to (Liu, 2012), SA is a process where we extract and analyse the emotional polarity in a given piece of text. Large amount of work focused on classifying the text into its sentiment class, which varies based on the granularity. SA is one of the research areas within the larger natural language processing (NLP) field. The interest in SA research was embarked by the advent of user-driven platforms such as social media websites. Research on SA started with the early work of (Pang et al., 2002), where they analysed the sentiment in movie reviews. Since then, the work has developed and spanned different topics and fields such as social media analysis, computational social science and others. Most of the work is focused on English, whereas Arabic did not receive much attention until after 2010. The work on Arabic SA was kicked off by (Abdul-Mageed et al., 2011), but it still lacks behind the progress in English. This can be attributed to the many challenges of Arabic language; including the large variety in dialects (Habash, 2010; Darwish et al., 2014) and the complex morphology of the language (Abdul-Mageed et al., 2011).

As the work on SA systems developed, researchers started analysing the intricacies of such systems in order understand their performance and where they fail. There are many challenges when doing SA, such as negation handling, domain dependence, lack of world knowledge and 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 correlated with expressing the opinion in an indirect way, where the intended meaning is different from the literal one (Wilson, 2006). Additionally, sarcasm is highly context-dependent, as it always takes part between parties where shared knowledge exist. Usually, a speaker will not use sarcasm unless he/she thinks that it will be understood as so (Joshi et al., 2017).

Sarcasm detection is a crucial task for SA. The reason for this is that a sarcastic utterance usually carries a negative implicit sentiment, while it is expressed using positive expressions. This contradiction between the surface sentiment and the intended one creates a complex challenge for SA systems (Bouazizi and Ohtsuki, 2016).

There has been lots of work on English sarcasm detection, those include datasets such as the works of (Abercrombie and Hovy, 2016; Barbieri et al., 2014a; Barbieri et al., 2014b; Filatova, 2012; Ghosh et al., 2015; Joshi et al., 2016) and detection systems such as (Rajadesingan et al., 2015; Joshi et al., 2015; Amir et al., 2016).

Work on Arabic sarcasm is yet to follow. Up to our knowledge, work on Arabic sarcasm is limited to the work of (Karoui et al., 2017), a shared task on irony detection (Ghanem et al., 2019) along with the participants' submissions and a dialectal sarcasm dataset by (Abbes et al., 2020). Currently, there is no publicly available dataset for Arabic sarcasm detection. The data in (Karoui et al., 2017) is not publicly available and most of the tweets provided in (Ghanem et al., 2019) were deleted.

In this paper, we present ArSarcasm dataset, a new Arabic sarcasm detection dataset. The dataset was created using previously available Arabic SA datasets and adds sarcasm and dialect labels to them. The dataset contains 10,547 tweets, 1,682 (16%) of which are sarcastic. In addition, we analyse annotators' subjectivity regarding sentiment annotation, hoping to promote finding better procedures for collecting and annotating new datasets. The analysis shows that annotators' biases could be reflected on the annotation. Moreover, we provide an analysis of the performance of SA systems on sarcastic content. Finally, our BiLSTM based model , which serves as a baseline for this dataset, achieves

an F1-score of 0.46 on the sarcastic class, which indicates that sarcasm detection is a challenging task.
ArSarcasm is publicly available for research purposes, and it can be downloaded for free[1].

2 Background

2.1 Sarcasm and Irony Detection

The literature has a large amount of work on sarcasm and irony detection, which vary from collecting datasets to building detection systems. However, researchers and linguists cannot yet agree on a specific definition of what is considered to be sarcasm. According to (Grice et al., 1975) sarcasm is a form of figurative language where the literal meaning of words is not intended, and the opposite interpretation of the utterance is the intended one. Gibbs Jr et al. (1994) define sarcasm as a bitter and caustic from of irony. According to Merriam Webster's dictionary[2], sarcasm is "a sharp and often satirical or ironic utterance designed to cut or give pain", while irony is defined as " the use of words to express something other than and especially the opposite of the literal meaning". These definitions are quite close to each other, yet each of them gives a different definition of sarcasm. While most of the literature assumes that sarcasm is a form of irony, Justo et al. (2014) argues that it is not necessarily ironic. Thus, sarcasm is always confused with other forms of figurative language such as metaphor, irony, humour and satire.
One of the early works on English sarcasm/irony detection is the work of (Davidov et al., 2010), where the authors created a dataset from Twitter using specific hashtags such as #sarcasm and #not, which indicate sarcasm. This way of data collection is called distant supervision, where data is collected based on some specific content that it bears. Distant supervision is the most common approach to collect sarcasm data from Twitter, where the hashtag #sarcasm and others are used. Some other works that utilised distant supervision to create Twitter datasets include (Barbieri et al., 2014a; Bamman and Smith, 2015; Bouazizi and Ohtsuki, 2016; Ptáček et al., 2014). Davidov et al. (2010) mention that the use of the #sarcasm hashtag is possible but not reliable, and they used it as a search anchor. In addition, such hashtags can be useful in the cases of subtle sarcasm which might not be easily understood. Khodak et al. (2018) proposed a dataset collected from Reddit. They used a similar distant supervision approach, but they relied on "/s" marker which indicates sarcasm.
The other way to create a dataset is through manual labelling. This is done by collecting a large amount of data and asking annotators to manually label it. Works that relied on this approach include (Riloff et al., 2013; Van Hee et al., 2018). According to (Oprea and Magdy, 2019a), this approach of creating datasets captures only the sarcasm that the annotators could perceive and misses the intended sarcasm. Intended sarcasm is when the text is considered to be sarcastic by its author. In their work, they experimented with the benefits of the context in detecting perceived and intended sarcasm. In another work (Oprea and Magdy, 2019b), the authors propose a new dataset that captures intended sarcasm. They collected their data using an online survey, where they asked the participants to provide sarcastic and non-sarcastic tweets. They also asked them to provide an explanation for the sarcastic text and how would they convey the same idea in a direct way.
The work on Arabic sarcasm is scarce and limited to few attempts. It is also worth mentioning that researchers on Arabic inherited the aforementioned confusion about sarcasm definition. The earliest work on Arabic sarcasm/irony is (Karoui et al., 2017), where the authors created a corpus of Arabic tweets, which they collected using a set of political keywords. They filtered sarcastic content using distant supervision, where they used the Arabic equivalent of #sarcasm such as #سخرية, #مسخرة, #تهكم and #استهزاء. The result was a set of 5,479 tweets distributed as follows: 1,733 ironic tweets and 3,746 non-ironic. However, this corpus is not publicly available. Ghanem et al. (2019) organised a shared task competition for Arabic irony detection. They collected their data using distant supervision and used similar Arabic hashtags. In addition, they manually annotated a subset of tweets, which were sampled from ironic and non-ironic sets. The dataset provided in the shared task contained 5,030 tweets with almost 50% of them being ironic. It is worth mentioning that at the time of writing this paper around 1,300 tweets were still available. Finally, Abbes et al. (2020) proposed a dialectal Arabic irony corpus, which was also collected from Twitter.

2.2 Arabic Sentiment Analysis

In contrast to the recent attention coming to irony and sarcasm detection, Arabic SA has been under the researchers' radar for a while. There is a reasonable amount of Arabic SA resources that include corpora, lexicons and datasets.
Early work on Arabic such as (Abdul-Mageed et al., 2011; Abbasi et al., 2008), focused on modern standard Arabic (MSA). Later, attention started moving to dialects such as the work of (Mourad and Darwish, 2013), where the authors introduced an expandable Arabic sentiment lexicon along with a corpus of tweets. El-Beltagy (2016) introduced a lexicon, which contains around 6000 sentiment terms that are taken from the Egyptian dialect and MSA. The Arabic Sentiment Tweets Dataset (ASTD) (Nabil et al., 2015) contains 10,006 tweets mainly in the Egyptian dialect. It is distributed over 4 classes: positive (799), negative (1,684), neutral (832) or objective (6,691). The tweets were collected over the period between 2013 and 2015, based on the most trending topics at that time.
Elmadany et al. (2018) introduced ArSAS dataset, which is annotated for Arabic speech-act and sentiment analysis. The dataset consists of around 21K tweets, that cover multiple topics. The data was manually annotated using CrowdFlower[3] crowd-sourcing platform. The annotation scheme for the sentiment analysis task was 4-way sentiment classification, as each of the tweets is labelled with one of the following: positive (4,543), negative (7,840), neutral (7,279), or mixed (1,302). Badaro et al. (2014) introduced ArSenL, an Arabic sentiment lexicon. The lexicon was built

[1] ArSarcasm is available at: https://github.com/iabufarha/ArSarcasm
[2] https://www.merriam-webster.com
[3] Currently known as Figure-Eight

using different resources such as Arabic WordNet and English sentiment WordNet. In SemEval 2016, Arabic was included in the sentiment analysis task for multiple languages (Kiritchenko et al., 2016), where they introduced a small dataset of 1,366 tweets. In 2017, Arabic was also a part of SemEval with a larger dataset of 9,455 Arabic tweets annotated with 3 labels: positive, negative or neutral (Rosenthal et al., 2017). Other datasets and lexicons were proposed in the works of (Ibrahim et al., 2015; Refaee and Rieser, 2014; Aly and Atiya, 2013; Mahyoub et al., 2014).

3 Proposed Dataset

In this work, we present ArSarcasm, a new dataset for Arabic sarcasm detection. The dataset consists of a combination of Arabic SA datasets, where we reannotated them for sarcasm. In addition to that, we also provide labelling for the dialect and sentiment.

3.1 Resources

In this work, we relied on a set of well-known Arabic SA datasets. The reason for this choice is that sarcasm is highly subjective and always mentioned as one of the main reasons that degrades sentiment analysers' performance. The datasets we are using are SemEval's 2017 (Rosenthal et al., 2017) and ASTD (Nabil et al., 2015) datasets. ASTD dataset consists of 10,006 tweets labelled as shown in Table 1. The dataset contains tweets that date back to the period between 2013 and 2015. The tweets are mostly in Egyptian dialect and they were annotated using Amazon's Mechanical Turk. In our work, since we are aiming to annotate for sarcasm, we decided to eliminate the objective class and we took our sample from the other subjective classes.

Class	Count
Positive	799
Negative	1,684
Neutral	832
Objective	6,691
Total	10,006

Table 1: ASTD statistics.

The other dataset we are using is the one provided in SemEval's 2017 task for Arabic SA (Rosenthal et al., 2017). This dataset consists of 10,126 tweets distributed over different sets as shown in Table 2. The data was annotated using CrowdFlower[4] crowd-sourcing platform. The new dataset contains 10,543 tweets, most of which were taken from SemEval's dataset.

Set	Positive	Negative	Neutral	Total
Training	743	1,142	1,470	3,355
Validation	222	128	321	671
Testing	1,514	2,222	2,364	6,100
Total	2,479	3,492	4,155	10,126

Table 2: SemEval 2017 Task 4-A dataset statistics.

[4]Currently Figure-Eight.

3.2 Annotation

For the annotation process, we used Figure-Eight[5] crowd-sourcing platform. Our main objective was to annotate the data for **sarcasm** detection, but due to the challenges imposed by dialectal variations, we decided to add the annotation for **dialects**. We also include a new annotation for **sentiment** labels in order to have a glimpse of the variability and subjectivity between different annotators. Thus, 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).

To keep the sentiment annotation process consistent, we used the same guidelines that were used to annotate SemEval's dataset. Regarding sarcasm, we define it as *an utterance that is used to express ridicule, where the intended meaning is different from the apparent one.*
Only annotators who have Arabic language in their profiles and come from an Arab country were allowed to participate. 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 question has the correct label for sentiment, sarcasm and dialect. If the performance of an annotator in these test questions dropped below 80%, this annotator is eliminated and all the labels he provided are also ignored. Agreement among annotators was 80.7% for sentiment, 89.3% for sarcasm and 86.7% for dialects.

4 Statistics and Analysis

4.1 Dataset Statistics

The new dataset contains 10,547 tweets, 8,075 of them were taken from SemEval's dataset while the rest (2,472 tweets) were taken from ASTD. Each of the tweets has three labels for sarcasm, sentiment and dialect. Table 3 shows the statistics of the new dataset, where we can see that 16% of the data is sarcastic (1,682 tweets). The new annotation shows that most of the data is either in MSA or the Egyptian dialect, while there are few examples of the Maghrebi dialect. Figure 1 shows the ratio of sarcasm in the tweets belonging to each dialect. Maghrebi dialect has the largest percentage, but this is an outlier due to the small number of Maghrebi tweets (only 32 tweets). Thus, sarcasm is more prominent in the Egyptian dialect with 34% of the Egyptian tweets being sarcastic . Also, from the table, it is noticeable that the Egyptian dialect comprises most of the sarcastic tweets (799 tweets, 47.5% of the sarcastic tweets). Table 4 provides examples of sarcastic tweets from different dialects.

4.2 Sentiment in Sarcasm

Figure 2 shows the sentiment distribution over the sarcastic tweets. It is clear that most of the sarcastic tweets

[5]https://www.figure-eight.com/

Dialect	Non-Sarcastic	Sarcastic	Negative	Neutral	Positive	Total
Egyptian	1,584	799	1,179	733	471	2,383
Gulf	397	122	200	218	101	519
Levantine	433	118	239	178	134	551
Maghrebi	20	12	18	10	4	32
MSA	6,431	631	1,893	4,201	968	7,062
Total	8,865	1,682	3,529	5,340	1,678	10,547

Table 3: Dataset statistics for sarcasm and sentiment over the dialects.

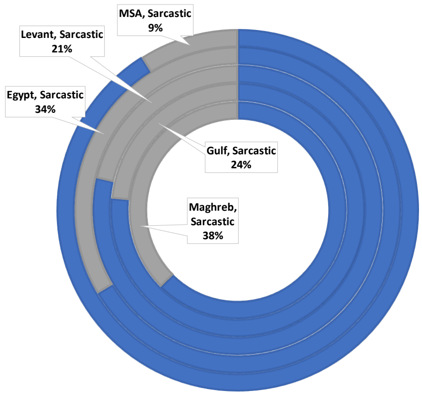

Figure 1: Ratio of sarcasm over the dialects.

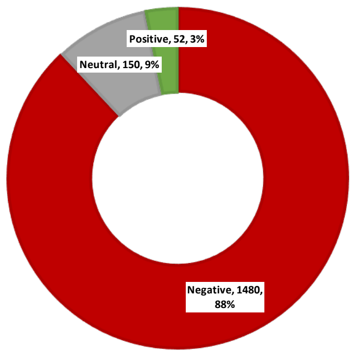

Figure 2: Sentiment distribution over the sarcastic tweets.

have negative sentiment, and this agrees with the definition we adopted, which implies that sarcasm includes making ridicule of someone or something. However, there are some neutral and positive sarcastic tweets, which could be due to the highly subjective nature of sarcasm. In addition, this could be attributed to the fact that some other metaphoric or figurative expressions might fall under the sarcasm definition. An example of that is understatement, where a person describes a good thing using negative terms such as "This was an extremely hard exam". This phenomenon is demonstrated in example 2 in Table 4, where the speaker is bragging about his success in being a presenter, and he mentions that this had happened because his mother wished him to be embarrassed and looked at as a weird person.
Table 4 provides examples of sarcastic tweets from different dialects along with their sentiment. Those examples show some aspects of the sarcasm nature, such as referencing real world items or figures. The examples show how challenging sarcasm can be, as some of them are expressed using positive expressions, yet having negative sentiment and vice versa. This, in turn, makes it extremely challenging for an SA system to analyse such examples, which urges the need for sarcasm detection systems. They also show that sarcasm relies heavily on world knowledge and context, thus incorporating such information is necessary to correctly identify sarcasm.

4.3 Annotation Subjectivity

We also studied the difference between the original and new sentiment labels. Figure 3 shows how the new labels are different from the original ones, labels above the charts are the original ones. It is clear that there is an extreme change

ID	Tweet	Sentiment	Dialect
1	كنت أعتقد أن خدمة غوغل ترجمة سيئة جدا إلى أن جربت بينغ. ملك جمال الترجمة غوغل (I was thinking that Google translate is bad, till I tried Bing. Google is Mr. Translation)	Negative	MSA
2	واضح إن أمي دعت عِليا وأنا صغير وقالتلي روح ربنا يفرج عليك خلقو، قام ربنا طلعني مذيع (It is clear that my mother was mad at me and wished that I get embarrassed and looked at by people, Now I am a TV presenter)	Positive	Egyptian
3	بالصيفيات الحلوه محد يقرر ينزلني على لبنان لما وصلت درجه الحراره تحت الصفر امي تقول نفكر نروح لا شكرا (When it is summer, no one suggests going to Lebanon. Now, when it is below zero, my mother considers going there. No, thanks)	Negative	Levantine
4	الناس المؤمنين بالسحر كان لازم نوضحلهم ان هاري بوتر مو فلم وثائقي (We should have explained for those who believe in magic that Harry Potter is not a documentary)	Negative	Gulf

Table 4: Examples of some sarcastic tweets from different dialects.

in the labels. This is empirical proof of the highly subjective nature of sentiment analysis annotation. We can see that in the case of the positive class, more the 50% of the labels has been changed, Table 5 provides examples of these cases. From the table, it is noticeable that these cases can be attributed to different reasons. For example, in the second tweet, the original annotator failed to perceive the sarcasm intended by the author. This can be due to either a misunderstanding of the intentions, or a mismatch between the author's intention and the annotator's preference. The other reason that might have caused the labels to change

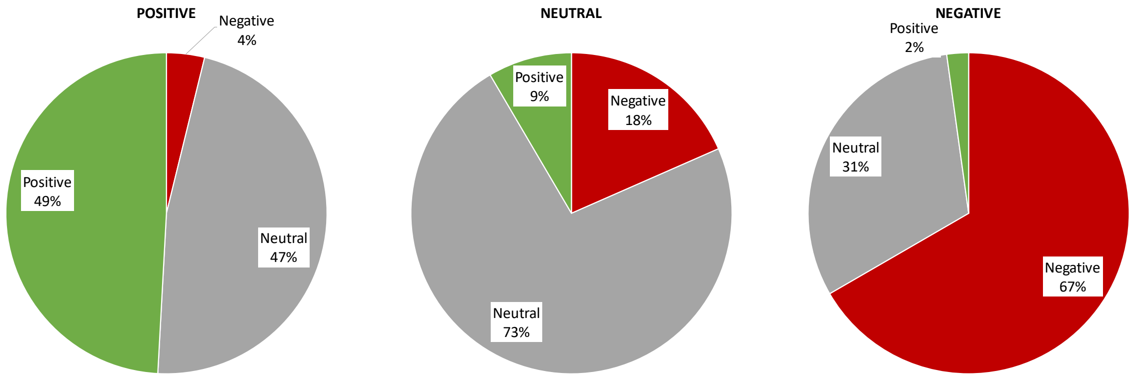

Figure 3: The change in sentiment labels between the original and new annotation. The labels above the charts are the original labels.

is the different perspectives that a text can been looked at from. For example, some annotators might annotate news as neutral, considering the view of the news agency, while others might reflect their own preference. The same thing occurs if the text is about two conflicting parties, where the annotators are likely to take one side. In addition to that, the available Arabic SA datasets are highly political and they contain different dividing topics. Having all of these factors together would result in the high presence of the annotator's biases and personal views.

Moreover, in the case of most sentiment and sarcasm datasets, they were annotated using crowd-sourcing platforms. These platforms provide multiple annotations for each data point, but they do not ensure having the same annotators to annotate all the data. This would provide inconsistent labels for the subjective text, where different conflicting biases are reflected on the assigned label. Thus, having multiple people annotating a dataset would probably give conflicting labels for different related instances within the data. These phenomena impose challenges for sentiment analysis systems, since the boundaries between the labels are not clear.

Based on the previous statistics and examples, we can see that the current annotation schemes and procedures are not robust enough against bias, and they do not ensure the consistency among different annotators. In addition, the current approach of considering sarcasm as binary text classification problem is not precise. Sarcasm is highly related to the context, cultural background, world knowledge and personal traits of its author. We believe that more sophisticated data collection and annotation approaches should be used to have a proper computational representation of sarcasm.

5 Effect of Sarcasm on Sentiment Analysis

To better understand how sarcasm can be disruptive for SA systems, we conducted an experiment on the newly annotated data. This was done through comparing the performance of an available SA system on both sarcastic and non-sarcastic tweets. In this experiment, we used Mazajak (Abu Farha and Magdy, 2019), state-of-the-art Arabic sentiment analyser. In order to have an informative comparison, we separated the dataset into two sets, sarcastic

ID	Tweet	Original label	New label
1	جوجل تنافس أبل وسامسونج بهاتف جديد (Google is competing Apple and Samsung with a new phone)	Positive	Neutral
2	مبروك عليك ويندوز ١٠ .. ربنا يخلهوللللللك (Congratulations on Windows 10, God keeeeep it for you)	Positive	Negative
3	اخس مشغلين اغنيه جستن بيبر (Shame, they are playing a Justin Bieber song)	Neutral	Negative
4	سيتم الرد علي حضرتك في اقرب وقت يا فندم. (Sir, we will respond to you soon)	Neutral	Positive
5	سمعة أبل على المحك.. مشكلة حقيقية في آيفون ٧ (Apple's reputation is on the line ... A real problem in iPhone 7)	Negative	Neutral
6	طقس كاذب يقولو ثلوج ويطلع حر (deceitful weather, they say it will snow and it is warm)	Negative	Positive

Table 5: Examples of some tweet that have its labels changed.

(1,682) and non-sarcastic (8,865). The performance was compared using the original and new sentiment labels. Table 6 shows the achieved macro F1-score. It is clear that there is a gap between the performance on sarcastic and non-sarcastic. Mazajak achieved F1-scores of 0.43 (new labels) and 0.44 (original labels) on sarcastic tweets, and F1-scores of 0.64 (new labels) and 0.61 (original labels) on the non-sarcastic ones.

Although Mazajak was trained on samples from the same dataset, the results on the sarcastic tweets are much lower than those on the non-sarcastic ones. The low performance on the sarcastic tweets indicates that SA systems rely mostly on the surface sentiment expressed by the words. This, in turn, means that sarcasm, which is an indirect implicit expression tool, is a major challenge for SA systems.

Set	F-score (new)	F-score (original)
Sarcastic	0.43	0.44
Non-Sarcastic	0.64	0.61

Table 6: Mazajak's performance on sarcastic and non-sarcastic tweets. The references are the original and the new sentiment labels.

6 Sarcasm Detection Baseline System

In this section, we conduct an experiment to set a baseline system for the new dataset. We tested a deep learning model, which consists of a bidirectional long short-term memory (BiLSTM) followed by a fully connected layer. We used the hyper-paremeters shown in Table 7. For text representation, we utilised the embeddings provided by (Abu Farha and Magdy, 2019).

#LSTM cells	128
Recurrent dropout	0.2
Dropout	0.2
#Hidden units	64
Activation	ReLU
Optimiser	Adam
Learning rate	0.0001
Batch size	512

Table 7: Hyper-parameters used for BiLSTM model.

The data was divided using an 80/20 split to create training and testing sets. Table 8 shows the results achieved by the model on the sarcastic class. As shown, the system detected sarcasm with precision 62%, but quite low recall of only 38%, which demonstrates that it is not straightforward to spot sarcasm. The overall F1-score is 0.46, which empirically proves that sarcasm detection is a challenging task that requires additional investigation. An example of that is the use of contextual information alongside the text itself, which proved to be effective in English sarcasm detection (Oprea and Magdy, 2019a).

Metric	**Result**
Precision	0.62
Recall	0.38
F1-score	0.46

Table 8: Baseline results on the sarcastic class.

From the previous experiment, we conclude that sarcasm is a challenging task, and it relies heavily on the context, world knowledge and cultural background. Thus, having better performance or good detection systems relies heavily on how these aspects are incorporated into the training and preparation of these systems (Oprea and Magdy, 2019b).

7 Conclusion and Future Work

Sarcasm is an important aspect of any language. It includes expressing ideas, opinions and emotions in an indirect implicit way. This nature of implicitness makes sarcasm problematic for SA systems which mostly rely on the surface meaning/features.

In this work, we presented ArSarcasm, a new Arabic sarcasm dataset. The dataset was created through the re-annotation of available Arabic sentiment datasets. The new dataset contains sarcasm, sentiment and dialect labels. Analysis shows that sarcasm is highly prominent in sentiment datasets with 16% of them being sarcastic. We also show the high subjective nature of such datasets, which was demonstrated by the change in sentiment labels in the new annotation. The experiments show the gap between SA systems' performance on non-sarcastic tweets compared to sarcastic tweets, which urges the need to study such phenomena. Finally, our initial experiments on sarcasm detection show that it is a challenging task.

We believe that this dataset is a starting point in the direction of full study of sarcasm and figurative language in Arabic. However, due to the highly subjective nature of sarcasm, its reliance on world knowledge, cultural background and the perspectives of the communication parties, we believe that the data collection procedure should incorporate more signals about these information. In the future, we hope to prepare a new dataset that incorporates more textual information. We also hope to study and analyse the differences and similarities among sarcastic expressions used by Arabic speakers in different countries.

Acknowledgements

This work was supported by the D&S Programme of The Alan Turing Institute under the EPSRC grant EP/N510129/1.

8 References

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

Abbes, I., Zaghouani, W., and El-Hardlo, O. (2020). Daict: A dialectal arabic irony corpus extracted from twitter. *LREC 2020*.

Abdul-Mageed, M., Diab, M. T., and Korayem, M. (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.

Abercrombie, G. and Hovy, D. (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.

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

Aly, M. and Atiya, A. (2013). 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)*, pages 494–498.

Amir, S., Wallace, B. C., Lyu, H., Carvalho, P., and Silva, M. J. (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, August. Association for Computational Linguistics.

Badaro, G., Baly, R., Hajj, H., Habash, N., and El-Hajj, W. (2014). A large scale arabic sentiment lexicon for arabic opinion mining. In *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*, pages 165–173.

Bamman, D. and Smith, N. A. (2015). Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.

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

Barbieri, F., Saggion, H., and Ronzano, F. (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.

Bouazizi, M. and Ohtsuki, T. O. (2016). A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488.

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

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics.

El-Beltagy, S. R. (2016). Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic. In *LREC*.

Elmadany, A. A., Mubarak, H., and Magdy, W. (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.

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

Ghanem, B., Karoui, J., Benamara, F., Moriceau, V., and Rosso, P. (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.

Ghosh, D., Guo, W., and Muresan, S. (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.

Gibbs Jr, R. W., Gibbs, R. W., and Gibbs, J. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

Grice, H. P., Cole, P., and Morgan, J. L. (1975). Syntax and semantics.

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

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330 – 338.

Ibrahim, H. S., Abdou, S. M., and Gheith, M. (2015). Mika: A tagged corpus for modern standard arabic and colloquial sentiment analysis. In *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, pages 353–358. IEEE.

Joshi, A., Sharma, V., and Bhattacharyya, P. (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.

Joshi, A., Tripathi, V., Bhattacharyya, P., and Carman, M. (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.

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

Justo, R., Corcoran, T., Lukin, S. M., Walker, M., and Torres, M. I. (2014). Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133.

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

Khodak, M., Saunshi, N., and Vodrahalli, K. (2018). A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kiritchenko, S., Mohammad, S., and Salameh, M. (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.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Mahyoub, F. H., Siddiqui, M. A., and Dahab, M. Y. (2014). Building an arabic sentiment lexicon using semi-supervised learning. *Journal of King Saud University - Computer and Information Sciences*, 26(4):417 – 424. Special Issue on Arabic NLP.

Mourad, A. and Darwish, K. (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.

Nabil, M., Aly, M., and Atiya, A. (2015). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

Oprea, S. and Magdy, W. (2019a). Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy, July. Association for Computational Linguistics.

Oprea, S. and Magdy, W. (2019b). isarcasm: A dataset of intended sarcasm. *arXiv preprint arXiv:1911.03123*.

Pang, B., Lee, L., and Vaithyanathan, S. (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.

Ptáček, T., Habernal, I., and Hong, J. (2014). Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223.

Rajadesingan, A., Zafarani, R., and Liu, H. (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.

Refaee, E. and Rieser, V. (2014). An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714.

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

Van Hee, C., Lefever, E., and Hoste, V. (2018). Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

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