The Effect of Arabic Dialect Familiarity on Data Annotation

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

Abstract

Data annotation is the foundation of most natural language processing (NLP) tasks. However, data annotation is complex and there is often no specific correct label, especially in subjective tasks. Data annotation is affected by the annotators' ability to understand the provided data. In the case of Arabic, this is important due to the large dialectal variety. In this paper, we analyse how Arabic speakers understand other dialects in written text. Also, we analyse the effect of dialect familiarity on the quality of data annotation, focusing on Arabic sarcasm detection. This is done by collecting third-party labels and comparing them to high-quality first-party labels. Our analysis shows that annotators tend to better identify their own dialect and they are prone to confuse dialects they are unfamiliar with. For task labels, annotators tend to perform better on their dialect or dialects they are familiar with. Finally, females tend to perform better than males on the sarcasm detection task. We suggest that to guarantee high-quality labels, researchers should recruit native dialect speakers for annotation.

1 Introduction

Many natural language processing (NLP) tasks rely on training machine learning (ML) models on labelled data. The labels are assigned in different approaches, amongst the most common ones is human annotation. These labels are sometimes highly subjective and might be affected by annotators' backgrounds and beliefs. Such subjectivity would have minimal effects for objective tasks where people have consensus (Plank et al., 2014). However, these differences can be disruptive when considering subjective tasks such as sentiment analysis (Medhat et al., 2014; Abu Farha and Magdy, 2021), sarcasm detection (Abu Farha et al., 2022a), hate speech (MacAvaney et al., 2019) and many others. This applies to all languages, but for Arabic, it is more important due to the large dialectal variety amongst Arab annotators. Arabic has three variants; the first is classical Arabic (CA), which is the language of Quran and early literature. The second is modern standard Arabic (MSA), which is standardized and mainly used in news and books. The third is dialectal Arabic (DA), which is the colloquial language spoken in everyday life and it varies from one region to another. DA differs from MSA in the sense that these dialects are not standardized. Arabic dialects substantially differ from MSA and each other in terms of phonology, morphology, lexical choice and syntax (Habash, 2010). These variations affect how speakers of different dialects understand each other; where some words or maybe complete sentences can be incomprehensible.

Previous works on Arabic dialects focused on dialect identification either in text or speech such as the works of (Zaidan and Callison-Burch, 2014; Elfardy et al., 2014; Bouamor et al., 2014; Salameh et al., 2018; Elaraby and Abdul-Mageed, 2018; Bouamor et al., 2019; Abdul-Mageed et al., 2020, 2021). Other works focused on higher level tasks exploiting dialectal data such as sentiment analysis (Abdul-Mageed et al., 2014), emotion (Alhuzali et al., 2018), offensive language (Mubarak et al., 2020), and sarcasm (Abu Farha and Magdy, 2020). Most of these datasets are created through manual data annotation. Those annotations are collected by either recruiting designated annotators or through crowd-sourcing platforms. Especially in the case of crowd-sourced annotations, the annotators are usually from different regions and speak different dialects.

In this paper, we argue that dataset creators should take into consideration the effects of annotators' native dialect and dialect familiarity on the annotation process. Due to the differences between Arabic dialects, annotators might be assigning inaccurate labels to texts written in dialects they do not fully understand. In our work, we aim to analyse

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 399 - 408

how a speaker of one dialect understands another. Also, we study the effect of dialect familiarity on the data annotation process, taking Arabic sarcasm as a case study of a highly subjective task.

In our paper, we investigate the following research questions:

- **RQ1**: How do speakers of different dialects understand text written in other dialects?
- **RQ2**: How do speakers of different dialects perform on the sarcasm detection task?
- **RQ3**: Is there a correlation between gender and the performance of an annotator on the sarcasm detection task?

In this paper, we answer these questions through collecting third-party annotations for SemEval's 2022 task 6 (iSarcasmEval) dataset (Abu Farha et al., 2022a). This dataset has first-party labels for both sarcasm and dialect, where the text authors provided the labels. Thus, we argue that those labels are of a higher quality compared to traditional third-party labels. In our work, we collect both sarcasm and dialect labels from third-party annotators, and we analyse the variation of performance based on annotators' mother dialect, familiarity with other dialects, and gender. Our analysis shows that: (1) annotators tend to better understand and identify their own dialect; (2) annotators are prone to confuse dialects with each other; (3) Egyptian dialect and MSA are the easiest to identify in written text; (4) sarcasm annotations are more trustworthy if they are provided by native dialect speakers; and (5) females tend to perform better than males on the sarcasm detection task. We hope that our findings in this study would work as guidelines for future work on labelling Arabic datasets. Data used for this work with all labels are made publicly available[1].

2 Related Work

2.1 Data Annotation and Subjectivity

Most NLP applications rely on manually annotated data. These annotations are collected from annotators from different cultures and backgrounds. Previous works acknowledged the effects of subjectivity on the quality of datasets. However, the literature lacks in-depth analyses or attempts to mitigate this issue. (Rottger et al., 2022) tried to approach this issue through suggesting new paradigms for data annotation. In their work, they suggest that dataset creators follow either descriptive or the prescriptive paradigm. Descriptive paradigm encourages annotator subjectivity, whereas prescriptive paradigm discourages it. They also argue that dataset creators should explicitly aim for one or the other. For Arabic, dialect intelligibility and understanding can be one of the subjective factors affecting the data annotation process. The literature of Arabic NLP lacks in-depth analyses on the effects of dialect familiarity on the quality of data annotations or how people understand different dialects. Habash et al. (2008) approached the dialectal variety focusing on creating standard annotation guidelines identifying dialect switching between MSA and at least one dialect. Zaidan and Callison-Burch (2014) mentioned that annotators tend to over-identify their dialect. We add to this line of work by exploring how annotators understand different dialects. We also analyse the quality of their labels on one of the most subjective tasks, sarcasm detection.

2.2 Dialectal Arabic NLP

One of the major challenges when studying dialectal Arabic (DA) was the lack of resources. For this reason, early works focused on creating resources that cover a few regions or countries (Jarrar et al., 2017; Khalifa et al., 2016; Sadat et al., 2014; Harrat et al., 2014; Al-Twairesh et al., 2018), while others focused on creating multi-dialect resources (Zaidan and Callison-Burch, 2011; Elfardy et al., 2014; Bouamor et al., 2014; Mubarak and Darwish, 2014; Cotterell and Callison-Burch, 2014) In addition, some previous works on Arabic dialects focused on dialect identification either in text or speech (Zaidan and Callison-Burch, 2014; Salameh et al., 2018; Abdul-Mageed et al., 2021, 2020; Bouamor et al., 2019; Elaraby and Abdul-Mageed, 2018; Elfardy et al., 2014; Bouamor et al., 2014).

Most of the works targeted the five major Arabic dialects: Egyptian (Nile Basin), Levantine, North African (Maghrebi), Gulf, and modern standard Arabic (MSA). However, in recent years, there has been an interest in a more fine-grained categorisation. Some of the significant works in this area are NADI shared tasks (Abdul-Mageed et al., 2020, 2021). The organisers provided data annotated on country and provenance levels, covering 21 countries and 100 provenances. Other works focused on higher level tasks exploiting dialectal data such as sentiment analysis (Abdul-Mageed et al., 2014),

[1] https://github.com/iabufarha/arabic-dialect-familiarity

emotion (Alhuzali et al., 2018), offensive language (Mubarak et al., 2020), and sarcasm (Abu Farha and Magdy, 2020). Most of the multi-dialectal resources were annotated either by designated annotators or crowd-sourced annotations. In most cases, annotators' familiarity with the dialects at hand is not taken into consideration. In our work, we aim to show that such information is necessary and should be one of the considerations when creating dialectal resources.

2.3 Sarcasm Detection

Sarcasm is a form of verbal irony that is often used to express ridicule or contempt. It is usually correlated with expressing an opinion in an indirect way where there would be a discrepancy between the literal and intended meaning of an utterance (Wilson, 2006). Sarcasm is one of the most subjective tasks that relies heavily on cultural references and the cultural background of the author. To understand sarcasm, a person needs to understand the context in which it is used, and language/dialect is part of that (Oprea and Magdy, 2019; Abercrombie and Hovy, 2016; Wallace et al., 2014). Most of previous work on sarcasm detection falls into one of two branches: creating datasets (Ptáček et al., 2014; Khodak et al., 2018; Barbieri et al., 2014; Filatova, 2012; Riloff et al., 2013; Abercrombie and Hovy, 2016; Oprea and Magdy, 2020a; Abu Farha and Magdy, 2020; Abu Farha et al., 2021) or creating detection models (Campbell and Katz, 2012; Riloff et al., 2013; Joshi et al., 2016; Wallace et al., 2015; Rajadesingan et al., 2015; Bamman and Smith, 2015; Amir et al., 2016; Hazarika et al., 2018; Oprea and Magdy, 2019). A few works focused on analysing the effect of including context in sarcasm detection models (Oprea and Magdy, 2019; Abercrombie and Hovy, 2016; Wallace et al., 2014). Wallace et al. (2014) showed that annotators tend to need context to provide judgements about ironic content. They showed that there is a correlation between that and the misclassified cases. Oprea and Magdy (2019) explored the effect of contextual information to detect sarcasm, and Oprea and Magdy (2020b) analysed the effect of cultural background and age on sarcasm understanding. Their analysis indicates that age, English language nativeness, and country are significantly influential on sarcasm understanding and should be considered in the design of sarcasm detection systems. Similar results were confirmed in the case of spoken sarcasm, where Puhacheuskaya and Järvikivi (2022) found that having a foreign accent had a negative impact on irony understanding.

Recently, Abu Farha et al. (2022b) compared human and machine performance on sarcasm detection for both English and Arabic. In their work, they compared human and machine performance on iSarcasmEval's dataset (Abu Farha et al., 2022a), a first-party annotated sarcasm dataset, where labels were provided by the authors of text themselves. Their analysis shows that sarcasm detection is challenging for humans, who perform nearly as well as state-of-the-art models. They also analysed error patterns for both humans and machine models. Based on their analysis they suggest avoiding third-party annotations for subjective tasks, building models and datasets that are better able to represent and utilise contextual information, and building better representations for proverbs and idioms which are heavily used to express sarcasm.

Our study adds to this line of work by focusing on Arabic and its dialects. In our work, we study how dialectal variation and familiarity affect human's ability to understand sarcasm.

3 Methodology

In this section, we describe our methodology for the analysis of dialects comprehension during data annotation tasks. We initially discuss the dataset we used and its ground-truth labels. Then we explain collecting third-party labels from annotators of different dialects, which will be compared later to the ground-truth labels for the analysis process.

3.1 Dataset

We use SemEval-2022 Task 6, iSarcasmEval, datasets (Abu Farha et al., 2022a). The shared-task includes three subtasks: sarcasm detection (subtask A), sarcasm category classification (subtask B), and pairwise sarcasm identification (subtask C). Subtasks A and C cover both English and Arabic, while subtask B is English only. The reason we chose iSarcasmEval's dataset is that the labels were provided by the authors themselves, which would make them more reliable than if they were provided by third-party annotators. For this work, we use the test set of Arabic subtask A (sarcasm detection). The test set consists of 1400 sentences, 200 of which are sarcastic and 1200 non-sarcastic. Each of the sentences has two labels provided by the author of the sentence: the dialect of the sentence (out of five dialects) and whether the sentence

is meant to be sarcastic or not. Table 1 shows the statistics over the available dialects.

Dialect	Total	Sarcastic	Non-sarcastic
Nile Basin	520	131	389
MSA	482	16	466
Gulf	176	10	166
Levant	168	22	146
Maghreb	54	21	33

Table 1: Distribution of the dataset over the dialects.

3.2 Third-party Annotations

To analyse the performance of speakers of different dialects, we collected third-party annotations using Appen[2] platform. For each sentence, we collected *five annotations*. We allowed only native Arabic speakers to participate. Before starting the annotation process, each annotator is presented with test questions and only those who answer all the questions correctly would be allowed to participate in the annotation process. The test questions were sampled from a set of sentences that are clearly sarcastic/non-sarcastic. We used this approach to make sure that the annotators are not giving random answers and to avoid introducing any bias before the annotation. For each sentence, we asked annotators to provide the following:

- Sarcasm label indicating whether the text is sarcastic or not.
- Dialect label out of five: MSA, Egyptian (Nile), Gulf, Levantine, and Maghrebi.
- Mother dialect, which is the dialect the annotator grew up speaking.
- Known dialects, which are the dialects the annotator is familiar with.
- Gender of the annotator (either male or female).

A total of 22 annotators participated in our survey, 15 males and 7 females. Table 2 provides the distribution of the annotators according to their mother dialect and the dialects they are familiar with.

In the following sections, we provide an in-depth analysis of how each group of annotators of a given dialect performed in the labelling task of dialects and sarcasm.

[2] https://appen.com

Dialect	Mother dialect	Known by
Nile Basin	11	21
Levant	6	10
Gulf	1	7
Maghreb	4	5
MSA	-	16

Table 2: Annotators' details. The table shows the number of annotators who speak a specific dialect as a mother tongue and the number of annotators who mentioned that they know a specific dialect.

4 Results and Analysis

4.1 Dialect Identification

Figure 1 shows the accuracy of annotators in identifying the dialects. From the figure, it is clear the annotators, except Egyptian speakers, were able to identify MSA. Egyptian and Gulf speakers performed best on their dialect. Levantine and Maghrebi speakers performed better on dialects other than their own. Figure 2 shows the distribution of assigned dialect labels compared to the original ones. The results show that Egyptian and MSA are the easiest to identify. However, the annotators seem to confuse other dialects, especially Levantine and Maghrebi. Figure 3 provides a clearer picture of how speakers of one dialect identified other dialects. As shown in Figures 3a and 3c, Egyptian and Gulf speakers excel at identifying texts in their dialect. Figure 3d shows that Maghrebi speakers seem to confuse their dialect with MSA. Levantine speakers (Figure 3b) seem to confuse their dialect with the Gulf dialect. Similar to Figure 2, most annotators tend to easily identify MSA, except for Egyptian speakers who confuse it for Egyptian dialect. Gulf speakers seem to confuse Levantine and Maghrebi for the Gulf dialect.

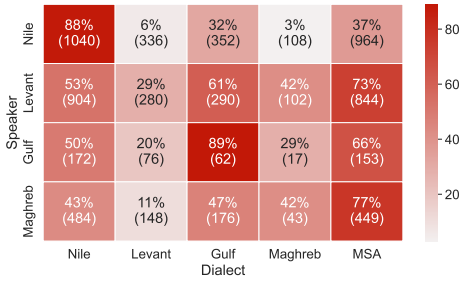


Figure 1: Dialect identification accuracy of annotators speaking different dialects. Annotation counts are indicated in brackets.

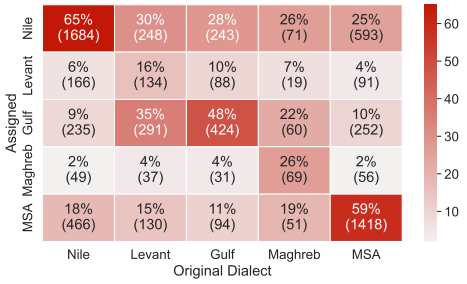


Figure 2: Assigned dialect labels vs the original ones. Annotation counts are indicated in brackets.

4.2 Sarcasm Detection

The effect of identifying dialects might be mild on the task of annotation itself. Thus, we examined the annotators' performance on the subjective task of sarcasm detection, which requires annotators to be able to understand the text to provide correct labels and is found to be a highly challenging task for annotators in different languages (Abu Farha et al., 2022b). Table 3 shows the annotators' performance on sarcasm detection. From the table, Levantine speakers seem to perform better on this task, followed by Gulf speakers. In order to have a better understanding, we analyse the performance over each dialect. Figure 4 shows the performance of speakers of a specific dialect on all the dialects. The figure shows $F1^{sarcastic}$ score and the number of annotations for the respective dialect. The results show that speakers of the Egyptian (Nile) dialect struggle to detect sarcasm written in MSA. Also, speakers of Maghrebi and Egyptian dialects struggle to identify sarcasm expressed using the Gulf's dialect. The results show that Levantine and Gulf speakers perform relatively well on all the dialects. Generally, the annotators achieved the highest score when the text was in Egyptian or their mother dialect.

Speaker's dialect	F1-sarcastic
Nile Basin	0.50
Gulf	0.53
Levant	0.58
Magreb	0.48

Table 3: Sarcasm detection performance (F1-sarcastic) of speakers of different dialects.

4.3 Sarcasm Detection - Dialect Familiarity

Figures 5a and 5b show the performance of annotators in two cases: when the text's dialect is one that they are familiar with and when it is not. When considering the case when the text's dialect is one that the annotators are familiar with (Figure 5a), the annotators have the highest performance on the Egyptian (Nile) dialect. These scores indicate that the annotators are truly familiar with the Egyptian (Nile) dialect. When looking at the cases where people are unfamiliar with the dialect, the performance is inconsistent. For example, the performance of Maghrebi speakers on texts in Levantine is higher for annotators who indicated that they are not familiar with the Levantine dialect. Another example is Levanbine speakers' performance on Maghrebi texts. Such inconsistencies indicate that some annotators might have provided a guess regarding the sarcasm label or that they underestimated their familiarity with the respective dialect.

Figures 6a and 6b show the performance when the annotators identified the dialects either correctly or incorrectly. The figures show that the performance is generally higher when the annotators identify the dialect correctly. This goes along with the previous observation that the annotators performed better on dialects they are familiar with. The exceptions are the performance of Levantine speakers on Maghrebi dialect, Maghrebi speakers on Levantine, and Nile speakers on Gulf dialect. Levantine speakers performed slightly better on MSA when they incorrectly identified the dialect. This goes along with the previous observation that indeed some annotators might be guessing the labels.

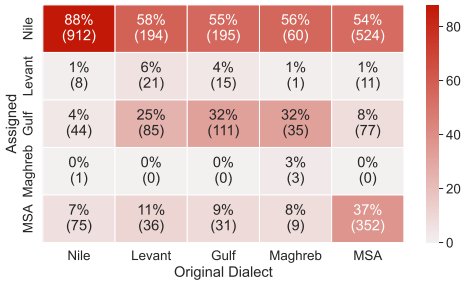
4.4 Sarcasm and Gender

We further analysed the performance of annotators based on their gender. Figure 7 shows the performance over dialects based on the annotators' gender. From the figure, it is noticeable that females perform better than males at detecting sarcasm. Females performed better than males on all dialects except MSA where the performance is quite comparable.

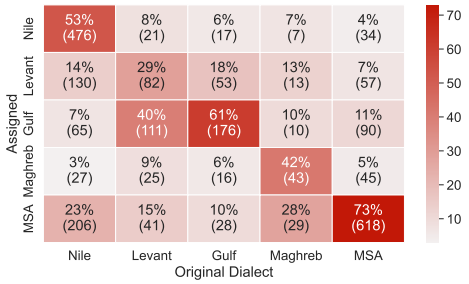
5 Discussion

In this section, we provide a discussion of the results mentioned in Section 4. We also revisit and answer our research questions as follows:

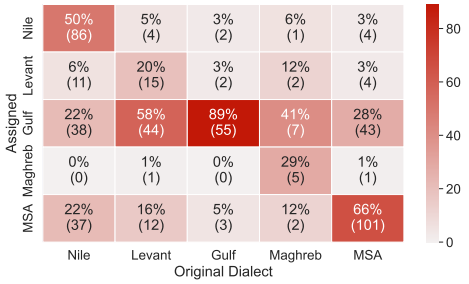
RQ1: How do speakers of different dialects understand other dialects? There are some similarities between dialects and, to some extent, people speaking different dialects can understand each other.



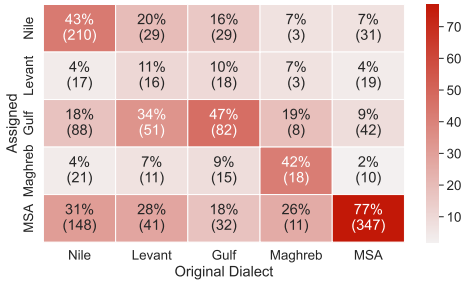
(a) Egyptian (Nile) speakers.



(b) Levantine speakers.



(c) Gulf speakers.



(d) Maghrebi speakers.

Figure 3: Dialect identification performance of speakers of different dialects. The table shows the assigned dialect labels vs the original ones for speakers of each dialect. Annotation counts are indicated in brackets.

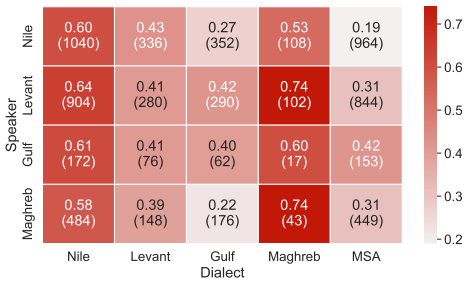
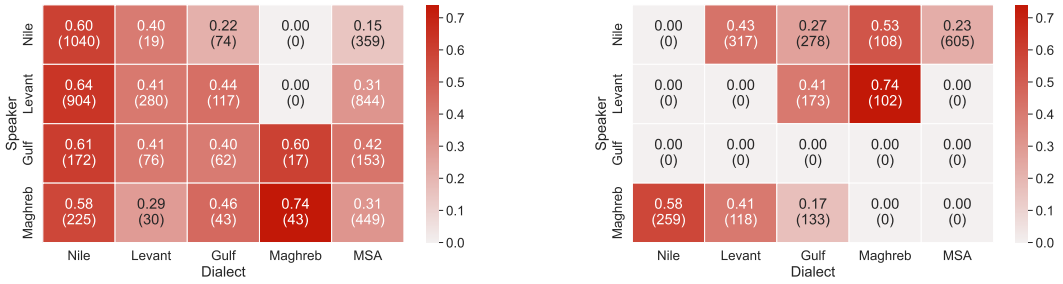


Figure 4: Sarcasm detection performance (F1-sarcastic) of different dialects speaker on each dialect. Original dialect labels were used. Annotation counts are indicated in brackets.

However, as shown in Section 4.1, annotators tend to confuse some dialects for different ones. For example, Egyptian speakers tend to over-identify their own dialect, assuming that more than 50% of other dialects to be Egyptian. This observation is similar to the behaviour observed in (Zaidan and Callison-Burch, 2014). Similar behaviour is observed with Gulf speakers towards Levantine. Such over-identification behaviour, and given the large number of Egyptian annotators, might introduce bias into datasets. Egyptian, Gulf, and Maghrebi speakers tend to perform better on their dialect. Levantine speakers' performance was inconsistent and they seemed to confuse Levantine for Gulf. This could be due to the spectrum of variation within the Levant countries from north to south, where the southern Levantine dialect is closer to the Gulf dialect.

The confusion between the dialects might be due to the fact that these dialects share many words or the differences are mostly phonological. Also, due to the slight differences between dialects' orthography, annotators might confuse sentences in dialects they are unfamiliar with and assign them to a different one. This phenomenon is clear in section 4.3, where Levantine speakers had better performance on MSA for sarcasm detection, but they assigned an incorrect dialect label.

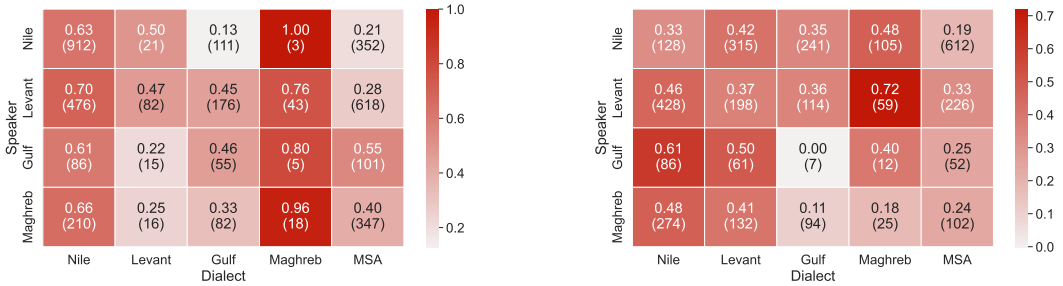
RQ2: How do speakers of different dialects perform on the sarcasm detection task? As discussed in Sections 4.2 and 4.3, annotators tend to better understand sarcasm expressed in their dialect. This is due to the fact that annotators unfamiliar with a dialect would struggle to grasp the complete meaning of a sentence. Also, the fact that sarcasm usually relies on cultural references that can be specific to



(a) Dialect is known.

(b) Dialect is unknown.

Figure 5: Sarcasm detection performance (F1-sarcastic) of speakers of different dialects. Annotation counts are indicated in brackets.



(a) Correctly identified the dialect.

(b) Incorrectly identified the dialect.

Figure 6: Sarcasm detection performance (F1-sarcastic) when based on their prediction of the dialect. Annotation counts are indicated in brackets.

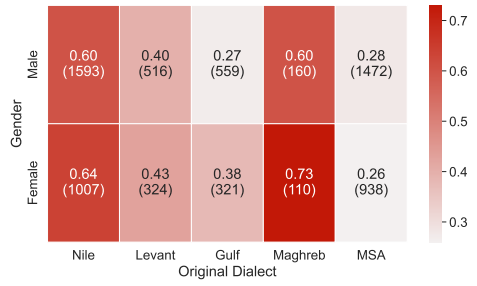


Figure 7: Sarcasm detection performance (F1-sarcastic) based on the annotators' gender. Annotation counts are indicated in brackets.

a region/dialect means that people unfamiliar with the dialect would not be able to understand such references. This observation aligns with the findings in (Oprea and Magdy, 2020b), where the authors found that English language nativeness and country are significantly influential on sarcasm understanding. Indeed, these factors should be considered when collecting third-party annotations for Arabic data. Although there are many shared linguistic and cultural aspects among Arabic speakers, there are still some local differences. Those are embodied in culture, traditions, and dialects. Thus, it is necessary to have native speakers, who are aware and familiar with these differences, annotating subjective and linguistically complex data like sarcasm.

RQ3: Is there a correlation between gender and the performance of an annotator on the sarcasm detection task? Based on the results in Section 4.4, female annotators seem to detect sarcasm better than male annotators. With the small number of annotators and the available data, we cannot provide an explanation for this observation. Future works should consider studying this in a better-designed setup that considers other factors such as educational background and personality traits.

We hope the findings of our study here will be of large benefits for researchers who work in the field of Arabic NLP, especially when applying data annotations. We have shown that dataset creators need to be careful when appointing annotators for

labelling Arabic data. Our findings can act as a guide to appoint annotators with the suitable dialectal background for annotating data in each dialect.

6 Conclusions

In this paper, we analyse how Arabic speakers understand and identify other dialects in written text. We also analyse human performance on sarcasm detection and compare it across different dialects. We use SemEval's 2022 task 6 dataset, which has first-party sarcasm and dialect labels. Our analysis shows that the performance of annotators varies based on the annotators' familiarity with the text's dialect. Also, our analysis shows that annotators might not be familiar with the text's dialect and would confuse it with a different one. Our results also show that females are more likely to understand sarcasm compared to males. Based on the analysis, it is clear that dialect familiarity affects how annotators understand texts and their performance on a specific task. Consequently, we recommend that Arabic dataset creators should consider collecting annotations from native dialect speakers, which would guarantee higher-quality labels.

Limitations

The main limitation of our work is the number of annotators. In our work, we had only one speaker of the Gulf dialect. Future works should consider a larger sample size with a uniform distribution over the dialects. Another limitation is that we used the five major dialects. However, there are dialectal variations within these regions which should be considered. Finally, we only analysed the quality of the labels on sarcasm detection; future works should consider other tasks.

Acknowledgements

This work was partially supported by the Defence and Security Programme at the Alan Turing Institute, funded by the UK Government.

References

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech Language*, 28(1):20–37.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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. ACL.

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. A comparative study of effective approaches for arabic sentiment analysis. *Information Processing Management*, 58(2):102438.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022a. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Ibrahim Abu Farha, Steven R. Wilson, Silviu Vlad Oprea, and Walid Magdy. 2022b. Sarcasm Detection is way too easy! An Empirical Comparison of Human and Machine Sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nora Al-Twairesh, Rawan Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, Nourah Al-Mutlaq, Nada Almanea, Waad Bin Huwaymil, Dalal Alqusair, Reem Alotaibi, Suha Al-Senaydi, and Abeer Alfutamani. 2018. Suar: Towards building a corpus for the saudi dialect. *Procedia Computer Science*, 142:72–82. Arabic Computational Linguistics.

Hassan Alhuzali, Muhammad Abdul-Mageed, and Lyle Ungar. 2018. Enabling deep learning of emotion with first-person seed expressions. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 25–35, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mario J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *CoNLL*, pages 167–177. ACL.

David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. In *ICWSM*, pages 574–577. AAAI Press.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2014. Italian irony detection in twitter: a first approach. In *CLiC-it*, page 28. AILC.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

John D Campbell and Albert N Katz. 2012. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, 49(6):459–480.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for Arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. AIDA: Identifying code switching in informal Arabic text. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 94–101, Doha, Qatar. Association for Computational Linguistics.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *LREC*. ELRA.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53.

Nizar Y Habash. 2010. Introduction to Arabic Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaïli. 2014. Building Resources for Algerian Arabic Dialects. In *15th Annual Conference of the International Communication Association Interspeech*, Singapour, Singapore. ISCA.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. In *COLING*, pages 1837–1848. ACL.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: an annotated corpus for the palestinian arabic dialect. *Language Resources and Evaluation*, 51(3):745–775.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *EMNLP*, pages 1006–1011. ACL.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Hamdy Mubarak and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7, Doha, Qatar. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020.

Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

Silviu Oprea and Walid Magdy. 2019. 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. Association for Computational Linguistics.

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

Silviu Vlad Oprea and Walid Magdy. 2020b. The effect of sociocultural variables on sarcasm communication online. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *COLING*, pages 213–223. ACL.

Veranika Puhacheuskaya and Juhani Järvikivi. 2022. I was being sarcastic!: The effect of foreign accent and political ideology on irony (mis) understanding. *Acta Psychologica*, 222:103479.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *WSDM*, pages 97–106. ACM.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, pages 704–714. ACL.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Fatiha Sadat, Farzindar Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. Association for Computational Linguistics.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic Dialect Identification. *Computational Linguistics*, 40(1):171–202.