Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Towards Generalization of Machine Learning Models: A Case Study of Arabic Sentiment Analysis

Samir Abdaljalil[1], **Shaimaa Hassanein**[2*], **Hamdy Mubarak**[1], **Ahmed Abdelali**[1]

[1] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[2] Zewail City of Science and Technology, Giza, Egypt
{sabdaljalil, hmubarak, aabdelali}@hbku.edu.qa, s-shaimaa_said@zewailcity.edu.eg

Abstract

The abundance of social media data in the Arab world, specifically on Twitter, enabled companies and entities to exploit such rich and beneficial data that could be mined and used to extract important information, including sentiments and opinions of people towards a topic or a merchandise. However, with this plenitude comes the issue of producing models that are able to deliver consistent outcomes when tested within various contexts. Although model generalization has been thoroughly investigated in many fields, it has not been heavily investigated in the Arabic context. To address this gap, we investigate the generalization of models and data in Arabic with application to sentiment analysis, by performing a battery of experiments and building different models that are tested on five independent test sets to understand their performance when presented with unseen data. In doing so, we detail different techniques that improve the generalization of machine learning models in Arabic sentiment analysis, and share a large versatile dataset consisting of approximately 1.64M Arabic tweets and their corresponding sentiment to be used for future research. Our experiments concluded that the most consistent model is trained using a dataset labelled by a cascaded approach of two models, one that labels neutral tweets and another that identifies positive/negative tweets based on the Arabic emoji lexicon after class balancing. Both the BERT and the SVM models trained using the refined data achieve an average F-1 score of 0.62 and 0.60, and standard deviation of 0.06 and 0.04 respectively, when evaluated on five diverse test sets, outperforming other models by at least 17% relative gain in F-1. Based on our experiments, we share recommendations to improve model generalization for classification tasks.

Introduction

Social media aggressively occupies a significant part of our daily life (Alshehri et al. 2018). This phenomenon affects all worldwide societies equally, including the Arab region. In fact, according to Radcliffe and Abuhmaid (2020), "More than 10 million users are active on Twitter in Saudi Arabia, akin to 38% of the population." Users tend to share different types of information through social media platforms like Twitter. This includes sharing opinions and thoughts about topics or companies (Sharma et al. 2022). As a result of such activities, platforms like Twitter holds a wealth of data that could be mined for opinion tracking and understanding people's sentiment towards topics, concerns or entities (Kharde and Sonawane 2016).

Researchers in both the public and private sectors use this data to build tailored systems using the information gathered from these platforms. Later, such systems are in turn used to analyze the platforms' content and act as a proxy in making recommendations and decision. Performing such analysis "...Identifies the polarity, relevance and objectivity of the text. With the help of these tools, text can be classified into categories like positive, negative and neural" (Sharma and Ghose 2020). While this process exemplifies the automation that is being sought after using machine learning (ML) and artificial intelligence based algorithms, such process comes with a plethora of challenges. Many of the challenges are related to the data itself, as well as the the annotation quality. Others are related to the technology used to build such systems.

Traditionally, machine learning classifiers such as Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression, ensemble of voting classifiers, are amongst the technologies used in investigating the data and developing the system. Conversely, due to the abundance of available data and computational power in recent years, deep learning methods such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Transformers are becoming the de facto players in the field. This is mainly credited to outperforming traditional approaches in many tasks and applications. To build these systems, humans are typically involved at some stage to annotate a sample or a subset of the data to be used in training these systems. With this type of processing additional challenges arises, some are inherent from the language itself. In the case of **Arabic**, the language includes complex and rich morphology (derivational and inflectional) (Mubarak et al. 2019), rare use of diacritics (similar to short vowels in English) which increases word ambiguity, as well as the presence of many dialects that are used across different Arab countries and cities, and not all of them are mutually intelligible (Shaalan et al. 2018). Others are related to the approaches used for the classification (Namugera, Wesonga, and Jehopio 2019).

In machine learning, one of the challenges faced by the

*Shaimaa Hassanein contributed to this work during her internship at Qatar Computing Research Institute.

community is the lack of generalization of models, which has been discussed in the literature (Barbiero, Squillero, and Tonda 2020). In fact, "there is little consensus on what interpretability in machine learning is let alone how to evaluate it for benchmarking or reason about how it may generalize to other contexts" (Doshi-Velez and Kim 2018). However, as far as we know, not much has been done on investigating the generalization of machine learning models for Arabic data, specifically in the task of sentiment analysis.

In this paper, we investigate the generalization of models and data for Arabic Sentiment Analysis. It is almost unfeasible to have a dataset that includes every possible dialect, topic and/or country represented on Twitter, so this research attempts to gain insight into the role that the versatility of the training data plays on the performance of its corresponding model when tested on unseen data. We use several diverse test sets to evaluate our experiments. Furthermore, we apply the techniques on a large corpus consisting of approximately 2M random tweets to ensure diversity in dialects and topics introduced to the model.

The main contributions of this research are: (i) We show that the performance of different ML models trained using current datasets in the Arabic sentiment analysis task drop significantly when they are tested on different test sets; (ii) We create and share a large balanced corpus of approximately 1.6M Arabic tweets of different topics and dialects and their corresponding sentiment, that were automatically annotated using multiple approaches discussed in detail and compared throughout the paper through Zenodo (Abdaljalil et al. 2023); and finally, (iii) We perform experiments to understand which techniques to use to improve model generalization, and evaluate those approaches on multiple diverse test sets, while providing a list of recommendations gained from our experiments that would enhance model generalization for other classification tasks and potentially other languages.

Related Work

Arabic Sentiment Analysis

Approaches used to tackle the sentiment analysis task varied between machine learning classification models, deep learning models, lexicon based models, and semantic understanding approaches.

Al-Ayyoub, Bani Essa, and Alsmadi (2014) collected Arabic user comments by crawling the web. They used Khoja's stemmer (Khoja and Garside 1999) to get the word roots, and normalized the text to its Modern Standard Arabic (MSA) equivalent. They built their own lexicon of 120K Arabic terms by using publicly available content[1] and extended the lexicon with more distinct stems used in Arabic news articles, translated them into English, then used online English Twitter sentiment analysis tool, Sentiment140[2], to determine the sentiment value of each word. They achieved 87% accuracy on 900 MSA tweets. With the preprocessing techniques used by Al-Ayyoub, Bani Essa, and Alsmadi (2014), dialectal language, the most common form of communication used on social media, was not explored in this case.

AraSenti is an Arabic sentiment keyword and emoticon lexicon specific to Twitter created by Al-Twairesh, Al-Khalifa, and Al-Salman (2016). Using that as a basis for their research, Al-Thubaity, Alqahtani, and Aljandal (2018) created their own lexicon, the Saudi dialect sentiment lexicon, SauDiSenti, using a labeled dataset of 5400 tweets comprising Saudi dialect and MSA. They compared the results achieved using their lexicon with the largest known AraSenTi dictionary of 225,329 words. They achieved a 0.478 F-1 score in comparison with the 0.39 F-1 score achieved using the large AraSenTi dictionary.

Semantic similarity, a similar approach to the keyword lexicon based approach was also explored in the literature. Alowaidi, Saleh, and Abulnaja (2017) used a two-part feature extraction methodology. They translated the emojis with their descriptive words, and applied a Bag of Words model. However, due to the "main limitation of this model [being] that it is semantically weak; where words considered as independent features and ignore the semantic associations between them..." Alowaidi, Saleh, and Abulnaja (2017) also used the Arabic WordNet (AWN), which is the equivalent of WordNet (Miller 1995). This essentially maps each word within a tweet to its corresponding related concepts, and the concepts related to each word within the tweet are incorporated as features. Experimenting with SVM and NB classifiers, they found that the SVM approach achieved an F-measure value of 95.63%.

On the other hand, researchers used purely supervised approaches by applying deep learning. Heikal, Torki, and El-Makky (2018) proposed a deep learning approach to sentiment analysis of Arabic text, which combines a Convolutional Neural Network (CNN) model, and Long Short-term Memory (LSTM) model. As part of their data preparation step, they used AraVec, which is trained on Arabic tweets, to produce a vector embedding for each word within the tweet. Nabil, Aly, and Atiya (2015) introduced ASTD, a collection of 10k top trending tweets in 2013, they built a CNN model to identify sentiment on this data set, which had an accuracy of 64.30%. They also experimented with a Bidirectional LSTM (Bi-LSTM) architecture, which achieved an accuracy score of 64.75%. To improve their results, both architectures were combined to create an ensemble model that uses soft voting. This approach resulted in an accuracy of 65.05%.

Mhamed et al. (2021) used a CNN approach to the sentiment analysis task. They explored two CNN models, in which they use both the ASTD dataset (Nabil, Aly, and Atiya 2015), which has four classes, as well as ATDFS, a binary sentiment dataset (Alharbi and Aljaedi 2019). Their first model made use of global average pooling function with two layers; [Their second model] is a CNN using bidirectional gated recurrent units (GRUs)" (Mhamed et al. 2021). They achieved an accuracy of 73.17% and 85.58% on the ASTD multi-classification dataset, and ATDFS binary sentiment dataset, respectively.

[1] http://diab.edublogs.org/dataset-for-arabic-document-classification/

[2] http://www.sentiment140.com/

Many researchers applied multiple supervised and unsupervised techniques when investigating the sentiment analysis of Arabic text. Using a dataset of 2,000 tweets written in the Syrian dialect, Aloqaily et al. (2020) initially applied a lexicon-based approach by using SentiWordNet (Esuli and Sebastiani 2006), a publicly accessible lexical tool, to appropriately label the sentiment of each individual word within a tweet, to ultimately calculate an overall sentiment score for each tweet. This approach achieved an F-1 score of 0.22. They then used the labelled dataset to train five different classifiers, including Logistic Model Trees, and tested it on a dataset of 1,600 similar tweets, achieving an F-1 measure score of 0.92.

Similarly, Abdulla et al. (2013) used a lexicon-based approach to predict the sentiment of 2000 Arabic tweets, using keywords found on the SentiStrength website[3]. When comparing their predicted labels with the actual labels, they achieved 59% accuracy. They then built classifiers using the labelled dataset, and achieved an accuracy of 85% using SVM and NB models. Furthermore, Abuelenin, Elmougy, and Naguib (2017) deployed a combination of stemmers as well as feature extraction techniques in the preprocessing stage of their research. They used cosine similarity to compare words within their training data to a 400-word Arabic Slang Lexicon, and annotated 2,000 tweets accordingly, and to be used for trained different types of classifiers. They found that the best combination included using countVectorizer and an ISRI stemmer (Syarief et al. 2019) to preprocess the data, while setting the cosine similarity threshold at 0.7, and training a LinearSVC classifier, which had an overall accuracy of 92.98%.

Although the literature has explored Arabic sentiment analysis, generalization of the resulting trained models within different contexts is yet to be heavily explored.

Model Generalization in Sentiment Analysis

Model generalization is an active topic within the research community. According to Barbiero, Squillero, and Tonda (2020), "In many real-world cases, it is of utmost importance to estimate the capabilities of a machine learning algorithm ... to provide accurate predictions on unseen data, depending on the characteristics of the target problem." In terms of the sentiment analysis task specifically, several researchers explored model generalization, and proposed different ways to address it.

Using movie review datasets, Ashir (2021) investigated the use of lexicon combined with unsupervised machine learning to increase model generalization. The author also incorporated rule-based techniques, and found that lexical-based rules such as grammatical rules, emoticons, and Part of Speech (POS) tagger, reinforce the model generalization greatly instead of just considering the word embeddings as a feature for training a machine learning model. Wang et al. (2017) proposed Select-Additive Learning (SAL) approach to address the challenges that come with model generalizability in sentiment analysis for videos. Their work takes into account "Sentiment-associated features (i.e. people are smiling while expressing positive sentiment) more than the identity-related features (i.e. wearing glasses)" (Wang et al. 2017). Through this approach they found an increase in the generalization of the model when evaluated on unseen data.

While in the current literature more relevant work related to model generalization for sentiment analysis exists, there is a clear gap when addressing Arabic content. As a result, we aim to focus on Arabic textual data to explore generalization of the sentiment analysis task.

Data

Training Datasets

In this paper, we are using the largely available Arabic twitter sentiment analysis datasets, namely:

ASAD (Alharbi et al. 2021) is comprised of a total of 95K Arabic tweets, and their corresponding sentiment classification. The dataset contains a total of three classes including: Positive, Negative and Neutral. Using the Twitter API, they randomly collected Arabic tweets posted between May 2012 and April 2020. The tweets were then annotated by a group of Arabic native annotators, where each tweet had an average of three independent annotations to ensure reliability. We were able to access to approximately 55K of the tweets since the authors shared the dataset by tweet IDs and by 2022-07, many tweets were not available on Twitter anymore. We used 50K for training, and 5K as the testing data.

ArSAS (Elmadany, Mubarak, and Magdy 2018) consists of 21K Arabic tweets, that are labelled under four classes of sentiment: Positive, Negative, Neutral, and Mixed. The authors specified a set of 20 topics (mostly political), and used Twitter API to collect tweets posted in November 2017 containing any of the topics. Using crowdsourcing, 500 Arab annotators participated in the task and each tweet was annotated by 3 annotators. Quality of the annotations was controlled by utilizing 70 test questions embedded randomly within the tweets presented to each annotator (success threshold was 70%). The Inter-Annotator Agreement (IAA) was 0.79 which indicates annotations of a high quality. We had access to approximately 16K of the tweets used for training, while 5K are used for testing.

Different Testing Datasets

In addition to the ASAD and ArSAS datasets, we included other datasets with a variety of topics to accurately test the generalizability of our methods and proposed datasets.

ATSAD: The Arabic Tweets Sentiment Analysis Dataset (ATSAD) by Abu Kwaik et al. (2020) consists of 36K tweets extracted in April 2019 using Twitter API. They used a two-point sentiment scale that consists of either Positive or Negative. The main approach they employed is using "Distant supervision using emojis as weak labels to annotate the entire dataset" (Abu Kwaik et al. 2020). From the full dataset, the authors asked two annotators, an NLP expert and an Arabic Native speaker, to manually annotate 8K tweets that they consider as the gold standard dataset, where both annotators agreed on 90% of the labels. In case of any disagreement, the NLP expert's label was considered as the gold label. As a result, we use the 8K gold tweets and their corresponding

[3]http://sentistrength.wlv.ac.uk/

sentiment label to test our different approaches throughout this paper.

Saudi_data: Alyami and Olatunji (2020) curated a dataset by extracting tweets based in Saudi Arabia discussing several social issues, such as women's rights and unemployment. We will refer to this dataset as 'Saudi_data' throughout this paper. They opted for a manual annotation of the tweets based on the sentiment of the text, in which they use a two-point classification scale of positive or negative. We consider the dataset of 4.25K tweets as a test set in our case.

Egypt_data: Kora and Mohammed (2019) built a corpus of Arabic Egyptian tweets and their corresponding sentiment. We will be referring to this dataset as the 'Egypt_data' throughout this paper. This dataset consists of 40K tweets, where 20K are labelled as negative tweets, while the other 20K are labelled as positive. The authors manually labelled the dataset, and "the collected tweets covered a blend of different general topics discussed on Twitter" (Kora and Mohammed 2019). We consider the full dataset (40K tweets) as a test set throughout this paper.

Experiments & Results

We consider the ArSAS and ASAD datasets as the starting points for exploring the generalization. This consideration was motivated by the fact that these datasets had the highest average human evaluation scores (will be further detailed in the subsection *Human Evaluation*), and have consistent labels used within each dataset (Positive, Negative, and Neutral/Mixed). Nonetheless, the other datasets are used for testing the models we built.

Differences between ASAD and ArSAS

Upon further analysis of both the ASAD and ArSAS datasets separately, there seems to be a large discrepancy in the topics discussed within each dataset. As shown in Figure 1, the most frequent topics discussed in the ASAD set (as shown on the left of Figure 1) are topics related to social/local topics, such as Corona Virus and Saudi Arabian cities, while the ArSAS dataset (as shown on the right) is more related to sports/politics such as the World Cup, Mohamed Salah, and Arabic Spring revolution. Such differences can be related to firstly: The different time period when the different data was collected, and secondly, the approach used for crawling the data in which the keywords were different.

Furthermore, there is a difference in the class distribution in each dataset. In ASAD, a large portion of the data is neutral (71%), while the rest is either positive or negative. While ArSAS consists of 41% neutral, 37% negative, and 22% positive tweets, as shown in figure 2.

Human Evaluation

To evaluate the quality of the labels of the tweets in each of the datasets, we hired two independent annotators who are native Arabic speakers from different countries and familiar with Arabic dialects. They were given the same set of 200 random tweets from each dataset and asked to label each of the random samples using a three-point classification scale: Positive, Negative, Neutral/Mixed. Once the annotations were complete, we found that both annotators had an agreement of approximately 91% in their labels. Each set of annotations provided by the annotators was compared to the reference labels provided within the datasets, and we calculated a percentage of overlap between the two. The results of this evaluation are shown in Table 1.

Dataset	Classes	Av. Overlap	Agreement (%)
ASAD	P, N, T	0.75	89
ArSAS	P, N, T, M	0.68	94
ATSAD	P, N	0.31	87
Egypt_data	P, N	0.61	91
Saudi_data	P, N	0.66	92

Table 1: Human Evaluation of Different Datasets. Classes are P: Positive, N: Negative, T: Neutral, M: Mixed

One interesting remark to note is that the human evaluation score of ATSAD is much lower than the rest of the datasets, being at 0.31. This is due to the fact that the ATSAD dataset uses a two-point classification scale of only positive and negative labels, and does not take into consideration any neutral or mixed classes. However, both annotators seemed to think that there were many neutral examples throughout ATSAD, which resulted in the low evaluation score when compared to the reference labels.

Experimenting with BERT using ArSAS

We experimented with fine tuning different BERT models using ArSAS. Although we used both ArSAS and ASAD as our starting point for this research, we started with ArSAS since it was a smaller dataset, and upon further analysis, we found that the class distribution in ArSAS is more balanced than the distribution found in ASAD, as shown in figure 2 and discussed previously. These BERT models are Marbert[4], Arabert[5], Arabert-Twitter[6], Qarib[7] and Camelbert[8]. When evaluated on the ArSAS test set, Arabert-Twitter did the best, as it evaluated at an F-1 score of 0.77, so we decided to continue using Arabert-Twitter when comparing other BERT experiments with our SVM experiments throughout this paper.

Merging ASAD and ArSAS Data

To improve on the generalizability of the models, we experimented with merging the training datasets of ASAD and ArSAS since they include different topics and dialects, as discussed previously. Initially, we merged both training datasets, amounting to 66K tweets and their corresponding sentiment, to train the **merged_ArSAS_ASAD** model, and test it on the different test sets. As shown in Table 2, this model achieved an F-1 score that was essentially an average of the scores achieved by the models trained on the two

[4]https://huggingface.co/UBC-NLP/MARBERT

[5]https://huggingface.co/aubmindlab/bert-base-arabertv02

[6]https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter

[7]https://huggingface.co/qarib/bert-base-qarib

[8]https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix

Figure 1: Topics comparison using keywords in ASAD (left) and ArSAS (right) training sets

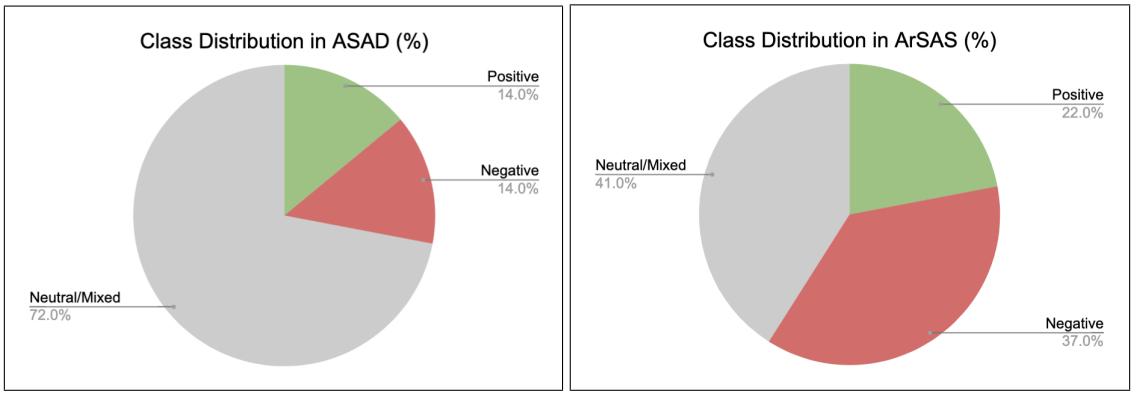

Figure 2: Class distribution (%) in ASAD (left) and ArSAS (right) training sets

datasets separately, at an average of 0.53 and relatively high standard deviation of 0.19, which meant that this wasn't the optimal solution to the generalization task.

In this case, we hypothesized that the difference in class distribution may have had a major impact on model performance. As a result, we extracted a portion of ASAD in order to balance the class distribution with ArSAS. This approach resulted in a major increase in F-1 score when **merged_balanced_distribution** was evaluated on the unseen test sets, when compared to **merged_ArSAS_ASAD**, as shown in Table 2.

To understand the generalizability of models further, we also experiment with a counter-labelling approach. This included using **ArSAS_model** to label ASAD training set, and comparing the actual labels in the training set with **ArSAS_model** predicted labels, and keeping the ones that are agreed on. If one of the labels is Neutral/Mixed, while the other label is non-neutral, the non-neutral label is kept as the 'gold label'. This was done to ensure that any characteristics of non-neutral sentiment were appropriately represented in the resulting training dataset. The merged_counter-labeling had a slightly improved F-1 score when evaluated on the unseen test data. However, the F-1 score was much lower than previous experiments when evaluated on ASAD dataset. This could be due to the fact that ASAD dataset contained a high number of Neutral/Mixed datapoints, while this model performed well on identifying tweets of a Positive or Negative sentiment.

These experiments showed improvement in generalizability of the resulting models, where the **merged_balanced_distribution** achieved the highest average f-1 score of 0.59, with a standard deviation of 0.07. However, since the experiments focused on ArSAS and ASAD datasets, which are limited in their topics and/or dialects coverage, we decided to create models using an entirely different and larger training dataset, to effectively investigate model generalization.

Training on a New Dataset

Dataset Characteristics To train on completely unseen data, we used Twitter API with language filter set to "AR" and collected approximately 2M tweets that were created between 2009 and 2020. Our main exploration is regarding the generalization of machine learning models, no other attributes, such as geographical area or keywords, were specified, to ensure that the data extracted would be as versatile as possible.

Data Preprocessing Each of the tweets went through standard preprocessing to ensure that none of the usual noisy data that comes with social media posts such as tweets would be considered when performing our experiments. This included removing any English characters/words from the tweets, to allow the processing of Arabic words only. We only considered tweets that contain at least 5 Arabic words.

Exp.	Model	#Tweets	ASAD	ArSAS	ATSAD	Egypt_data	Saudi_data	Av. F-1	Av. SD
1	ArSAS_model	16.2K	0.36	**0.75**	0.51	0.58	0.59	0.55	0.13
2	ASAD_model	50K	**0.80**	0.42	0.13	0.25	0.30	0.38	0.23
3	merged_ArSAS_ASAD	66.2K	**0.80**	0.72	0.40	0.31	0.42	0.53	0.19
4	merged_balanced_distribution	35.4K	0.66	0.68	0.50	0.54	0.55	**0.59**	**0.07**
5	merged_counter-labeling	32.3K	0.30	**0.75**	**0.59**	**0.62**	**0.60**	0.51	0.14

Table 2: Detailed evaluation results for models merging ASAD and ArSAS datasets

Furthermore, all Arabic diacritics were removed, and decorated and Farsi letters were mapped to original letters. In addition, all characters were normalized (ex: mapping different shapes of Alif to plain Alif). - Text normalization of hamza, and different forms of (ا، ي، ة)

Experiments[9] We experimented with different approaches to automatically label the dataset, then a linearSVC model with n-gram count vectorizer followed by a TF-IDF transformer was trained and evaluated on the five different test sets. We report the macro-average F-1 scores for all the experiments throughout this paper, since the class distributions within the test sets are not equal. We also report average F1 (Av. F1) and average standard deviation (Av. SD).

Firstly, we labelled the dataset using **ASAD_model** and **ArSAS_model** separately, and kept the tweets that were given the same label by both models. This resulted in a dataset of approximately 450K tweets and their corresponding sentiment. A linearSVC model, ArSAS_ASAD_SVM_agreement, was then trained on the resulting dataset. As shown in Table 3, when evaluated on all the test sets, there was a slight improvement in the F-1 scores in terms of consistency across all five test sets, when compared to the previous models shown in Table 2.

Using the same dataset, we also fine-tuned bert-base-Arabertv02-Twitter[10] model, to see whether using BERT would improve the quality of the predictions. We split the 450K dataset into 400k training and 50k validation to perform this experiment.

Although the BERT model reached an F-1 score of 0.98 on the validation set, when evaluated on the test sets, there was only a slight improvement in the F-1 scores, 0.02-0.04, as shown in Table 3. Although the slight increase in scores is promising, the long training time, complexity and large size of the finetuned BERT model are things to take into consideration when working with BERT.

Next, we proceeded to experiment with using lexicon-based approaches to automatically label the training data, which included both keyword and emoji based lexicons. In this case, we initially used a random portion of the dataset, amounting to approximately 200K tweets, or 10% of the full dataset. The goal was to perform multiple experiments with several variations. And due to the computational complexity of the conducted experiments, using the full 2M dataset would not be efficient at this point.

For the keyword-based approaches, we used the Arabic keyword sentiment lexicon presented by Kiritchenko, Mohammad, and Salameh (2016) in SemEval-2016, which contained phrases and words extracted from multiple sources including Twitter. This experiment entailed going through each of the tweets in the training set, and each tweet was given an average score (between -1 and 1) depending on the words/phrases it contains, and their corresponding scores in the lexicon. If no keyword was found in the tweet, it was removed, which resulted in a dataset of 125k tweets. Using the scores, if any tweet had an average score of more than 0.2, it was considered positive, while any tweet less than -0.2 was negative, and everything in between was considered neutral. Those scores were used since we conducted multiple experiments and these scores gave us the most balanced distribution between the classes. Once the **keyword_model** was trained on the labeled data, the scores, shown in Table 4, were much lower than previous experiments, such as the ones detailed in Table 2 and Table 3, which resulted in exploring other options such as emoji-based approaches, since "considering Emoji in sentiment analysis [can] help improve overall sentiment scores" (Ayvaz and Shiha 2017).

We started with a large emoji sentiment lexicon proposed by Kralj Novak et al. (2015) in which they analyzed 1.6M English/European tweets. In the **emoji_score_model**, we deployed a similar technique to the one used for the **keyword_model** which entailed calculating an average sentiment score of the emojis within each tweet, and used 0.2 and -0.2 as the threshold scores. As shown in Table 4 the **emoji_score_model** achieved higher scores for test sets that do not contain any neutral sentiment, such as ATSAD and Egypt_data, in comparison with ASAD and ArSAS which contain neutral tweets.

Since the **keyword_model** performed better on identifying neutral sentiment (F-1 scores on the ASAD and ArSAS test sets were 0.50 and 0.33 respectively), we decided to take both keywords and emojis into account when determining a sentiment score for each tweet using a similar technique. Although the **emoji_keyword_model** approach slightly improved identifying neutral sentiment, the F-1 scores were still lower than previous experiments shown in Table 3, as it slightly hindered the ability of the emojis to identify non-neutral sentiment, and the results were not entirely consis-

[9]Best scores within each table are in bold, while best performing scores across all experiments are underlined.

[10]The BERT model is trained on approx. 60M Arabic Tweets, and can be found here; https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter

Exp.	Model	#Tweets	ASAD	ArSAS	ATSAD	Egypt_data	Saudi_data	Av. F-1	Av. SD
6	ArSAS_ASAD_SVM_agreement	450K	0.66	0.64	**0.43**	0.57	0.52	0.56	**0.08**
7	ArSAS_ASAD_BERT_agreement	450K	**0.71**	**0.65**	0.41	**0.59**	**0.55**	**0.58**	0.11

Table 3: Detailed evaluation results for 2M random tweets models agreement experiments

tent across all test sets, as shown in Table 4.

Furthermore, we decided to explore using an Arabic-based emoji sentiment lexicon that was created by Hakami, Hendley, and Smith (2021). We refer to this experiment as **Ara_emoji_score**. We applied the same technique we previously used with the **emoji_score_model**, and found a substantial increase in F-1 score for the binary test sets without any neutral sentiment. For instance, the **emoji_score_model** and **Ara_emoji_score** had F-1 scores of 0.42 and 0.62 when evaluated on the Saudi_data test set, respectively.

As an extra step, we decided to only take into consideration the top 50 most common/used positive and negative emojis, as well as the top 100 neutral emojis from the lexicon, and discard of the rest of the emojis, depending on the total occurrence of each emoji according to the lexicon created by Hakami, Hendley, and Smith (2021). In doing so, **Ara_common_emoji_score** had a higher average F-1 score of 0.50, and showed an increase particularly when it comes to identifying neutral sentiment, since when evaluated on the ASAD test set, for instance, it scored a 0.33, which is approximately 0.14 more than that of the **Ara_emoji_score** (experiment 11 in table 4). It is also important to note that it had consistent evaluation scores on the non-neutral test sets, namely: ATSAD, Egypt_data, and Saudi_data.

As a result, we applied this approach on the full set of 2M tweets, to create a larger training dataset. In doing so, we ended up with a training set of approximately 399K, which we then used to train a LinearSVC model. As shown in Table 5, the model named **Ara_common_emoji_score** achieved a consistent score of approximately 0.6 on non-neutral test sets. However, the major issue was its low ability in appropriately identifying neutral tweets.

Therefore, we decided to create a cascaded model that consists of two parts. Due to its effectiveness in appropriately predicting neutral sentiment, the first model is a binary classification model that applies the same technique used to train **merged_balanced_disribution**, referenced in Table 2. In this case, however, it was trained on labelling tweets as either neutral or not neutral. When evaluated on its own, this binary model achieved a relatively high F-1 scores across all test sets, with an average F-1 score of 0.84, which shows its effectiveness in differentiating between neutral and non-neutral sentiment. This model was then given the 2M tweets, and any tweets labelled as 'NOT NEUTRAL' were then given to the second model, **Ara_common_emoji_score**, to be labelled as either positive or negative. Any tweets from the 'NOT NEUTRAL' tweets labelled as neutral by the **Ara_common_emoji_score** model were ignored. Once the full 2M tweets were labelled using both models, the dataset was then used to train a new SVM model named **cascaded_models**, which showed major improvement in the F-1 score when evaluated on ASAD and ArSAS test sets that mostly contain neutral tweets.

As discussed earlier previously, balancing class distribution positively affects the model performance, so we balanced the positive and negative classes in the dataset and trained a new model named **cascaded_models_balanced** on 1.64M tweets; 470K positive, 470k negative, and 690K neutral. The resulting F-1 score is the most consistent across all test sets with an average F-1 score of around 0.6, and a standard deviation of 0.04, which is the lowest out of all the models, as shown in experiment 16 of table 5.

To further evaluate the strength of the training set used for **cascaded_models_balanced** model, we fine-tuned a BERT model, **BERT_cascaded_models_balanced**, using the same training set, and found that when evaluated on the test-sets, the F-1 scores of the BERT model were consistent with the scores of **cascaded_models_balanced**, both outperformed the rest of the models in terms of consistency and general accuracy. However, it is important to note that the 2.2 MB SVM-based model is much lighter than the transformer-based model with a size of 516 MB, **BERT_cascaded_models_balanced**, and achieved comparable evaluation results with less computational needs.

Ethics and Social Impact

With this exploration, there are ethical and social considerations to keep in mind.

User Privacy

To comply with Twitter Privacy Policy, we share the tweets by IDs, and ensure that the corresponding tweet texts as well as the user handles, are not shared.

Biases

It is important to note that any biases in the proposed dataset were unintentional. Due to the random nature of data collection over the specified period of time, we understand that some countries and/or dialects could be represented within the data more than others. However, this could be due to the fact that Twitter usage in certain countries is much higher than that of other countries within the region. The same applies to the topics represented within the dataset, as we tried to minimize bias towards a certain topic by having a prolonged data collection time-frame between the years 2009 and 2020.

Recommendations

Throughout our investigation, we found some aspects that are essential to keep in mind when considering improving

Exp.	Model	#Tweets	ASAD	ArSAS	ATSAD	Egypt_data	Saudi_data	Av. F-1	Av. SD
8	keyword_model	125K	0.50	0.33	0.22	0.35	0.30	0.34	0.09
Using Emoji lexicon extracted from analyzing European tweets									
9	emoji_score_model	56.1K	0.21	0.12	**0.37**	**0.34**	**0.42**	0.29	0.11
10	emoji_keyword_model	71K	**0.49**	**0.34**	0.28	**0.34**	0.34	**0.36**	**0.07**
Using Arabic Emojis									
11	Ara_emoji_score	59.1K	0.19	0.24	**0.64**	0.53	**0.62**	0.44	0.19
12	Ara_common_emoji_score	37.3K	**0.33**	**0.35**	0.62	**0.63**	0.58	**0.50**	**0.13**

Table 4: Detailed evaluation results for 200K random tweets using lexicon-based approach

Exp.	Model	#Tweets	ASAD	ArSAS	ATSAD	Egypt_data	Saudi_data	Av. F-1	Av. SD
13	Ara_common_emoji_score	399K	0.32	0.37	0.66	0.64	0.58	0.51	0.14
15	cascaded_models	2M	0.66	0.49	0.63	0.60	0.56	0.59	0.06
16	**cascaded_models_balanced**	1.64M	**0.67**	**0.53**	0.63	0.61	0.58	0.60	**0.04**
17	**BERT_cascaded_models_balanced**	1.64M	0.66	0.51	**0.67**	**0.66**	**0.60**	**0.62**	0.06

Table 5: Results of the best performing models trained using 2M tweets

model generalization in Arabic sentiment analysis, as well as other classification tasks in Arabic and potentially other languages. First, simple merging of different datasets with variant characteristics may not be the ideal solution, and best results can be obtained from training a new model from scratch on data that many classifiers agree on their labels (compare the scores of experiments 3 and 6 on unseen test datasets in Table 2 and Table 3 respectively). This reduces the biases in each model and its data, and can capture the actual words and semantics that contribute to predicting the correct sentiment in the studied language.

Moreover, when exploring lexicon-based labelling approaches, such as emoji-based ones, using lexicons that are tailored to the specific language being explored proves to be beneficial when evaluating the model. In the case of Arabic sentiment analysis, using the Arabic emoji lexicon proved to be significantly better than the European-based emoji lexicon (compare the scores of experiments 9 and 11 in Table 4). It is also important to keep in mind that emojis are generally better than textual data, such as keywords, when used in automatic labeling of non-neutral sentiment, as they are more consistently used across different countries and allow to overcome the problems of dialectal variations and non-standard orthography, as shown in experiments 8 and 12 in Table 4.

Furthermore, as discussed previously, since three of the test datasets use a two-point classification system, while our models are trained using a three-point classification system, this negatively affected our evaluation scores, so it would be beneficial to evaluate any future approaches on datasets with similar classification systems to accommodate such difference, to accurately estimate their true capabilities in identifying sentiment. Consequently, we recommend using standard annotation labels for different classification tasks. Also, we recommend ensuring that classes distribution within the training dataset are equal since it has a positive effect on the generalization of the model on unseen data, as shown in the differences between experiments 3 and 4 in Table 2, and experiments 15 and 16 in Table 5.

Conclusion

In conclusion, we investigated the generalization of models and data in the context of the Arabic sentiment analysis task. We performed many experiments, including lexicon-based and counter-labelling, to test the performance of the models when tested on multiple diverse datasets that contain different topics and dialects, which are ASAD, ArSAS, ATSAD, Saudi_data, and Egypt_data.

Our best performing - and most consistent - approach, **cascaded_models_balanced**, takes many things into consideration. This includes the fact that the training data used was labelled using two separate models, one solely responsible for identifying neutral examples, while the other model, based on the Arabic emoji lexicon, was used to identify positive/negative tweets. Furthermore, the class distribution within the dataset was taken into consideration, to ensure that the classes are equally represented within the training data, specifically between the positive and negative classes. We share the dataset of approximately 1.64M tweets (through their IDs) and their corresponding sentiment classification used to train the cascaded_models_balanced with the community through Zenodo (Abdaljalil et al. 2023).

In future, we would like to expand this study further to cover other classification tasks as well as exploring other transformer models, both in Arabic and other languages. This will allow us to validate some of our conclusions as well as to refine these recommendations and learned lessons.

References

Abdaljalil, S.; Hassanein, S.; Mubarak, H.; and Abdelali, A. 2023. Towards Generalization of Machine Learning Models:

An Arabic Sentiment Analysis Dataset. https://doi.org/10.7910/DVN/LBXV9O.

Abdulla, N.; Ahmed, N. A.; Shehab, M.; and Al-Ayyoub, M. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. 1–6. ISBN 978-1-4799-2305-2.

Abu Kwaik, K.; Chatzikyriakidis, S.; Dobnik, S.; Saad, M.; and Johansson, R. 2020. An Arabic Tweets Sentiment Analysis Dataset (ATSAD) using Distant Supervision and Self Training. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 1–8. Marseille, France: European Language Resource Association. ISBN 979-10-95546-51-1.

Abuelenin, S.; Elmougy, S.; and Naguib, E. 2017. Twitter Sentiment Analysis for Arabic Tweets. 467–476. ISBN 978-1-4799-2305-2.

Al-Ayyoub, M.; Bani Essa, S.; and Alsmadi, I. 2014. Lexicon-Based Sentiment Analysis of Arabic Tweets. *International Journal of Social Network Mining (IJSNM)*, X: 0–0.

Al-Thubaity, A.; Alqahtani, Q.; and Aljandal, A. 2018. Sentiment lexicon for sentiment analysis of Saudi dialect tweets. *Procedia Computer Science*, 142: 301–307. Arabic Computational Linguistics.

Al-Twairesh, N.; Al-Khalifa, H.; and Al-Salman, A. 2016. AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 697–705. Berlin, Germany: Association for Computational Linguistics.

Alharbi, A. R.; and Aljaedi, A. 2019. Predicting Rogue Content and Arabic Spammers on Twitter. *Future Internet*, 11(11).

Alharbi, B.; Alamro, H.; Alshehri, M.; Khayyat, Z.; Kalkatawi, M.; Jaber, I.; and Zhang, X. 2021. ASAD: A Twitter-based Benchmark Arabic Sentiment Analysis Dataset. arXiv:2011.00578.

Aloqaily, A.; Al-hassan, K., Malak andSalah; Elshqeirat, B.; and Almashagbah, M. 2020. SENTIMENT ANALYSIS FOR ARABIC TWEETS DATASETS: LEXICON-BASED AND MACHINE LEARNING APPROACHES.

Alowaidi, S.; Saleh, M.; and Abulnaja, O. 2017. Semantic Sentiment Analysis of Arabic Texts. *International Journal of Advanced Computer Science and Applications*, 8(2).

Alshehri, A.; Nagoudi, E. M. B.; Abdul-Mageed, M.; and Alhuzali, H. 2018. Think Before You Click: Data and Models for Adult Content in Arabic Twitter. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Alyami, S. N.; and Olatunji, S. O. 2020. Application of Support Vector Machine for Arabic Sentiment Classification Using Twitter-Based Dataset. *Journal of Information & Knowledge Management*, 19(01): 2040018.

Ashir, A. 2021. A Generalized Method for Sentiment Analysis across Different Sources. *Applied Computational Intelligence and Soft Computing*, 2021: 1–8.

Ayvaz, S.; and Shiha, M. 2017. The Effects of Emoji in Sentiment Analysis. *International Journal of Computer and Electrical Engineering*, 9: 360–369.

Barbiero, P.; Squillero, G.; and Tonda, A. 2020. Modeling Generalization in Machine Learning: A Methodological and Computational Study. *ArXiv*, abs/2006.15680.

Doshi-Velez, F.; and Kim, B. 2018. *Considerations for Evaluation and Generalization in Interpretable Machine Learning*, 3–17. Cham: Springer International Publishing. ISBN 978-3-319-98131-4.

Elmadany, A.; Mubarak, H.; and Magdy, W. 2018. An Arabic Speech-Act and Sentiment Corpus of Tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, OSACT3 ; Conference date: 08-05-2018.

Esuli, A.; and Sebastiani, F. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Hakami, S. A. A.; Hendley, R.; and Smith, P. 2021. Arabic Emoji Sentiment Lexicon (Arab-ESL): A Comparison between Arabic and European Emoji Sentiment Lexicons. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 60–71. Kyiv, Ukraine (Virtual): Association for Computational Linguistics.

Heikal, M.; Torki, M.; and El-Makky, N. 2018. Sentiment Analysis of Arabic Tweets using Deep Learning. *Procedia Computer Science*, 142: 114–122. Arabic Computational Linguistics.

Kharde, V. A.; and Sonawane, S. 2016. Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications*, 139(11): 5–15.

Khoja, S.; and Garside, R. 1999. Stemming Arabic Text. Technical report, Lancaster University, Computing Department.

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)*, 42–51. San Diego, California: Association for Computational Linguistics.

Kora, R.; and Mohammed, A. 2019. Corpus on Arabic Egyptian tweets. https://doi.org/10.7910/DVN/LBXV9O.

Kralj Novak, P.; Smailović, J.; Sluban, B.; and Mozetič, I. 2015. Sentiment of Emojis. *PLOS ONE*, 10: 1–22.

Mhamed, M.; Sutcliffe, R.; Sun, X.; Feng, J.; Almekhlafi, E.; and Retta, E. A. 2021. Improving Arabic Sentiment Analysis Using CNN-Based Architectures and Text Preprocessing. *Computational Intelligence and Neuroscience*, 2021 5538791.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11): 39–41.

Mubarak, H.; Abdelali, A.; Darwish, K.; Eldesouki, M.; Samih, Y.; and Sajjad, H. 2019. A System for Diacritizing

Four Varieties of Arabic. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 217–222. Hong Kong, China: Association for Computational Linguistics.

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*, 2515–2519. Lisbon, Portugal: Association for Computational Linguistics.

Namugera, F.; Wesonga, R.; and Jehopio, P. 2019. Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. *Computational Social Networks*, 6.

Radcliffe, D.; and Abuhmaid, H. 2020. Social Media in the Middle East: 2019 in Review. *SSRN Electronic Journal*.

Shaalan, K.; Siddiqui, S.; Alkhatib, M.; and Monem, A. 2018. *Challenges in Arabic Natural Language Processing*, 59–83. ISBN 978-981-322-938-9.

Sharma, A.; and Ghose, U. 2020. Sentimental Analysis of Twitter Data with respect to General Elections in India. *Procedia Computer Science*, 173: 325–334. International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020.

Sharma, A.; Kedar, A.; Jadhav, H.; Gunjan; and Ospanova, A. 2022. A Survey on Sentiment analysis of Twitter using Machine Learning. *International Journal of Innovative Research in Science Engineering and Technology*, 11: 3721–3725.

Syarief, M. G.; Kurahman, O. T.; Huda, A. F.; and Darmalaksana, W. 2019. Improving Arabic Stemmer: ISRI Stemmer. In *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*, 1–4.

Wang, H.; Meghawat, A.; Morency, L.-P.; and Xing, E. P. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 949–954.