Love Me, Love Me Not: Human-Directed Sentiment Analysis in Arabic

Abraham Israeli,[1,2] Aviv Naaman,[1] Yotam Nahum,[1] Razan Assi,[3] Shai Fine,[1] Kfir Bar[1]
[1]The Data Science Institute, Arabic NLP Lab, Reichman University, Israel
[2]Ben-Gurion University of the Negev, Israel
[3]Hebrew University of Jerusalem, Israel
{abraham.israeli,aviv.naaman,yotam.nahum,kfir.bar}@post.idc.ac.il,
razan.assi@mail.huji.ac.il
shai.fine@idc.ac.il

Abstract

Gauging the emotions people have toward a specific topic is a major natural language processing task, supporting various applications. The topic can be either an abstract idea (e.g., religion) or a service/product that someone writes a review about. In this work, we define the topic to be a person who writes a post on a social media platform. More precisely, we introduce a new sentiment analysis task for detecting the sentiment that is expressed by a user toward another user in a discussion thread. Modeling this new task may be beneficial for various applications, including hate-speech detection, and cyber-bullying mitigation. We focus on Arabic, which is one of the most popular spoken languages worldwide, divided into various dialects that are used on social media platforms. We compose a corpus of 3,500 pairs of tweets, with the second tweet being a response to the first one, and manually annotate them for the sentiment that is expressed in the response toward the author of the main tweet. We train several baseline models and discuss their results and limitations. The best classification result that we recorded is 82% F1 score. We release the corpus alongside the best-performing model.

1 Introduction

Sentiment analysis (SA) is one of the most popular tasks in natural language processing (NLP). It is the task of classifying a given piece of text according to the emotions expressed by its author. In its simplest form, the sentiment is classified as positive, negative, or neutral. Aspect-based sentiment analysis (ABSA), a variant of sentiment analysis, is the task of mining opinions from texts, expressed toward specific entities and their aspects (Cambria et al., 2013). For example, in the following review: "`Nice restaurant, a bit expensive but the food is great`", the entity is the restaurant and the aspects are the price and quality of food. While the author writes positively about the quality of food, he/she has some reservations about the price. ABSA is considered an active research area (Pontiki et al., 2016; Ma et al., 2019; Zhang and Qian, 2020; Zhang et al., 2021; Li et al., 2021). However, most of the studies are done with texts written in English.

In the last two decades, social networks have become the dominant written-communication platforms.[1] In most platforms, the users may engage with other posts by up-voting, also known as "like", and by replying with a nested post, thereby generating a discussion thread, open for all users. Most of the existing computational methods for SA do not encode this conversational structure into their prediction models.

With the recent growing interest in training NLP models for languages other than English, the Arabic language has become one of the most prominent among research groups. (Bouamor et al., 2018; Obeid et al., 2020). Nonetheless, the amount of effort invested in advancing sentiment-related technologies in Arabic, is still considered limited comparing to English (Farha and Magdy, 2019; Guellil et al., 2019; Abu Farha et al., 2021; Alhumoud and Al Wazrah, 2021). Therefore, in this work we have opted to work on Arabic, a Semitic language, highly inflected for different lingustic categories. Arabic has what is usually referred to as diglossia, which is a separation between the written and the spoken language. Modern Standard Arabic (MSA) is the language that people use in official settings, while spoken Arabic is considered to be a collection of regional dialects that may significantly differ from each other. In informal writing people often mix MSA with the relevant dialect, forming what is called Middle Arabic. Arabic tweets are typically written in that Middle Arabic, which is in fact described on a spectrum ranging from MSA to the relevant regional dialect. In this work, we

[1]Facebook reported on 2.9 Billion monthly active users (retrieved 09/12/2022), see: https://tinyurl.com/52h8b4mb

put a special focus on tweets written in a mixture of MSA and the Levantine dialects,[2] which are mostly spoken in Lebanon, Syria, Israel, Palestine, and Jordan.

In Section 6, we further elaborate on our future plans to expand this work to other dialects and potentially to other languages.

In this paper, we present a new sentiment analysis task, somewhat related to ABSA, which is about detecting the sentiment expressed by a user toward another user in a discussion thread. We call this task "human-directed sentiment analysis" (`HD-Sentiment`). The emotions that users express toward other users, may play an important role for many NLP applications, such as hate-speech detection (Waseem and Hovy, 2016; Mondal et al., 2017; Ziems et al., 2020), cyber-bullying (Whittaker and Kowalski, 2015; Rosa et al., 2019), and user-based recommendation systems (Han and Karypis, 2005; Da'u and Salim, 2020). To the best of our knowledge, this is the first study to define the `HD-Sentiment` task and to provide a manually annotated corpus that can be used computationally. Similar to other sentiment analysis tasks, we work with three labels: positive, negative, and neutral. To simplify the task, we define it to have an input composed of a pair of posts, the *main post* and the *response*, rather than the entire discussion thread. The goal of the task is to detect the sentiment expressed by the responder in the response post, toward the author of the main post. The model can only use the texts of both posts as input. Adding information to the input will be considered in future works. Figure 1 shows an example of such a pair of posts, written by two different users. In this example, it is clear that the sentiment expressed by the responder toward the author of the main post is positive.

In accordance with other ABSA-related corpora, while the overall sentiment expressed by the responder can be positive, the sentiment toward the main author can be expressed as negative.

`HD-Sentiment` is related to dialogue-level sentiment analysis (Li et al., 2017; Chen et al., 2018; Zhang et al., 2020) since the sentiment is expressed toward participants in a multi-user conversation. `HD-Sentiment` can be of special interest to dialogue-level sentiment researchers as this aspect of the conversation sheds light on the relations between users, which are yet to be addressed. Due to

[2]Both Northern and Southern Levantine dialects.

Figure 1: Example of a tweet and a response. We conceal all identities to preserve users' right to remain anonymous. The example was captured along with an English translation, suggested originally by Google Translate. In this example, we label the human-directed sentiment (`HD-Sentiment`) as positive.

the way the data were collected and annotated (see Section 3), we prefer to define `HD-Sentiment` as a special case of ABSA rather than a sub-topic within dialogue-level sentiment analysis.

At a first glance, the `HD-Sentiment` task seems fairly easy, especially for a response that looks like this: "`@[USER] I admire you`". However, many times responders tend to express their feelings implicitly, using humor, sarcasm, and other figures of speech. The nature of the platform may also affect the way people express themselves in posts (Fiesler et al., 2018). For example, Twitter is a platform for short messages, which forces people to depend on the broader context and compress their messages accordingly.

Table 1 provides some examples of pairs of posts and responses, taken from the corpus we are releasing with this work. The tweets were originally written in Arabic; we added English translations for convenience. For each pair, we provide the label that was assigned by a human annotator, reflecting the sentiment expressed by the responder toward the author of the main post. More details about the corpus are discussed in Section 3.1. Notably, some examples are more explicit than others. They use words that explicitly express emotions, as well as direct references to the author of the main post (e.g., first row). However, in other tweets it is harder to interpret the underlying sentiment. In the third row, it is due to the sarcastic style that is used by the responders. Additionally, like with other ABSA tasks, there are cases where the author does not refer to the aspect at all. The example in the second row is labeled as neutral since there is no evidence for addressing the main author (equivalent to the aspect in ABSA). However, even when explicitly referring to the main authors, responses do not necessarily convey emotions toward them.

Our contribution is threefold: (i) We define a new NLP sentiment analysis task, `HD-Sentiment`; (ii) We release the first annotated corpus designed for the `HD-Sentiment` task, consisting of 3.5K Arabic-written tweets. The dataset is available for download.[3]; and (iii) We report on some baseline results of models that we train for the task. We make the best model available for public use in the HuggingFace public repository.[4]

2 Related Work

Sentiment analysis has been an active research area in the past few decades (Agarwal et al., 2011; Rosenthal et al., 2017a; Sandoval-Almazan and Valle-Cruz, 2018; Lindskog and Serur, 2020). Commonly, an SA task is designed as a binary classification for positive/negative labels. There are a number of popular data sets for the binary classification version, such as IMDb (Maas et al., 2011), consisting of 50K reviews from the Internet Movie Database (IMDb), as well as the Stanford Sentiment Treebank 2 (SST-2) (Socher et al., 2013), which contains about 200K movie reviews. Another known data set is the Yelp Reviews (Asghar, 2016), consisting of more than 500K reviews.

Twitter has always been one of the main sources for acquiring data for SA, exposing some additional information about every tweet and the users beyond the plain text. The SemEval Workshop has a special track for sentiment analysis. Specifically, SemEval-2017 Task 4 (Rosenthal et al., 2017b) consists of five subtasks representing different variants of SA for tweets, written in English and Arabic. Subtask B is about classifying the sentiment expressed in the tweet toward a given topic.

There are a few data sets for the aspect-based SA (ABSA) task. The SemEval-2016 task is the most dominant one (Pontiki et al., 2016). It consists of four subtasks, which vary from the detection of the relevant aspects in the text to the detection of the polarity of a given aspect. The data set contains about 6K reviews.

Considering the information about the author of the input text has been a point of interest, as described several times. Tang et al. (2015) defined a task of SA on reviews in which the user who wrote the text, as well as the product for which the text is written for, are given as input. In another work (Welch and Mihalcea, 2016), a new task has been defined for understanding the sentiment that students hold toward courses and instructors, as expressed by students in their comments. Equivalently, in our work, we are interested in the sentiment that is expressed in a reply tweet, toward the author of the original tweet.

In this work, we focus on Arabic-written tweets. There is a surging amount of computational works on Arabic, especially works related to SA on tweets (Nabil et al., 2015; Abdellaoui and Zrigui, 2018) as well as on other genres (Al-Obaidi and Samawi, 2016). In a recent work (Al-Laith et al., 2021), there has been an attempt to automatically build a large corpus of Arabic texts, annotated for SA. None of these corpora address the task that we define in this work.

3 Data Collection

In this work, we collect data from Twitter. Twitter allows users to reply to posts written by other users. We use the official Twitter API to collect conversation threads of tweets and replies. We define a set of 61 Arabic expressions to limit our collection for tweets that are relevant to the area and dialect of interest. The expressions were carefully composed to cover a variety of topics, such as sports, politics, and economics. Table 2 lists some of them. Additionally, we compile a list of relevant Twitter accounts, known for writing posts with high engagement rates. Most of them are key opinion leaders (e.g., Saad Hariri who was the prime minister of Lebanon). The full list of expressions, as well as the Twitter accounts that we used, is released with the corpus.[5]

The collection was done in June 2021 and applied a full-archive crawling procedure, so the crawling procedure is essentially unlimited by time.

We filtered out conversation threads that *do not* meet at least one of the following three criteria: (i) The tweet language is predominantly Arabic. (ii) The main post contains more than ten characters. (iii) There are at least ten responses to the main post.

[3] https://github.com/idc-dsi/Human-Directed-Sentiment

[4] https://huggingface.co/DSI/human-directed-sentiment

[5] https://github.com/idc-dsi/Human-Directed-Sentiment

	Main Post	Response Post	L
1	اذا وصلت لمرحلة إنك ترى وتعرف كل شيء ولكنك تظهر لهم إنك غبي ولم تفهم شيء فأنت قد فهمت الحياة تماماً.. 😉 #صباحو_للعالم_بتدّعي_الذكا 😴 If you reach the stage in which you see and know everything but act as if you are ignorant and don't understand anything then you have fully understood life..😉 #Good Morning to the people who pretend to be smart 😴	دخل ذكاكي انت 😍😍 How clever you are 😍😍	POS
2	هذه الليلة توفي دونالد رامسفيلد، أحد معدي ومخططي اجتياح افغانستان والعراق .هو أحد أهم الرجال الدمويين في إدارة جورج بوش الإبن. Tonight, Donald Rumsfeld, one of the organizers and planners of the invasions of Afghanistan and Iraq, died. He is one of the most important and bloody men in the administration of George Bush Jr.	اليوم يسلم كتابه بشماله.. عند رب يقول انا منا نستنسخ ما كنتم تعملون ..ويقول في كتاب لا يغادر صغيرة ولا كبيرة إلا احصاها ...اليوم يرى عين الحقيقة المطلقة للآخرة Today he returns his soul… Facing the Lord he says, "I will not reproduce what you did." He will tell it all, big and small. Today he faces the eternal truth	NEU
3	الرئيس عون: ما حصل في الأيام الماضية أمام محطات المحروقات غير مقبول، وإذلال المواطنين مرفوض تحت أي اعتبار، وعلى جميع المعنيين العمل على منع تكرار هذه الممارسات سيّما وانّ جدولاً جديداً لأسعار المحروقات صدر، ومن شأنه أن يخفّف الأزمة President Aoun: What happened in the past few days in front of the gas stations is unacceptable, and the humiliation of citizens is rejected under any consideration, and all concerned should work to prevent the recurrence of these practices, especially since a new tariff of fuel prices has been issued, which would alleviate the crisis.	صرلو فترة هيك لازم تضرب ايدك عالطاولة وتقله لرئيس الجمهورية يحسن الوضع شوي It's been like that for some time, you ought to hit your fist on the table and tell the President of the Republic to make things a little better.	NEG

Table 1: Examples of pairs of a post and response. The examples are taken from our annotated corpus. POS, NEU, and NEG are the positive, neutral, and negative labels respectively. We added English translations, which were manually prepared by a native speaker.

Overall, we collected 20.1K threads, corresponding to a total number of 346.3K tweets.

As mentioned above, instead of working with full conversation threads, we define our task to focus only on pairs of tweets, the main post, and one of its responses. Therefore, we compile our corpus accordingly.

Expression	Translation	Domain
الامير حمزة	Prince Hamzah	Politics
فلسطين	Palestine	Politics
ارتفاع الأسعار	High Prices	Economics
اصوات من السماء	Voices from Heaven	Religious
بشار مراد	Bashar Murad[6]	Culture
جميلة عوض	Jamila Awad[7]	Culture

[6]A Palestinian singer, songwriter, and social activist.
[7]An Egyptian actress.

Table 2: Crawling expressions. A *sample* of the Arabic terms we use for crawling, provided with their English translation, and the domain they are most relevant to.

	Main Posts			Response Posts		
	Avg.	Med.	Std.	Avg.	Med.	Std.
Chars	175.12	179	83.41	109.16	85	73.11
Tokens	64.85	65	30.34	43.25	35	27.09
Hashtags	0.53	0	1.09	0.11	0	0.56
Emojis	0.01	0	0.12	0.45	0	0.68

Table 3: Corpus statistics. The numbers are calculated over the entire collection of 3,500 tweets. Avg., Med., and Std. are the average, median, and standard deviation respectively.

3.1 Human Annotation

We sampled 3,500 pairs uniformly from the main collection of conversational threads, and assigned them for human annotation. Specifically, we pair every main post with up to five responses, chosen randomly. We provide some additional information about the chosen tweets in Table 3. We learn from the table that main posts are significantly longer than responses. Additionally, the authors of the main posts tend to use hashtags more frequently than responders, while the latter use emojis in their tweets more than main authors do.

We hired three human annotators to label the 3,500 tweet pairs. All annotators are highly educated Arabic speakers, fluent in MSA and the relevant regional dialects. They were introduced to the definition of the task, and were given careful annotation guidelines alongside specific annotation examples. As a first phase, we started annotating a small set of 100 pairs for training the annotators and calibrating the guidelines. The guidelines were adjusted to handle cases of annotator disagreements. In the second phase, we asked two annotators to label the entire set of 3,500 pairs. The agreement of the two annotators was measured to be 74%, corresponding to a kappa (Cohen, 1960) value of 0.59. The third annotator was assigned with the adjudication task, where he was asked to label only pairs on which the two annotators disagreed (26% of the pairs), to have a final decision for each pair.

In 95.3% of the cases, the third annotator agreed with one of the annotators. For our final corpus we removed the pairs that had complete disagreement among all three annotators (43 cases). The distribution of the [positive, neutral, negative] labels in the corpus are [9.59%, 44.45%, 45.95%]. We believe that the relatively small number of positive pairs stems from the nature of the platform as well as the topics and geography that we decided to focus on.

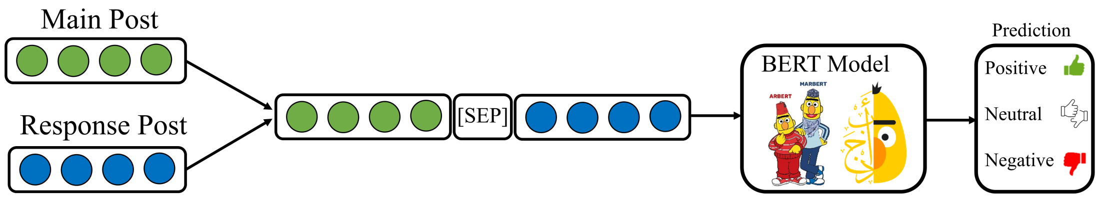

Figure 2: Our architecture for the fine-tuned BERT-based models. We concatenate the main post and the response and add the special token [SEP] in between.

4 Computational Approach

To validate the new annotated dataset and its usability, we trained three classifiers and compared their performance with two baseline approaches.

4.1 Experimental Setting

We preprocessed every tweet by replacing user mentions (formatted in Twitter as @<user>) with a placeholder word `[USER]`, and urls with `[URL]`. Hashtags remain untouched, as they may carry important information for SA. For evaluation, we used a 5-fold cross-validation approach. To get the most out of the new annotated resource, and due to the low support for the positive label, we do not split the corpus for train and test sets. We use the standard classification evaluation metrics. For each label, we calculate the precision, recall, and F1-score, as well as the macro and weighted-average scores over the three labels.

We fine-tuned different Arabic BERT (Devlin et al., 2019) models on the new `HD-Sentiment` corpus, during 5 epochs. To handle the skewed distribution of the labels, we used a weighted cross-entropy loss, with weights assigned according to the inverse proportion of their distribution.

4.2 BERT Based Classifiers

We preprocessed every input pair of tweets by concatenating the main post and the response with a special [SEP] token placed in between. The full architecture of our model is depicted in Figure 2. We used three different pre-trained Arabic language models,[8] using the transformers (Wolf et al., 2020) library by Hugging Face[9]: `AraBERT` (Antoun et al., 2020), `GigaBERT` (Lan et al., 2020), and `MARBERT` (Abdul-Mageed et al., 2020) that relies solely on Twitter data, which makes it a better fit for NLP tasks involving dialectical Arabic texts from social media, such as ours.

[8] Using the `BertForSequenceClassification` class.

[9] https://huggingface.co

4.3 Baseline models

We compared our classifiers with two baselines:

CAMeLBERT Sentiment Analysis. CAMeLBERT (Inoue et al., 2021) is a pre-trained language model, which has already been fine-tuned for several downstream Arabic NLP tasks, including sentiment analysis.[10] By the time of writing this paper, it is considered to deliver state-of-art results for SA in Arabic. The model was trained to classify texts with three labels: positive, negative, and neutral. We run the model on the response tweet to gauge its overall sentiment, which we return as a final predicted label.

Lexicon-Based Model. First, we look for mentions of the main author in the response, including references through 2nd-person pronouns. If none are found, the model returns “neutral”. However, if found, we use existing lexicons (Saif M. Mohammad and Kiritchenko, 2016) for detecting all instances of emotional words and related hashtags. Every word is assigned with a sentiment score,[11] which we average into an overall sentiment score assigned for the response. We predict “positive” (or “negative”) based on the sign of the overall score.

5 Results and Analysis

The results obtained by each model averaged over the five cross-validation folds, are summarized in Table 4. The best results in each column are in boldface. We add $*$ next to a number to indicate statistically significant results (p-value $< 10^{-4}$), using the Mann Whitney U-test (Mann and Whitney, 1947). The first two rows are the results of the baseline models (see Section 4.3). While the baseline models show competitive results in some of the individual labels, their overall results (measured as macro-F1 (M-F1) and weighted-F1 (W-F1)) are much worse than the results obtained by the fine-tuned models.

[10] CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment

[11] The score is not limited to a specific value range, which can also be negative

	Positive			Neutral			Negative			All	
	P	R	F1	P	R	F1	P	R	F1	M-F1	W-F1
Lexicon	0.11	0.52	0.19	0.74	0.67	0.7	0.6	0.21	0.31	0.4±0.01	0.48±0.01
CAMeLB	0.39	**0.68**	0.49	0.77	0.11	0.19	0.55	**0.91***	0.69	0.46±0.02	0.45±0.01
AraBERT	0.62	0.14	0.22	0.75	0.71	0.72	0.69	0.83	0.75	0.57±0.05	0.69±0.02
GigaBERT	**0.8**	0.3	0.43	0.78	0.77	0.78	0.74	0.84	0.78	0.66±0.04	0.75±0.02
MARBERT	0.79	0.67	**0.72***	**0.84***	**0.81***	**0.82***	**0.82***	0.87	**0.84***	**0.79 ± 0.02***	**0.82 ± 0.02***

Table 4: Results. P and R are precision and recall. M-F1 and W-F1 are the macro-F1 and weighted-F1 over the three labels. Lexicon and CAMelB are the lexicon-based and CAMeLBERT Sentiment Analysis models, respectively. Results are averaged over the five cross-validation folds. The standard deviation of the overall results is provided in the last two columns. The best results are in boldface while the second-best results are underlined. Statistically significant best results are marked with a *.

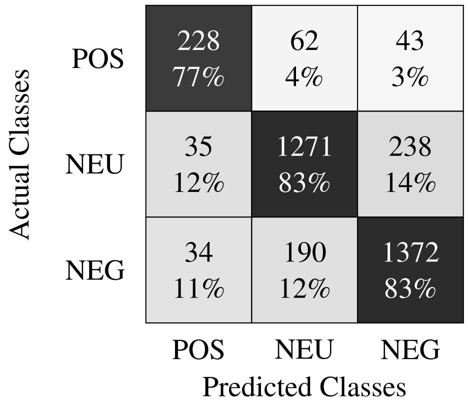

Figure 3: Confusion matrix for the best performing model (MARBERT). POS, NEU, and NEG are the positive, neutral, and negative labels, respectively. The percentage number in each cell is calculated columnwise.

Among the fine-tuned models, both AraBERT and GigaBERT perform well on the neutral and negative labels. However, their performance on the positive label, the one with the low support, is not as good. On the other hand, MARBERT outperforms all other models, on all labels' F1 scores as well as on the aggregated overall scores. This is unsurprising, considering that MARBERT was trained solely on Twitter data, and its size is larger than the other models' datasets.

We now take a closer look into the performance of the MARBERT model. Figure 3 is the confusion matrix we got by running MARBERT on the five cross-validation folds. It looks like the model has hard time distinguishing between the neutral and negative labels. On the other hand, the negative and positive labels are rarely "mixed up" by the model. As observed in both Table 4 and Figure 3, positive is the most difficult label to predict.

Quantitative analysis. Overall there are 602 misclassified pairs, out of which 317 (52.7%) were assigned with two different labels by the original human annotators. Disagreement at a rate of 52.7% is significantly higher than the disagreement rate of the entire corpus (26%, see Section 3.1), suggesting that the misclassified pairs are likely to be more difficult than the others even for human annotators.

6 Conclusion and Future Work

In this work we defined a new task, called Human-Directed Sentiment Analysis (`HD-Sentiment`). We collected and annotated the first `HD-Sentiment` corpus, and made it publicly available. Additionally, we fine-tuned a number of baseline models, discussed their results, and published the one that performed best.

`HD-Sentiment` may be considered as a special case of ABSA using only one aspect defined as the author of the main post. To some extent, `HD-Sentiment` extends previous works in the field of hate-speech detection and cyber-bullying; however, `HD-Sentiment` is more general as it aims at capturing a full range of emotions expressed in conversations, which are neither considered as bullying nor as expressing hate towards someone.

Part of the challenge in `HD-Sentiment` is the fact that the users who are involved in the conversations are not necessarily known in advance and are not provided as input to the learning model. We do not store historical information about the users nor their previous interactions. In our corpus, we included interactions between users, who may or may not know each other in advance.

Finally, we decided to work with Arabic, one of the most popular spoken languages worldwide.Consequently, there is a growing interest in processing Arabic for various NLP tasks. However, we believe that the `HD-Sentiment` task can be applied in other languages and other social platforms.

Future work takes two trajectories: (i) Extending `HD-Sentiment` to other languages, including the collection and annotation of additional corpora, and (ii) Building an explainability component for `HD-Sentiment` classifiers to better interpret the model's output.

References

Houssem Abdellaoui and Mounir Zrigui. 2018. Using tweets and emojis to build tead: an arabic dataset for sentiment analysis. *Computación y Sistemas*, 22(3):777–786.

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

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*.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38.

Ali Al-Laith, Muhammad Shahbaz, Hind F Alaskar, and Asim Rehmat. 2021. Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus. *Applied Sciences*, 11(5):2434.

Ahmed Y Al-Obaidi and Venus W Samawi. 2016. Opinion mining: analysis of comments written in arabic colloquial. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1.

Sarah Omar Alhumoud and Asma Ali Al Wazrah. 2021. Arabic sentiment analysis using recurrent neural networks: a review. *Artificial Intelligence Review*, pages 1–42.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, 28(2):15–21.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Aminu Da'u and Naomie Salim. 2020. Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review*, 53(4):2709–2748.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

Imane Guellil, Faical Azouaou, and Marcelo Mendoza. 2019. Arabic sentiment analysis: studies, resources, and tools. *Social Network Analysis and Mining*, 9(1):1–17.

Eui-Hong Han and George Karypis. 2005. Feature-based recommendation system. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 446–452.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. Gigabert: Zero-shot transfer learning from english to arabic. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Sebastian Lindskog and Juan A Serur. 2020. Reddit sentiment analysis. *Available at SSRN 3887779*.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3538–3547.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th acm conference on hypertext and social media*, pages 85–94.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2515–2519.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017a. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017b. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Mohammad Salameh Saif M. Mohammad and Svetlana Kiritchenko. 2016. Sentiment lexicons for arabic social media. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Rodrigo Sandoval-Almazan and David Valle-Cruz. 2018. Facebook impact and sentiment analysis on political campaigns. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, pages 1–7.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning semantic representations of users and products for document level sentiment classification. 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 1014–1023, Beijing, China. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Charles Welch and Rada Mihalcea. 2016. Targeted sentiment to understand student comments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2471–2481, Osaka, Japan. The COLING 2016 Organizing Committee.

Elizabeth Whittaker and Robin M Kowalski. 2015. Cyberbullying via social media. *Journal of school violence*, 14(1):11–29.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3540–3549.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 504–510.

Yazhou Zhang, Zhipeng Zhao, Panpan Wang, Xiang Li, Lu Rong, and Dawei Song. 2020. Scenariosa: a dyadic conversational database for interactive sentiment analysis. *IEEE Access*, 8:90652–90664.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counter-hate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*.