AraDepSu: Detecting Depression and Suicidal Ideation in Arabic Tweets Using Transformers

Mariam Hassib **Nancy Hossam** **Jolie Sameh** **Marwan Torki**
Faculty of Engineering, Alexandria University, Egypt
{mariamhassib1990,nancyhossam441999,joliesameh99}@gmail.com
mtorki@alexu.edu.eg

Abstract

Among mental health diseases, depression is one of the most severe, as it often leads to suicide which is the fourth leading cause of death in the Middle East. In the Middle East, Egypt has the highest percentage of suicidal deaths; due to this, it is important to identify depression and suicidal ideation.[1] In Arabic culture, there is a lack of awareness regarding the importance of diagnosing and living with mental health diseases. However, as noted for the last couple of years people all over the world, including Arab citizens, tend to express their feelings openly on social media. Twitter is the most popular platform designed to enable the expression of emotions through short texts, pictures, or videos. This paper aims to predict depression and depression with suicidal ideation. Due to the tendency of people to treat social media as their personal diaries and share their deepest thoughts on social media platforms. Social media data contains valuable information that can be used to identify users' psychological states. We create the AraDepSu dataset by scrapping tweets from Twitter and manually labeling them. We expand the diversity of user tweets, by adding a neutral label ("neutral") so the dataset includes three classes ("depressed", "suicidal", and "neutral"). Then we train our AraDepSu dataset on 30+ different Transformer-based models. We find that the best-performing model is MARBERT with accuracy, macro-average precision, macro-average recall, and macro-average F1-score values of 91.20%, 88.74%, 88.50%, and 88.75%.

1 Introduction

The well-being of a person comprises physical health and mental health. The mental health of a person shows the individual's state of mind. Mental disorders are a worldwide health problem affecting a large number of people and causing numerous deaths every year (Musleh et al., 2022).

Depression is one of the most well-known mental health disorders and it is considered a major issue for mental health practitioners. Depression is a mood disorder that causes a persistent feeling of sadness and loss of interest. Also called a major depressive disorder or clinical depression. It affects how you feel, think and behave and can lead to a variety of emotional and physical problems. Fortunately, it is also treatable especially if we identify it in the early stage.[2] In Arabic culture, early diagnosis of mental illness is difficult, because of the stigma of mental illness and lack of awareness in the field of psychiatry.[3] Depression has become a silent killer as it increases suicide risk. [4]

People tend to express their feelings openly on social media, especially on Twitter. Twitter provides a platform where users share their thoughts, emotions, feelings, and expressions. These tweets can aid in determining a person's thought process, mental health, and behavioral traits.

In this paper, our objective is to come up with a methodology to accurately classify and analyze Arabic tweets. We classify whether they are suffering from depression or depression with suicidal ideation which can help prevent suicidal deaths. We focus on the potential of Natural language processing (NLP) and machine learning techniques that can be utilized in the mental health field. NLP is very helpful when it comes to understanding the context of natural human language. As a result, it extracts latent meaning from text and creates AI-based solutions using text data available on social

[1] Suicide: The Fourth Cause of Death Among Young People. URL: https://www.bbc.com/arabic/59568886.

[2] What is Depression? URL: https://psychiatry.org/patients-families/depression/what-is-depression.

[3] Egypt: Mental health barriers URL: https://english.ahram.org.eg/NewsContent/50/1209/422608/AlAhram-Weekly/Focus/Egypt-Mental-health-barriers.aspx.

[4] Mental Health and Substance Abuse: Does Depression Increase The Risk For Suicide? URL: https://www.hhs.gov/answers/mental-health-and-substance-abuse/does-depression-increase-risk-of-suicide/index.html.

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 302 - 311
December 8, 2022 ©2022 Association for Computational Linguistics

media platforms.

2 Background

Depression is considered a global concern. It is a very common illness, as it affects people across all nations. Approximately 280 million people have recently been afflicted with depression in the world.[5] Depression can cause the affected person to suffer greatly and function poorly at work, at school, and in the family. At its worst, depression can lead to suicide. Over 700,000 people die due to suicide every year.

Depression is different from usual mood fluctuations and short-lived emotional responses to challenges in everyday life.[6] It comes in many forms, each accompanied by its own symptoms. The most common and known form is major depressive disorder (MDD), which influences the ability of individuals to do daily tasks (Aldarwish and Ahmad, 2017). Depression does not have a target age, as it may begin at a young age. Curbing depression is essential to saving people's lives (Marcus et al., 2012).

In Arabic culture, the stigma on mental illness is deeply entrenched, and there is a lack of awareness regarding this issue. The review reveals that beyond society and culture, the persistence of mental illness stigma. In the Arab world may be explained by inefficient monitoring mechanisms of mental health legislation and policies within the healthcare setting (Merhej, 2019).

3 Related Work

This section presents a summary of prior studies that have been conducted on the prediction and monitoring of depression and suicide using social media.

Various related data have been in the literature for the prediction of depression via various approaches. Two main strategies were reported in the literature to collect data and detect depression through social media.

The first strategy is crowd-sourcing data collection from social media publicly available. This allows researchers to cheaply outsource simple tasks or questionnaires, and gather data in real-time. It also helps to obtain far more numerous and widespread observations than in traditional data collection given its relatively low cost. The crowd-sourcing strategy is mainly conducted in two stages (De Choudhury et al., 2013a,b). First, responses from an online clinical depression survey are gathered. Then, contents are collected by accessing the Twitter data of the consented participants. This main strategy limitation is time-consuming.

The second alternative strategy is characterized by gathering data quickly and cheaply (Coppersmith et al., 2014). As such, the data is collected directly from social media that are publicly available for participants with self-identified mental illnesses. The disadvantage of this strategy is its low reliability. Unfortunately, very few of these collected data and applied models were found in Arabic.

Conducting a sentiment analysis of texts in the Arabic language is more complex than that directed toward English texts. That is because the Arabic language is characterized by more forms than other languages. The formal variant of Arabic is Modern Standard Arabic (MSA), but this is rarely used in spoken interactions. The most frequently used informal variant is Dialectal Arabic (DA), especially for communication purposes. A total of 30 major Arabic dialects differ from MSA, the approaches used to translate difficult MSA terms are ineffective when applied to DA translation. Recently, Arabic researchers have developed solutions for different dialects, but these remain minimally inaccurate and cover only a few dialects (Al-Twairesh et al., 2017). Our work mainly focuses on Egyptian Dialectal with it the most studied and widely spoken DA.

A common challenge that faces most depression detection trials is to identify the symptoms of mental illness in online health communities. This is due to the symptom overlapping between multiple mental illnesses. To the best of our knowledge, no previous trials went deeply into the Arabic Twitter data for detecting whether a user's tweet is depressed or depressed with suicidal ideation. To fill this literature gap, our solution helps in detecting signs relevant to depression using Arabic language tweets. To avoid the limitations related to data collection we faced in the beginning, we combine low-cost and reliable data collection strategies. We collected more than 20k tweets data from public tweets and labeled them manually.

[5] Institute of Health Metrics and Evaluation. Global Health Data Exchange. URL: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b.

[6] World Health Organization: Depression. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

Dataset	Number of examples
ArTwitter (Abdulla et al., 2013)	3,543
TEAD (Abdellaoui and Zrigui, 2018)	2,000
BRAD (Elnagar et al., 2018)	2,000
ASTD (Nabil et al., 2015)	1,590
Total	**9,133**

Table 1: Number of examples from different datasets.

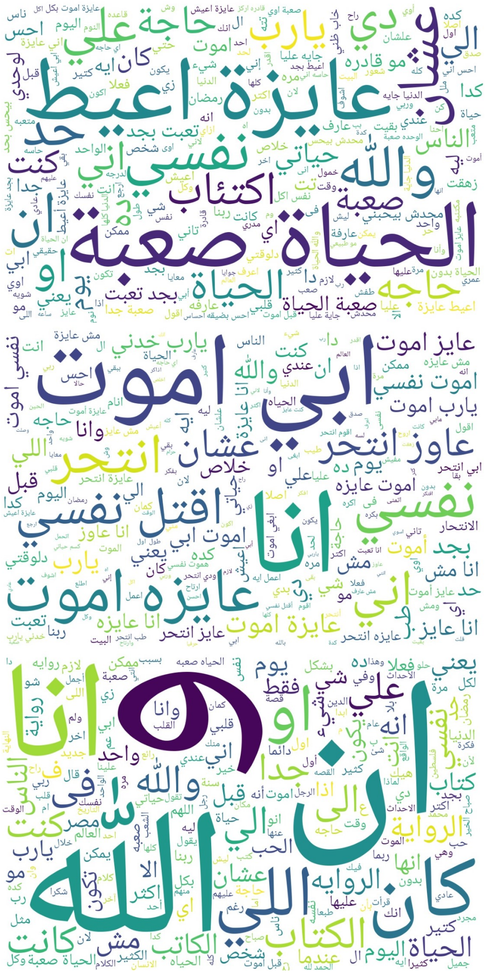

Figure 1: Word clouds for different classes in the AraDepSu dataset. Top: Depression words. Middle: Suicidal ideation words. words. Bottom: Non-depression words

4 Data

4.1 Data Collection

We extracted more than 10k tweets from different users with special keywords to get tweets with depression and suicidal ideation, posted between 2016 and 2022. We added to our dataset 1,230 records from the available data of the Modern Standard Arabic mood changing and depression dataset (Maghraby and Ali, 2022). Table 2 shows some of the depression keywords and Table 3 shows some of the depression with suicidal ideation keywords. Tweets in this study were a mixture of Modern Standard Arabic and Arabic dialects. Similar to the previous work on depression detection on English datasets (Babu and Kanaga, 2022), we collected data from different sentiment analysis datasets as shown in Table 1.

4.2 Cleaning and Pre-Processing Data

Our dataset annotation procedure includes two phases. In the first phase, we sanitized each tweet so that they do not contain irrelevant text, so they would be suitable input for our various models. First, we removed hyperlinks because they do not add much to the actual content of the tweet. Then, we removed empty columns and duplicate records.

4.3 Manually Labeling Process

In the second phase, each record is labeled by one category name, whether it is depression, depression with suicidal ideation, or non-depression. The annotators followed the authors' instructions in labeling the data. Each record was labeled by a single annotator. Then, the authors revised the annotated data sample by sample. In case of disagreement, the authors' decision is favored. Finally, we obtained a dataset with 20,213 tweets 5,472 classified with depression, 2,167 with suicidal ideation, and 12,574 as non-depression as shown in Table 4

In Figure 1 we show the word clouds for the different classes in AraDepSu dataset. The keywords for the depression class are highlighted in the depression words such as “life is hard”, “I want to cry" and similar keywords. We observe the same kind of keywords for the suicidal ideation class such as “ kill my self”, “I wan to die” and similar keywords. For the non-depression class, the highlighted keywords are not relevant to a specific

Keyword	Example
تعبت Exhausted	تعبت و زهقت و جبت اخري من كل حاجة I am exhausted and fed up with everything.
مكتئب Depressed	انا مكتئب أكتئاب حاد وبحاول افضل واقف على رجلي عشان اهلي بس I am severely depressed, just trying to withstand it for my family’s sake
منهارة Broken	قد يُخيل لكم إني قوية وثابتة بينما أنا في الحقيقة منهارة وبعيط You might think I’m strong and steady when in fact I’m broken and weeping.
حزينة Miserable	أنا حزينة ومقهورة أوي والله حتي مش عارفه اهرب من كل ده وانام I’m so miserable and defeated,I do not even know how to escape from all this and just sleep
اكتئاب Depression	فيني اكتئاب حاد I have severe depression
محدش بيحبني No one loves me	كنت عايزة اعرف بس انا ليه محدش بيحبني زي مابحبه ولا بفرق مع حد؟ I just want to know, why no one loves me as much as I do and I do not matter to anyone?
عايزه اعيط Want to Cry	انا عايزه اعيط او انام I want to cry or sleep

Table 2: Examples for depression from the annotated corpus.

Keyword	Example
عايز انتحر I want to commit suicide	عايز انتحر و اموت نفسي و اخلص I want to commit suicide and just end my life
اقتل نفسي kill myself	اقتل نفسي او اقتل نفسي مافيه خيار ثالث It is either killing myself or killing myself there is no other option.
ودي انتحر I want to commit suicide	ودي انتحر صراحه تعبت والله I want to commit suicide, honestly I’m tired, I swear
عايز اموت I want to die	لا ماضي ولا حاضر عايز اموت No past no future, I want to die.
مش عايزة اعيش I don’t want to live	موتني يا رب أنا مش عايزة اعيش كفاية كدة Just kill me God, I do not want to live anymore that is enough.
بفكر انتحر Thinking about committing suicide	قاعدة بعيط وبفكر انتحر Crying and thinking about committing suicide
يارب خدني Just take me God	يارب خدني من العيلة دي في اقرب وقت Just take me God from this family as soon as possible

Table 3: Examples for depression with suicidal ideation from the annotated corpus.

topic.

5 Experiments and Results

5.1 Dataset

The final dataset consists of 20,213 tweets divided into 15,159 training tweets and 5,054 testing tweets. Table 4 provides the statistical details of the dataset.

5.2 Models

In our experiments, we use the following models:

5.2.1 mBERT

Multilingual BERT model (Devlin et al., 2018), is a single language model pre-trained from monolingual corpora on data from the Wikipedia dumps of 104 languages.

5.2.2 GigaBERT

GigaBERT is a customized bilingual BERT for English and Arabic. We use two variants of this model, GigaBERT-v3 and GigaBERT-v4. GigaBERT-v3 is a customized bilingual BERT for English and Arabic. It is pre-trained on a large-scale corpus with 10B tokens. GigaBERT-v4 is a continued pre-training of GigaBERT-v3 on code-switched data (Lan et al., 2020).

5.2.3 XLM-RoBERTa

XLM-RoBERTa is an Unsupervised Cross-lingual Representation Learning at Scale (Conneau et al., 2019). This model is pre-trained on 2.5TB of filtered data containing 100 languages. We use two variants of this model, XLM-RoBERTa-base, and

Set	Non-depression	Depression	Depression With Suicidal Ideation	Total
Training	9,408	4,117	1,634	15,159
Testing	3,166	1,355	533	5,054
Total	12,574	5,472	2,167	20,213

Table 4: Distribution of depression and depression with suicidal ideation.

Model	Precision	Recall	F1-Score	Accuracy
mBERT	84.25	87.04	85.55	87.93
GigaBERT(v3)	86.21	87.44	86.80	88.90
GigaBERT(v4)	87.05	87.59	87.32	89.35
XLM-RoBERTa-base	86.32	87.31	86.79	89.06
XLM-RoBERTa-large	85.95	87.28	86.59	88.88
AraBERT-base(v01)	86.23	86.39	86.30	88.66
AraBERT-base(v1)	85.78	86.78	86.27	88.29
AraBERT-base(v02)	87.42	87.75	87.58	89.73
AraBERT-base(v02)-twitter	87.02	88.66	87.81	89.73
AraBERT-base(v2)	86.36	86.15	86.25	88.68
AraBERT-large(v02)-twitter	87.48	88.33	87.90	89.93
AraELECTRA(discriminator)	86.36	87.97	87.14	89.24
AraELECTRA(generator)	82.82	87.27	84.78	87.34
Arabic BERT-base	86.05	86.78	86.41	88.52
Arabic BERT-mini	83.80	86.23	84.94	87.67
Arabic BERT-medium	84.97	84.82	84.89	87.57
Arabic BERT-large	86.64	86.91	86.76	88.74
Arabic ALBERT-base	85.86	86.38	86.12	88.43
Arabic ALBERT-large	86.43	86.01	86.20	88.48
Arabic ALBERT-xlarge	86.82	85.62	86.21	88.70
MARBERT	**88.74**	88.50	**88.75**	**91.20**
MARBERT(v2)	87.75	88.50	88.12	90.07
ARBERT	86.42	86.21	86.31	88.60
QARiB	88.20	88.26	88.23	90.13
AraGPT2-base	83.34	85.70	84.45	86.94
AraGPT2-medium	81.08	83.57	82.31	83.83
AraGPT2-large	83.97	84.60	84.26	84.67
AraT5-base	84.44	88.68	86.35	88.70
AraT5-msa-base	82.74	88.66	85.26	87.73
AraT5-tweet-base	86.06	**88.93**	87.40	89.65
AraT5-msa-small	74.90	82.77	77.74	81.58
AraT5-tweet-small	80.47	85.95	82.12	85.26

Table 5: Performance comparison of different models on our dataset.

XLM-RoBERTa-large.

5.2.4 AraBERT

AraBERT is an Arabic pretrained language model based on Google's BERT architecture and uses the same BERT-Base config (Antoun et al.). There are many versions of the model. AraBERTv0.1 and AraBERTv1, with the difference being that AraBERTv1 uses Farasa Segmenter (Durrani and Mubarak). AraBERT(v01/1) was trained on 23GB of text while AraBERT(v02/2) was trained on 77GB of text. AraBERTv0.2-Twitter-base/large are two new models for Arabic dialects and tweets.

They are trained on 60M Arabic tweets with emojis in their vocabulary in addition to common words that were not present at earlier versions. We use many variants of this model, AraBERT-base(v01/1/02/2) and AraBERT-base/large(v02)-twitter.

5.2.5 AraELECTRA

ELECTRA is a method for self-supervised language representation learning. AraELECTRA was trained on the same 77GB of text used for AraBERT (Antoun et al., 2021a). We use two variants of this model, AraELECTRA generator and AraELECTRA discriminator.

5.2.6 Arabic BERT

Arabic BERT Base model was pretrained on 8.2 Billion words of the Arabic version of OSCAR (Suárez et al., 2020) filtered from Common Crawl and a recent dump of Arabic Wikipedia and other Arabic resources which sum up to 95GB of text (Safaya et al., 2020). We use four variants of this model, Arabic BERT-base/mini/medium/large.

5.2.7 Arabic ALBERT

An Arabic edition of ALBERT model which was pretrained on 4.4 Billion words from the Arabic version of the unshuffled OSCAR corpus (Suárez et al., 2020) and the Arabic Wikipedia (Safaya, 2020). We use three variants of this model, Arabic ALBERT-base/large/xlarge.

5.2.8 ARBERT and MARBERT

ARBERT and MARBERT are based on the BERT-base architecture. ARBERT is a language model that is focused on Modern Standard Arabic (MSA) and was trained on 61GB of text from news articles. MARBERT is a language model that is focused on both Dialectal Arabic (DA) and MSA. MARBERT was trained on randomly sampled 1B Arabic tweets from a dataset of about 6B tweets, the dataset makes up 128GB of text. (Abdul-Mageed et al., 2021). MARBERTv2 was further trained on the same data as ARBERT in addition to AraNews dataset (Ali et al., 2021).

5.2.9 QARiB

QARiB is a QCRI Arabic and Dialectal BERT model, which was trained on 420 Million tweets and 180 Million sentences of text (Abdelali et al., 2021).

5.2.10 AraGPT2

AraGPT2 is an advanced Arabic language generation model, trained from scratch on a large Arabic corpus of internet text and news articles (Antoun et al., 2021b). We use three variants of this model, AraGPT2-base/medium/large.

5.2.11 AraT5

AraT5 Text-to-Text Transformers for Arabic Language Generation that is focused on both Dialectal Arabic (DA) and MSA. AraT5-MSA was trained on 70GB of text. AraT5-Tweet was trained on 178GB of text (Nagoudi et al., 2022).

5.3 Hyper-parameters Setting and Evaluation

In our experiments, we use the implementation provided by HuggingFace Transformers library (Wolf et al., 2019). We train our models for 5 epochs with a learning rate of 2e–5 and a maximum sequence length set to 128 tokens. Table 5 shows the results of different models on our dataset. The best-performing model is MARBERT with a macro-average F1-score of 88.75%.

6 Discussion

Models pre-trained on multiple languages
Table 6 compares the models pre-trained on multiple languages. GiagBERT outperforms mBERT and XLM-RoBERTa on AraDepSu dataset. We think the reason is that AraDepSu contains Arabic dialectic tweets and GiagBERT is trained only on English and Arabic data.

Models pre-trained on tweets
Table 7 compares the models pre-trained on tweets with the nWords of the pre-trained dataset and the f1-score results. MARBERT outperforms AraT5 even though it is trained on more data. We think the reason is that the majority of AraT5-tweet data is MSA according to the analyses done by (Nagoudi et al., 2022), and the majority of our dataset is from dialect tweets.

Models pre-trained on Modern Standard Arabic
Table 8 compares the models pre-trained on MSA with the size of the pre-trained dataset. AraBERT-base(v02) outperforms models pre-trained on larger datasets. In these models, the performance relies more on the architecture than on the dataset size.

Qualitative Evaluation As shown in the study, pre-trained models produced reliable results and accuracy. However, there were some drastic differences in their training circumstances. As stated

Model	Pre-trained languages	F1-Score
mBERT	104	85.55
GigaBERT(v3)	En-Ar	86.80
GigaBERT(v4)	En-Ar	**87.32**
XLM-RoBERTa-base	100	86.79
XLM-RoBERTa-large	100	86.59

Table 6: Comparison of different models pre-trained on multiple languages.

Model	Pre-trained tweets	F1-Score
AraBERT-base(v02)-twitter	60M	87.81
AraBERT-large(v02)-twitter	60M	87.90
MARBERT	1B	**88.75**
QARiB	420M	88.23
AraT5-tweet-base	1.5B	87.40
AraT5-tweet-small	1.5B	82.12

Table 7: Comparison of different models pre-trained on tweets.

Model	**DataSet Size**	**F1-Score**
AraBERT-base(v01)	23GB	86.30
AraBERT-base(v1)	23GB	86.27
AraBERT-base(v02)	77GB	**87.58**
AraBERT-base(v2)	77GB	86.25
AraELECTRA(discriminator)	77GB	87.14
AraELECTRA(generator)	77GB	84.78
Arabic BERT-base	95GB	86.41
Arabic BERT-mini	95GB	84.94
Arabic BERT-medium	95GB	84.89
Arabic BERT-large	95GB	86.76
Arabic ALBERT-base	35GB	86.12
Arabic ALBERT-large	35GB	86.20
Arabic ALBERT-xlarge	35GB	86.21
AraGPT2-base	77GB	84.45
AraGPT2-medium	77GB	82.31
AraGPT2-large	77GB	84.26

Table 8: Comparison of different models pre-trained on MSA.

previously, the core difference is that MARBERT focuses on Dialectic data in its training, while AraBERT focuses on Modern Standard Arabic (MSA) data. Since AraDepSu dataset is mainly composed of scraped tweets, there were many different dialects. This justifies why MARBERT produced the best accuracy and is considered the best model for this study.

We show in Table 9 the predictions of MARBERT and AraBERT-base(v02) on some test tweets. We observe that MARBERT excels with different dialects and tricky tweets. Those tricky tweets may address depression or suicidal depression in general, but can not be used as evidence that the user is depressed, or define their current state. This may result in a conflict between the prediction and the ground truth. The main reason for this error was believed to be that the pattern the model was searching for to label the string, as depression, for example, was found successfully but the human

Sentence	Ground Truth	Prediction	Pre-trained Data	Model
حاسس ان انا بحب الصيف لوحدي والله I feel that I love summer alone, I swear to God.	Non-depression	Non-depression Depression	Dialectic MSA	MARBERT AraBERT-base(v02)
شكلي كدا هفضل لغاية ما اموت مش عارفة انا عايزة ايه It looks like I will not know what I want until I die.	Non-depression	Depression Depression	Dialectic MSA	MARBERT AraBERT-base(v02)
ترا النهايه ليلى بتموت مبنشوف غير بهار حزينه At the end Layla will die and you will only see Bahar sad.	Non-depression	Non-depression Depression	Dialectic MSA	MARBERT AraBERT-base(v02)
انا اجذب ناس مدمره نفسيا وهالشيء مزعج I attract psychologically destructive people and this is annoying.	Depression	Depression Non-depression	Dialectic MSA	MARBERT AraBERT-base(v02)
احس يدي بتنكسر ابي احمل التيك توك قبل اموت خلاص طفشت I feel my hand breaking, want to download Tik Tok before I die, I am so bored.	Depression	Depression Non-depression	Dialectic MSA	MARBERT AraBERT-base(v02)
مليت وانا اتضايق روحي حد يتضايق وياي I am bored of being upset, I want someone to be upset with me.	Depression	Depression Depression	Dialectic MSA	MARBERT AraBERT-base(v02)
انتحر و اموت نفسي و اخلص Is the only resort to commit suicide and end my life.	Suicidal Ideation	Suicidal Ideation Suicidal Ideation	Dialectic MSA	MARBERT AraBERT-base(v02)
بصراحة الحياة صعبه اوي .. انا مش عايز اكمل Life is too hard I do not want to continue with it.	Suicidal Ideation	Suicidal Ideation Suicidal Ideation	Dialectic MSA	MARBERT AraBERT-base(v02)
اتمنى اني اموت ولا احس ب الي احسه الحين يارب مو قادره ولله I wish to die and not feel what I am feeling now, Lord I just cannot anymore.	Suicidal Ideation	Suicidal Ideation Suicidal Ideation	Dialectic MSA	MARBERT AraBERT-base(v02)

Table 9: Qualitative Evaluation: Predictions of different models on sample tweets from the test data.

common sense factor was missing.

7 Conclusion

This study enables intelligent instruments to identify and predict depression symptoms and suicide ideation from Arabic text based on depression-related words. This paper proposed computational approaches for the utilization of Arabic tweets. We scraped data from tweeter with keywords that act as depression triggers and labeled them manually.

In conclusion and based on the results discussed above, Arabic people do share their feelings on Twitter. The results prove that depressed people show specific behaviors within their tweets. They often use negative words to describe their symptoms, like suicidal thoughts or sleeping disorders.

We built a predictive model to predict whether a user's tweet is depressed or depressed with suicidal ideation. We examined the performance of all the above classifiers using a dataset collected from Twitter and labeled manually with truth labels ("depressed", "suicidal", "neutral"). We found the best accuracy with the MARBERT classifier at 91.20%.

Acknowledgement

We thank many volunteers at Alexandria University for their help in labeling and cleaning our dataset, in particular, Maram Attia, Amina Mohamed, Sara Khaled, Khlod Mohamed, Hagar Abouroumia, Ziyad Mohamed and Karim Elsayed.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

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. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. 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 7088–7105, Online. Association for Computational Linguistics.

Nawaf Abdulla, N Mahyoub, M Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Corpus-based and lexicon-based. In *Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.

Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Al-Salman, and Yousef Al-Ohali. 2017. Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:63–72.

Maryam Mohammed Aldarwish and Hafiz Farooq Ahmad. 2017. Predicting depression levels using social media posts. In *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*, pages 277–280. IEEE.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. Arafacts: the first large arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021a. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021b. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nirmal Varghese Babu and E Kanaga. 2022. Sentiment analysis in social media data for depression detection using artificial intelligence: A review. *SN Computer Science*, 3(1):1–20.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ahmed Abdelali Kareem Darwish Nadir Durrani and Hamdy Mubarak. Farasa: A fast and furious segmenter for arabic.

Ashraf Elnagar, Leena Lulu, and Omar Einea. 2018. An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia computer science*, 142:182–189.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An empirical study of pre-trained transformers for arabic information extraction. *arXiv preprint arXiv:2004.14519*.

Ashwag Maghraby and Hosnia Ali. 2022. Modern standard arabic mood changing and depression dataset. *Data in Brief*, 41:107999.

Marina Marcus, M Taghi Yasamy, Mark van van Ommeren, Dan Chisholm, and Shekhar Saxena. 2012. Depression: A global public health concern.

Rita Merhej. 2019. Stigma on mental illness in the arab world: beyond the socio-cultural barriers. *International Journal of Human Rights in Healthcare*.

Dhiaa A Musleh, Taef A Alkhales, Reem A Almakki, Shahad E Alnajim, Shaden K Almarshad, Rana S Alhasaniah, Sumayh S Aljameel, and Abdullah A Almuqhim. 2022. Twitter arabic sentiment analysis to detect depression using machine learning. *CMC-COMPUTERS MATERIALS & CONTINUA*, 71(2):3463–3477.

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.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Ali Safaya. 2020. Arabic-albert.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv:2006.06202*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Appendix

In figure 2, we show the confusion matrices for the best model MARBERT and AraBERT. MARBERT is the best model based on Dialectal Arabic and AraBERT is the best model based on MSA. Both models produce close results for the depression class. However, the confusion between the non-depression and suicidal ideation is more present in the AraBERT confusion matrix.

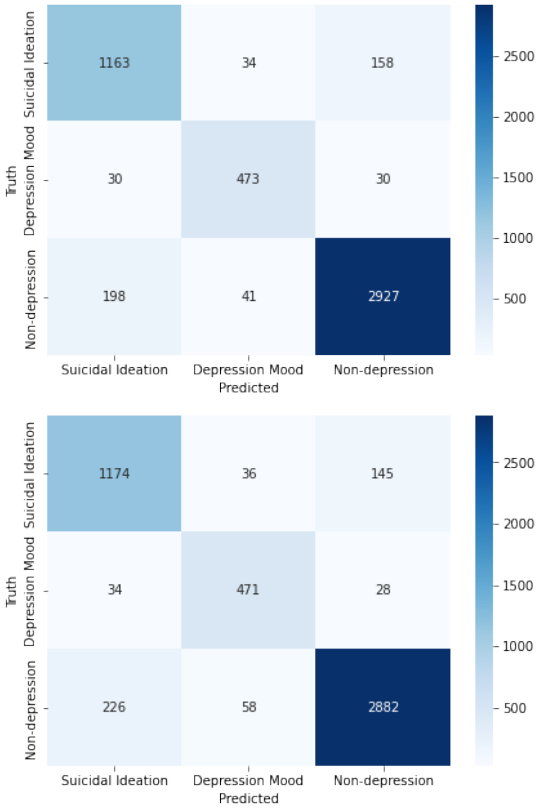

Figure 2: Top: MARBERT confusion matrix. Bottom: AraBERT-base(v02) confusion matrix.