ULRR

# Trustworthiness of health-related information on Arabic social media


| | |
|---|---|
| Item Type | Thesis |
| Authors | Albalawi, Yahya |
| Download date | 2026-01-18 05:06:35 |
| Item License | https://creativecommons.org/licenses/by-nc-sa/4.0/ |
| Link to Item | https://doi.org/10.34961/researchrepository-ul.21666269 |

# Trustworthiness of Health-Related Information on Arabic Social Media


Yahya Albalawi

Lero, the Irish Software Research Centre

Faculty of Science and Engineering

Department of Computer Science and Information Systems

University of Limerick


Submitted to the University of Limerick for the degree of

Doctor of Philosophy (PhD), 2022

1. Supervisor: Prof. Jim Buckley

   Lero, the Irish Software Research Centre

   University of Limerick

   Ireland

2. Supervisor: Dr. Nikola Nikolov

   Department of Computer Science and Information Systems

   University of Limerick

   Ireland

1. External Reviewer:

   Prof. Arsalane Zarghili

2. Internal Reviewer:

   Prof. Ita Richardson

1. Chair: Dr. Katie Crowley

Day of defense: 9th May 2022

Signature from head of PhD committee

# Abstract

Social media (SM) platforms play a vital role in disseminating health-related information. However, evidence suggests that Twitter posts (i.e., tweets) are often inaccurate; for example, research from Saudi Arabia indicates that 50% of health-related tweets contain inaccurate information. Previous studies also suggest that tweets do not need to be evidence-based or accurate to gain traction, which exacerbates the accuracy concern in the sphere of health information. The goal of the thesis is to develop a framework for automatically determining the accuracy of health-related tweets.

Knowing the accuracy of tweets offers the potential to recommend/promote accurate tweets while identifying/flagging/demoting inaccurate tweets. As a first step, this thesis employed a pilot study to identify possible metrics that may correlate with the accuracy of health-related tweets. The results showed that tweet meta-characteristics have some limited potential in the identification of inaccurate tweets and to inform on their dissemination potential.

The research then built past this work to develop a framework for automatically determining the accuracy of health-related tweets in Arabic. The first step was to develop a model to detect instances of *health-related tweets*. This was accomplished by determining the best pre-processing techniques for use with traditional machine learning and then developing traditional machine learning classifiers. The model was then compared with state-of-the-art pre-trained word embeddings. The findings from evaluating the pre-processing techniques with traditional machine learning showed that pre-processing techniques perform differently from one algorithm to another. In addition, most pre-processing methods highlighted in the literature were not included in the best combination. Pre-processing techniques specific to the Arabic language are more likely to improve classifier performance than other generalized pre-

processing techniques. However, ultimately the deep learning model outperformed the traditional machine learning models, even with optimized pre-processing.

After developing a model to detect the health-related tweets, the *accuracy of the health-related tweets* was to be determined. To develop classifiers for this step, we built data sets labeled "accurate" and "inaccurate." Two medical doctors labeled each tweet, and the data sets were used to evaluate pre-trained language models and word embeddings, to identify the best model for detecting health-related tweets' trustworthiness. The results suggest that pre-trained language models perform better than pre-trained word embeddings.

Results from both phases were impressive individually but suffer from the individual inaccuracy slightly when used in combination to detect accurate health tweets. However, we believe that the proposed process to identify health trustworthiness and the findings from these experiments will open the door for further research in this direction and may eventually result in an even more effective automatic prevention of incidents of health misinformation in Arabic.

# Declaration

I, Yahya Albalawi, declare that this thesis titled, "Trustworthiness of Health-Related Information on Arabic Social Media" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the University of Limerick.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed:
---------------------------------------------------------------------------
Date:

-------------------------------------------------------------------------------------

Yahya Albalawi, 2022

# Acknowledgements

First and foremost, I thank Allah the Almighty for His blessing, without which I could not have finished this research. Many, many thanks be to Allah. I would like to express my deep gratitude to my supervisors, Prof. Jim Buckley and Dr. Nikola Nikolov. I shall be eternally thankful for their exceptional assistance and for sharing their vast knowledge and expertise with me during my doctoral studies. The helpfulness of their recommendations, encouragement, assistance, counsel, discussions, and ideas for solving bottlenecks and other difficulties is incalculable. Without their dedication and time, my research would not have been possible.

I also wish to express my thanks to Lero and the Department of Computer Science and Information Systems at the University of Limerick, which provided continual and prompt support whenever I needed it. My thanks are extended to the administrators, colleagues, and teaching staff for creating an environment of trust, love, and cooperation. A special thank-you goes to my colleagues who were also studying for doctorates.

Moreover, I would like to acknowledge Taibah University in Saudi Arabia for providing funding and for supporting me during my PhD. I am grateful to my internal and annual examiner, Prof. Ita Richardson, and external examiner, Prof. Arsalane Zarghili. I would also like to thank my other annual examiners, Sarah Beecham and Prof. Bashar Nusaybah. Their feedback helped to improve the direction of my research and my thesis. I thank Dr. John Noll for his early advice and feedback in my journey, which were a beacon during a time of uncertainty.

Furthermore, I thank my oldest brother, Mohammed, who unconditionally supported me financially in my final year. I also thank my brother Dr. Yousef Albalawi MD, PhD (Ministry of Health, Saudi Arabia) and Dr. Khalid Shahat MD, PhD (Taibah University) for useful discussions and for helping me in recruiting the experts I needed for my research.

Most of all, I would like to express my profound gratitude to my true and faithful friend in this journey, my wife Nada Albalawi. Without her encouragement and compassion, I would never have finished this work. I thank my beloved kids, Mohammed and Abdullah, for being such a welcome distraction during my studies and for bringing love and joy to my life. Last but not least, I give special thanks to my beloved mother.

# Abbreviations

- **SM –** Social Media
- **CDC –** US Centers for Disease Control and Prevention
- **MERS –** Middle East Respiratory Syndrome
- **LDHs** – Local Health Departments
- **API –** Application Programming Interface
- **MWW** – Mann-Whitney-Wilcoxon
- **ML** – Machine Learning
- **SVM** – Support Vector Machine
- **NB –** Naive Bayes
- **MNB** – Multinominal NB
- **KNN -** K-Nearest Neighbors
- **LG** – Logistic Regression
- **CNN** – Conventional Neural Network
- **BTO –** Binary Term Occurrence
- **LSTM** – Long Short-Term memory
- **BLSTM** – Bidirectional LSTM
- **CBOW -** Continuous Bag of Words Model
- **SG -** Skip-Gram
- **TF-IDF -** Term Frequency — Inverse Document Frequency
- **GloVe** – Global Vectors
- **LIBSVM**: Library for SVMs

- **BERT –** Bidirectional Encoder Representations from Transformers
- **DL –** Deep Learning
- **WHO -** World Health Organisation
- **SLR –** Systematic Literature Review
- **NLP -** Natural Language Processing
- **RF –** Random Forest
- **CoVe –** Contextual Word Vectors
- **ELMo** - Embeddings from Language Model
- **acc** - accuracy

# Contents

# List of Tables

# List of Figures

# Chapter One

*"Misinformation is not like a plumbing problem you fix. It is a social condition, like crime, that you must constantly monitor and adjust to"*

TOM ROSENSTIEL

## 1 Introduction

In the last two decades, there has been a tremendous increase in the adoption of personal communication devices. This has been prompted by the rise in smartphone ownership and by the widespread popularization of social media (SM): By 2020, the number of users on various SM platforms was estimated to have surpassed 3.6 billion [1].

One of the most often discussed subjects among users is health [2-6] and research indicates that Twitter is the most frequently utilized SM platform for health-related communication [7-11]. Unfortunately, not every piece of health information on this medium is accurate. For example, according to some studies, 50% of health professionals' tweets are inaccurate [12].

Given the large volumes of text involved and the critical nature of health information, an approach is necessary to monitor misinformation in this instance and machine-learning models offers the potential to help authorities and researchers detect health misinformation on SM. This thesis focuses on deriving more-optimal machine-learning approaches to monitor health misinformation in Arabic tweets, a hugely popular SM communication channel in Saudi Arabia [13-15].

This chapter provides a contextual introduction to the thesis. It begins by discussing the rationale for this investigation and then proceeds to define the specific purposes and objectives. Additionally, it will describe the strategies used to accomplish the stated goals in overview. Finally, it summarizes the research contributions of this thesis before finishing off by describing the Thesis structure.

## 1.1 Motivation

The motivation for this study can be divided into two factors. The first is related to the importance of detecting the accuracy of health-related tweets, and the second is connected to the language being analyzed, which is Arabic.

The importance of misinformation in SM has attracted many researchers to study this phenomenon [16-19]. There is growing evidence that misinformation generated across SM platforms can have negative impacts, such as misinformation relating to people's health [20-22]. In particular, work during the initial stages of health disease outbreaks has shown that SM misinformation has had significant negative impacts on people’s health [16, 19, 23].

A notable example of this the outbreak of the Middle East Respiratory Syndrome (MERS). It caused 2,578 cases, with a morality rate of 34%, and 84% of the cases were reported in Saudi Arabia [24]. In that instance, misinformation on SM caused severe problems: According to experts’ reviews, misinformation discouraged people from going to the emergency department when they were in an acute condition, while others cancelled surgical appointments for other pressing conditions [25].

As this example illustrates, misinformation can be critical in a health context and, given the 50% inaccuracy figure reported above [12], misinformation would seem to be prevalent in the Arabic world. Yet analysis towards the identification of health misinformation seems to be at a very preliminary stage.

Several studies have pointed out that Twitter is a preferred medium for sharing health information [10, 26-28]. On this medium, people submit a large volume of text, which overwhelms researchers' abilities to study it, and so a machine-learning approach seems like an appropriate strategy to monitor such a situation.

Our literature review for Arabic natural language processing (NLP) on SM suggests that, while there are many studies about Arabic on SM, most of them are concerned with sentiment analysis. This is evidenced by the number of literature surveys and systematic literature reviews on sentiment analysis for Arabic [29-31], in addition to the number of studies concerned with Arabic NLP in more general contexts [32-34].

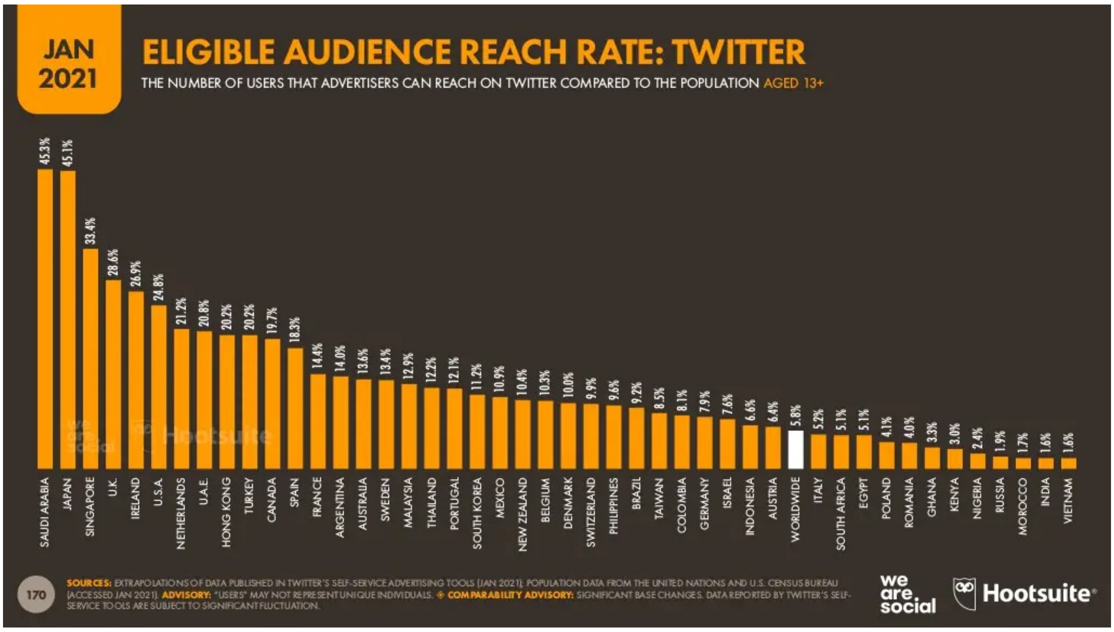


**Figure 1.1: Ratio of Twitter users relative to the population in each country**

Ultimately, given the popularity of Twitter for use in Saudi Arabia as shown in Figure 1.1 especially as a medium for health communication, an analysis of health misinformation on Twitter feeds from Saudi Arabia or in Arabic would be very useful endeavor. To allow the automatic detection of health-information quality offers the potential to filter or quality-assure Tweets at scale, and thus lessen the possibility of poor health outcomes for people who depend on Twitter for their health information.

Instead, five systematic literature reviews concerning SM and health outbreaks, or actionable diseases, have not reported any work concerning Arabic or Saudi tweets. Instead, four studies which were concerned with MERS analyzed the SM data for different languages and other countries [4, 17, 35-37]. Furthermore, Hagg et al. [35] specifically noted the absence of studies that analyze data concerning the use of SM for health-related purposes in the Middle East.

## 1.2 Aim and Objectives

Given the lack of work in this area, this thesis is driven by the overall goal of developing a framework to determine the trustworthiness of health-related tweets in Arabic. To explore and test the feasibility of using tweet-characteristics for this purpose, preliminary research was conducted with the following objectives:

- To study tweet-metrics that could possibly be associated with the accuracy of health-related tweets.
- To assess the methods used in extracting health-related tweets and labeling the accuracy of these tweets.

Although some meta-characteristics were potentially informative in this regard the decision was to design, configure and assess machine-learning (ML) models to detect health-related information in tweets (a precursor to detecting trustworthy health-related information in tweets). This was broken down into the following tasks:

- Evaluating various pre-processing techniques applied to Arabic texts for training a classifier in the identification of health-related tweets.
- Assessing state-of-the-art, pre-trained, word embeddings as an input layer to this identification task, with deep-learning models for this purpose.
- Comparing deep-learning methods with traditional (pre-processing-optimized) ML methods to determine the best model to detect health-related tweets in Arabic.

The subsequent aim was to develop an effective model, to assess the accuracy of health-related tweets. This was approached using the following steps:

- Developing a labeled data set based on health experts' opinions regarding accurate health tweets, that can be used to train a classifier in the accuracy of such tweets.
- Comparing the performance of pre-trained word embeddings with pre-trained language models to assess their accuracy in classifying accurate/inaccurate Arabic health-related tweets.

## 1.3 Research Methodology

The researcher conducted a pilot study to explore if tweet characteristics could inform on the accuracy of health-information in Arabic tweets and their subsequent dissemination. Specifically, the pilot study was constructed to assess the possibility of using (meta) data beyond tweets to augment the identification of trustworthy health tweets. It was also helped to inform on the subsequent design of the other two empirical studies reported on in this research.

While the results from the pilot study were encouraging, this research also wanted to probe ML for this classification purpose towards higher accuracy, given the high performing capabilities of ML in a wide range of classification techniques [38-41]. The second and third columns of Figure 1.2 illustrate the focus on that research: identifying high-performing models to detect health-information tweets and then building a model to detect the accuracy of those tweets. For these two steps, it was essential to justify the selection of the ML model to be applied in each phase, *considering the rapid changes in the field of machine learning* [42]. This research seeks to compare the methods and techniques available as state-of-the-art in the literature, towards building an appropriate classification approach model.

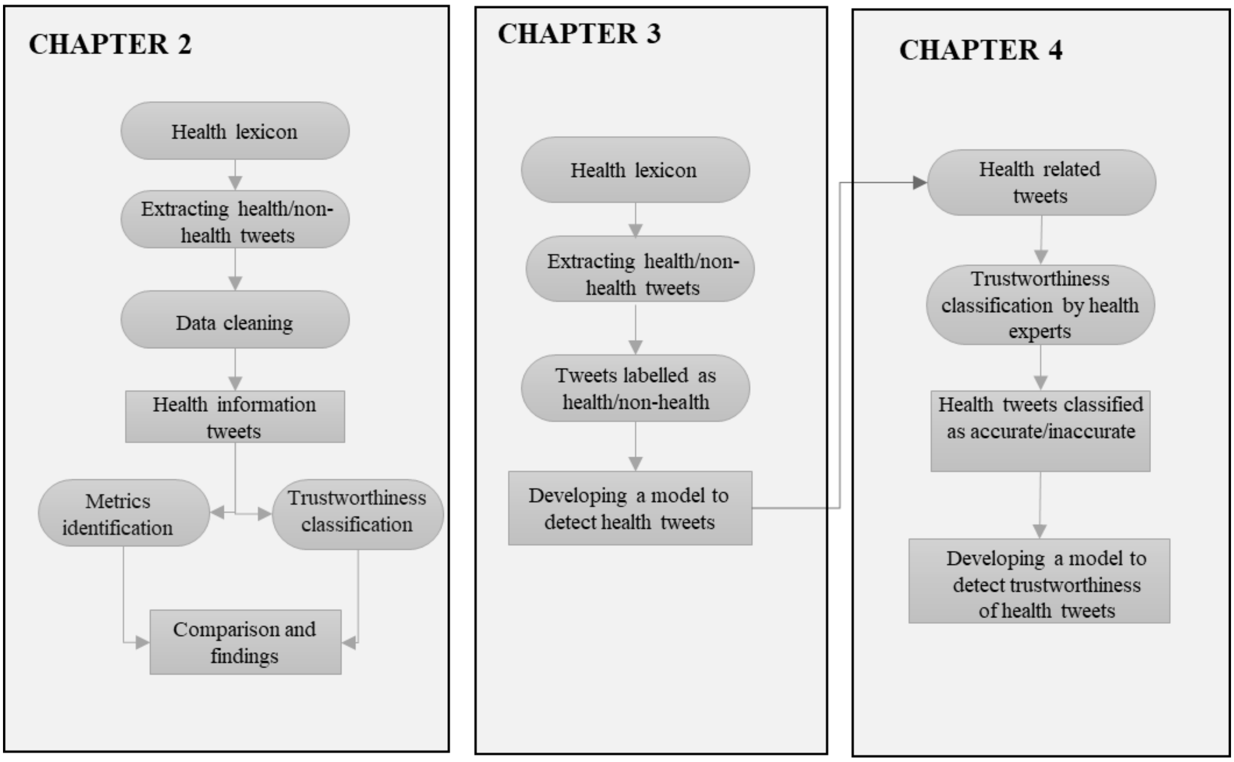


**Figure 1.2: Overview of the research flow**

For detecting health-related tweets (Chapter 4), the research starts as shown in Figure 1.2 by developing a health lexicon to extract tweets. These tweets were independently classified by two annotators as either “health-related” or “not health-related to construct data set to train the modeler. The next step was broken down into two experimental parts. In the first one, a number of pre-processing techniques were evaluated independently for traditional ML techniques. All pre-processing techniques that have been hypothesized to enhance model performance in Arabic text analysis were assessed, and a brute force approach was employed to determine the combination that most improved the model's accuracy in determining Arabic health tweets. In addition, two deep-learning architectures were used to compare several pre-trained, word-embedding techniques with the aim of detecting Arabic health-related tweets in this experiment. Finally, the two, best-performing deep-learning architectures were compared to traditional ML model to identify the one that was better at detecting Arabic health-related tweets.

In the third phase of this research, which aimed to develop a model that predicts the trustworthiness of such tweets, 1,800 tweets were individually labeled as accurate/inaccurate by two medical doctors. The tweets were considered trustworthy if both labeled them "accurate" and they were considered untrustworthy if labeled "inaccurate" by both doctors. We then obtained experimental results against this data set from 16 state-of-the-art, deep-learning models, including pre-trained word embedding and pre-trained language models towards identifying high-performing models.

## 1.4 Research Contributions

The main goal of this study was to develop a framework that can detect and classify health tweets as being accurate or inaccurate. This has resulted in several more granular contributions.

1- Finding factors (tweet meta-characteristics) that are associated with the trustworthiness of health-related tweets: The study performed identifies several factors that correlate with the accuracy of health-related tweets. For examples, retweets and tweets being favorited by crowds and the time of tweets being posted are correlated with accuracy of tweets.

2- Sources of accurate and inaccurate health tweets. This study has also identified sources more associated with accurate and inaccurate health tweets. For example, health organization and health professionals' accounts are more likely to tweet accurate information, followed by sources traditionally considered as trusted users. The least trustworthy category in the data set includes the users with exclusively SM-based accounts.
3- Evaluating pre-processing techniques for detecting health-related tweets and their impact on traditional ML. Not all pre-processing techniques improved the classifier performance for detecting health-related tweets. In fact, only four out of the 26 pre-processing techniques assessed seemed to improve the classifier accuracy significantly:  which are remove duplicate (letters), remove Kashida, replacing ة with ه and replacing ى with ي.
4- Health tweets identification: In terms of comparing the performance of traditional ML approaches with deep learning approaches using pre-trained word embeddings as an input layer for deep learning models, we found that the bidirectional long short-term memory (BLSTM) with Mazajak Continuous Bags of Words (CBOW) as an input layer performed better than the Convolutional Neural Network (CNN) and other traditional ML algorithms.
5- Classifying Health tweet's information as trustworthy/untrustworthy: In terms of accuracy, the research found that most pre-trained language models perform better than BLSTM in detecting the accuracy of Arabic health-related tweets. However, there were some cases where some pre-trained word models outperformed some pre-trained language models.

## 1.5 Thesis Structure

Chapter 2 discusses the research methodology. This includes research methods, data collection methods, and an overview of data analysis strategies. It also outlines how the researcher carried out the literature review. A section on inter-rater reliability is also included to demonstrate the consistency of the outcome of data annotated phases. It also expanded to talk about the annotators' experience. Lastly, this chapter describes the research process as it is undertaken in three phases of the research carried out in this thesis. Chapter 3 presents the paper published in the *Journal of Medical Internet Research* (JMIR) on the pilot study referred to section 1.3. That is, it focuses on the potential of the meta-data and tweet-characteristics in Arabic health tweets to inform on their trustworthiness and the likelihood of their dissemination. It also serves to inform the protocols adopted for the latter studies reported on in this thesis. Chapter 4 changes tack and presents the article published in the *Journal of Big Data*, focusing its attention on identifying Arabic health tweets, using ML through an empirical evaluation of several approaches and pre-processings. Chapter 5 presents a paper accepted for publication at the *JMIR Formative Research journal (JFR)*. It moves past the identification of Arabic health tweets to the classification of these tweets as accurate/inaccurate. These hapters (except Chapter 1 and 2) present all the published/accepted material as-is, except for the italics text, where the changes have been made in response to this thesis's internal and external examiners—also, acknowledging that the word 'paper' has been changed to the word 'chapter' and the figures numbers/tables numbers have been changed to reflect the thesis structure. The papers are contextualized by means of a preface before each of the three chapters.

# Chapter 2

# 2 Research methods

## 2.1 Introduction

The overall goal of this research was to develop a framework to determine the trustworthiness of health-related tweets in Arabic. To achieve this goal, this study was organized into clearly defined steps and used various research methods to collect and analyze the data. In this chapter we will focus on the literature review process, the annotators, health-lexicon creation and the ML analysis techniques used, culminating in a description of the overall research process that employs these components.

## 2.2 Literature review

This thesis adopted two review types—an informal traditional literature review and critical reviews—for each core Chapter (3, 4, and 5), as illustrated in Figure 2.1. A traditional review allows the researcher to overview a gap in the literature on a topic [43, 44], while a critical review allows the researcher to evaluate and critique the relevant literature on this gap [45]. By definition, critical reviews aim to provide "*a critical evaluation and interpretive analysis of existing literature on a particular topic of interest to reveal strengths, weaknesses, contradictions, controversies, inconsistencies, and/or other important issues with respect to theories, hypotheses, research methods or results*"[45].

Other types of literature reviews include systematic and scoping reviews. The author did not use a systematic literature review because it was unlikely to retrieve a sufficient number of studies. This was evidenced by different literature reviews on health information and social media (SM), as these studies did not mention any studies on machine learning (ML) or Arabic health information on SM: We note five systematic literature reviews on health and social media in the current study, including one conducted by the author, and several literatures reviews and surveys on Arabic natural language processing (NLP) and SM. None of these reviews mentioned any studies regarding the detection of Arabic health accuracy information on SM [4, 17, 29-37, 46].

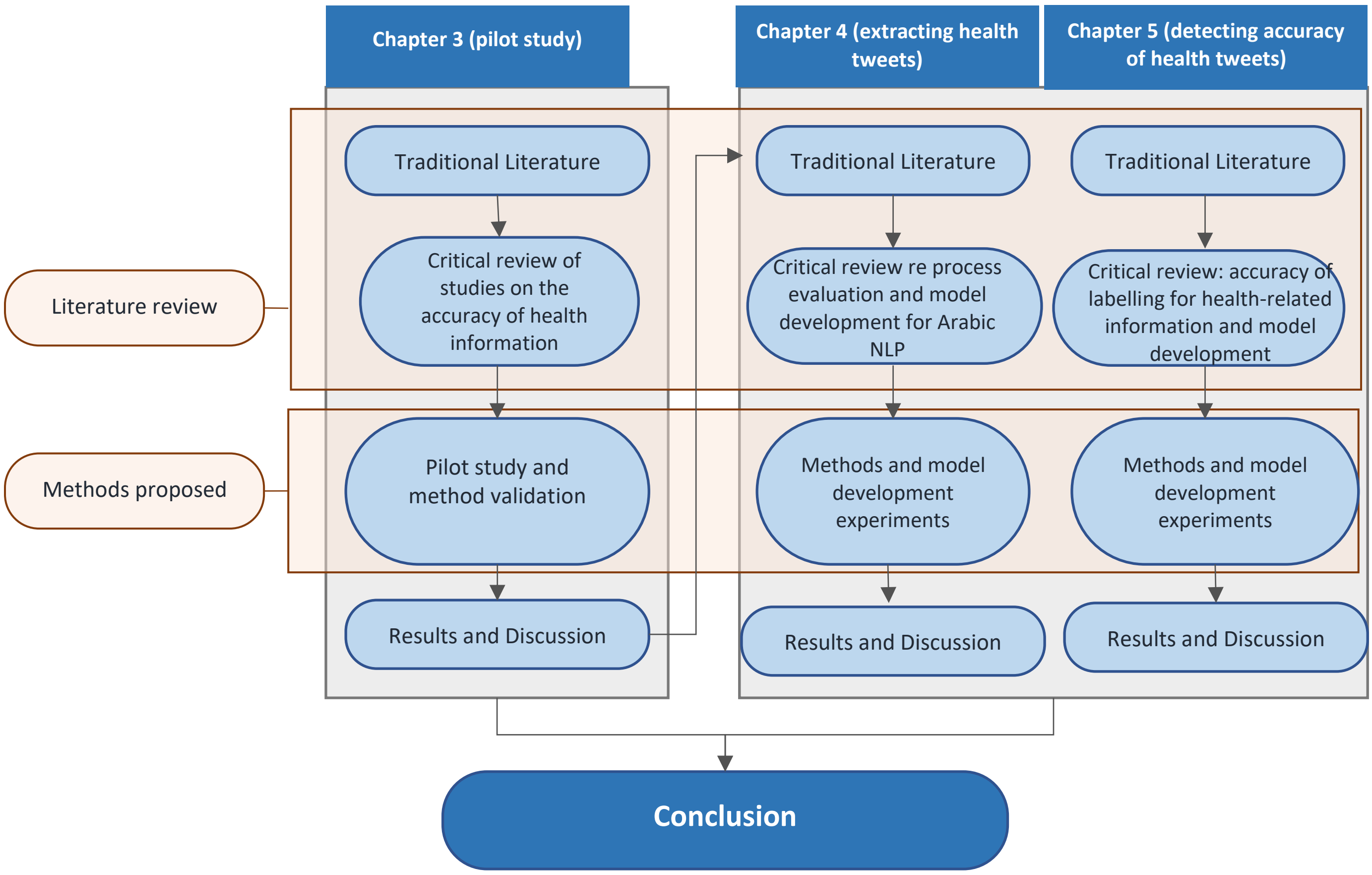


**Figure 2.1 Role of the literature review**

Critical reviews allow researchers to review an area of interest rather than answer specific research questions [45]. Also, critical reviews can occur throughout the project report as the research tackles various problems [47]. Thus, in Chapters 3, 4, and 5, the researcher reviews the related literature in three areas of interest, as explained in Table 2.1. These areas of interest were derived directly from the overall goal for each Chapter.

**Table 2.1 Different types of literature reviews conducted for each chapter in this thesis**

| **Chapters** | Focus of the chapter | Area of interest 1 | Area of interest 2 | Area of interest 3 |
|---|---|---|---|---|
| Chapter 3 | Arabic health information accuracy on SM | Arabic health on SM | Health information accuracy on SM | - |
| Chapter 4 | Extracting Arabic health tweets | Extracting health tweets | Arabic NLP for classification task on SM | - |
| Chapter 5 | Misinformation on SM detection + Arabic NLP | Misinformation on SM detection | Arabic NLP for classification task on SM | Methods to classify health tweets in terms of accuracy |

The structure of the critical review must demonstrate that the authors have read, understood, and evaluated the related literature, which is key to the research problem. Thus, in writing a critical review, the author should usually provide a summary of the existing research, evaluate the research, and link it to the current problem, which this thesis attempts to do in Chapters 3, 4, and 5.

In addition, although this study was not systematically documented, it used an intensive, iterative literature review that kept the research updated during the project, as illustrated in Figure 2.2.

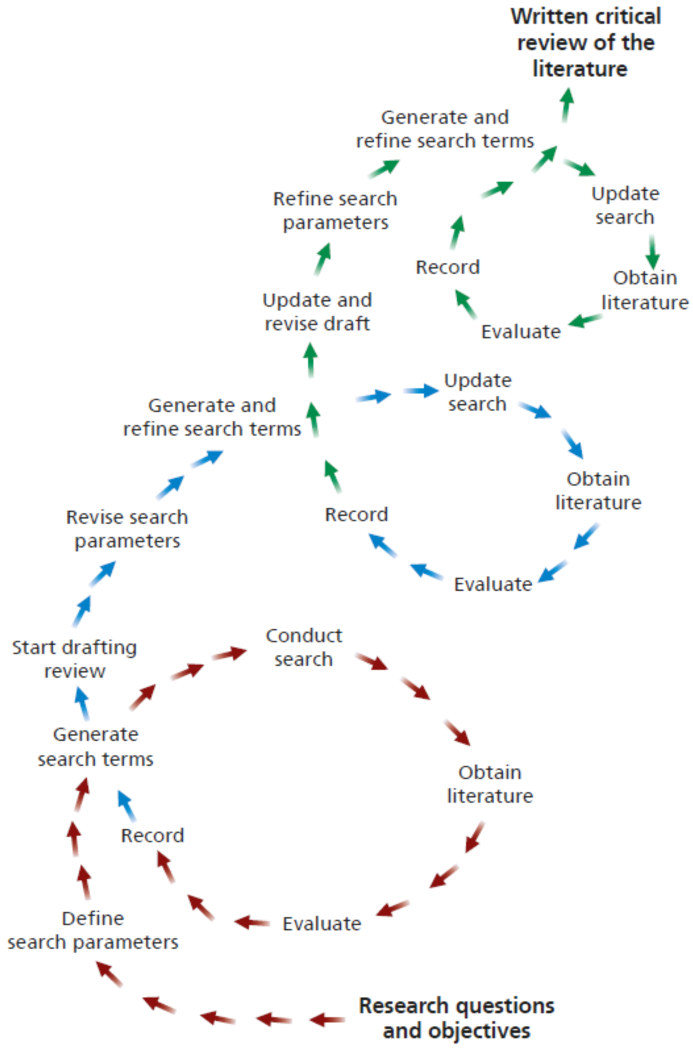


**Figure 2.2 The literature review process (source: Saunders et al. [47])**

According to Saunders et al. [47], conducting a critical review starts by defining the research's overall aim and objectives. If there is insufficient research, the solution is to broaden the research. Table 2.1 provides an example of how a particular area of interest was divided into two or three areas to expand the investigation. This was followed by identifying search parameters, including the subject of the research, which has an influence on the database choices as well as the language for the related articles which, in this case, was the English language[1]. Based on the search parameters and the research objectives, the study identified a number of keywords, followed by a search utilizing the databases mentioned above. After obtaining the results, the author assessed the titles and the abstract in terms of whether the study was relevant based on the study details and overall aims. In cases where the paper was considered marginal, the entire article was read. In addition, during the subsequent evaluation stage, which involves critically reading the articles and taking notes, abstracted articles that were not relevant to the research problem were excluded. All included and excluded articles are kept on an endnote database.

This allowed the researcher to repeat the steps by refining search terms (which will be explained further in the next section) or sometimes widening the research area due to insufficient reviews, as illustrated in Figure 2.2. By the end of each cycle, the researcher drafted a review until the authors were satisfied that they had read a sufficient amount of work relevant to the related area and the research agenda.

### 2.2.1 Search strategy

The researcher searched the following six electronic databases: Scopus, PubMed, Web of Science, Science Direct, IEEE, and the ACM library. As mentioned earlier all extracted papers were kept in an endnote databases during the study, which helped track the reviewed papers regularly. To identify the keywords in this string, the researcher began with published systematic literature reviews about SM and health ML, extracting keywords from these studies and updating the search string.

[1] We are unaware of any published peer-reviewed journal or conferences for Arabic NLP in other languages rather than English. This is evident in the language for leading conferences for Arabic NLP (https://aclanthology.org/venues/wanlp/ ) and also for the computer journals issued by Saudi universities whose primary language is English. Examples include https://www.journals.elsevier.com/journal-of-king-saud-university-computer-and-information-sciences and https://www.springer.com/journal/13369/aims-and-scope.

The researcher used snowballing, which refers to "using the reference list of a paper or the citations to the paper to identify additional papers," both forward (i.e., seeing where the paper was cited) and backward (i.e., seeing what citations the paper used) [48].

**Table 2.2 Examples of search strings used in this thesis**

| Topic | Search string |
|---|---|
| Health communication on social media | Health AND (Twitter OR Facebook OR Instagram OR "social media" OR "new media" OR "social network" OR "social networks" OR "web 2.0" OR "health 2.0") |
| Arabic health communication on social media | (Health OR healthcare) AND (Twitter OR Facebook OR Instagram OR "social media" OR "new media" OR "social network" OR "social networks" OR "web 2.0" OR "health 2.0") AND (Arabic OR Arabian) |
| Extracting health tweets | (Twitter OR tweets OR YouTube OR Facebook OR "social media" OR Instagram) AND ("MACHINE LEARNING" OR NLP OR classification OR classifier OR "data mining" OR detecting or detection) AND health |
| Arabic natural language processing for classification task on social media | (Twitter OR tweets OR YouTube OR Facebook OR "social media" OR Instagram) AND (Arabic OR Arabian) AND ("MACHINE LEARNING" OR NLP OR classification OR classifier OR detecting OR "data mining") |

When new papers were identified via snowballing, the researcher examined why those studies were not extracted using the search string. If the reason was that a keyword had not been included, the researcher updated the search string accordingly. Another reason for the lack of extraction might be that a paper is still in the preprinting phase or has not yet been published. Using this method, the researcher managed to identify a few papers related to pre-trained models that were not published at the time of writing Chapter 4.

## 2.3 Triangulation strategy for the Health Lexicon

To reduce the risk of bias that may have resulted from using one method, this study employed various methods to develop the health lexicon. Specifically, in creating the keywords, the study applied a triangulation strategy by employing three keyword sources and different methods in the process of identifying keywords. These sources included natural communication on Twitter, field experts' suggestions, and an existing health dictionary [50]. Applying the triangulation approach strengthened the study by minimizing possible biases [49]. Notably, while some words suggested by the annotators might have been specific to Saudi dialects, we did not consider the type of dialect for keywords, as this was beyond the study's scope.

## 2.4 Use of Annotators

### 2.4.1 Annotator experience

The annotators were employed twice, as implied in Figure 1.2.

The first task for the annotators was to label tweets as health-related or non-health-related. This task was considered a high-level task that did not require a strong background in health. For this high-level classification, the first annotator was the author of the thesis. The first annotator's experience included working as an information center supervisor for a hospital in Saudi Arabia. He also worked for a health organization on a project to detect public concerns regarding MERS in 2014/2015 in Saudi Arabia from tweets, including identifying tweets and analyzing the response to them. The second annotator was a graduate with a degree in Arabic linguistics. These two annotators played the main role in classifying the tweets in the first study as health-related or not health-related.

### 2.4.2 Medical experts

The medical experts in this thesis were used twice (Chapters 3 and 5) to judge the accuracy of the health-related tweets. To label a tweet as accurate or inaccurate, an agreement should be established between two experts. All medical experts in this study were certified to practice medicine in Saudi Arabia. For Chapter 3 (first paper), all 10 participating medical doctors were consultants. For Chapter 5, due to the high number of tweets being evaluated (1,800), each tweet was evaluated by one consultant and one specialist[2]. 18 consultants and 18 specialists were employed in the studies.

## 2.5 Inter-rater reliability

Given the importance of the correct coding of included studies in datasets, coding should be done independently by two or more persons, followed by the computation of an acceptable inter-rater reliability indicator. The degree to which various raters or judges provide consistent estimations of the same item is known as inter-rater dependability [50].

### 2.5.1 Cohen's Kappa

There are different measures of inter-rater reliability, with Cohen's kappa being the most used formula to calculate the agreement of fully crossed designs with two coders [51].

*The formula used to calculate the Cohen kappa statistics is* $K = \frac{p_0 - p_e}{1 - p_e}$. *Where* $p_0$ *is the relative observed agreement among annotators and* $p_e$ *is the probability chance of agreement.*

According to Cohen, scores of 0 indicate no agreement, 0.01–0.20 indicates none to minor agreement, 0.21–0.40 indicates fair agreement, 0.41–0.60 indicates moderate agreement, 0.61–0.80 indicates significant agreement, and 0.81–1.00 indicates virtually perfect agreement [51]. This view allows for little consensus among raters to be classified as "significant" [52]. Thus, Fleiss et al. proposed another benchmark to interpret Cohen's kappa [53] to evaluate agreement and indicate that such a coefficient is poor (< 0.40 is poor; 0.40–0.75 is intermediate–good; and > 0.75 is excellent).

[2] Specialist doctor means a senior doctor who has spent at least five years working in medical practice after graduating.

## 2.6 Machine learning

ML is “a collection of methods that enable computers to automate data-driven model building and programming through a systematic discovery of statistically significant patterns in the available data” [54]. There are three types of ML algorithms: unsupervised, reinforcement, and supervised.

Unsupervised learning occurs when a computer is given a training task based on unlabeled data with no supervision.

When a system or agent interacts with its environment, takes actions, and learns via trial and error, this is known as reinforcement learning.

Supervised learning requires a present output attribute. Algorithms seek to predict and classify based on these specified attributes. Their accuracies, misclassification, and other performance metrics are based on the counts of predetermined attributes that are correctly predicted or categorized.

### 2.6.1 Algorithm selection and justification

From an ML perspective, the problem of extracting and detecting the accuracy of health-related tweets is a classification problem—a type of supervised learning. This thesis utilized the conventional selection algorithm approach to find the best algorithm. The literature from the critical review justifies the choice of algorithms, followed by experiments. Figure 1.2 illustrates the process from this perspective.

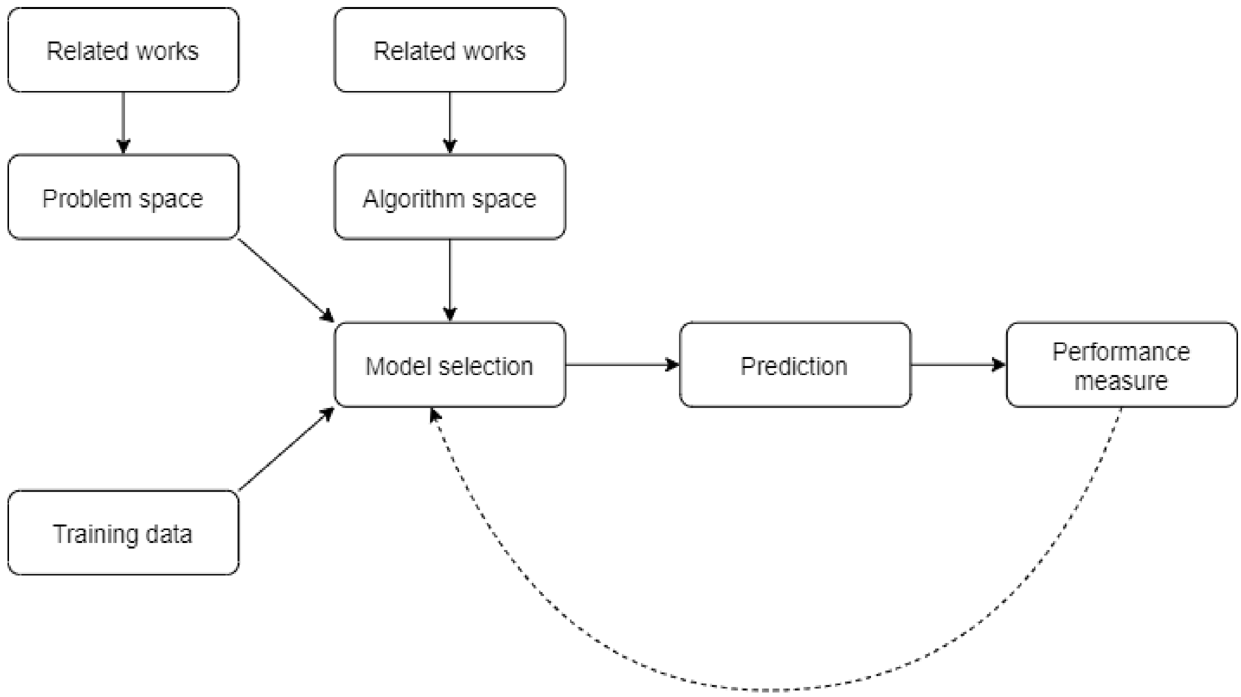


**Figure 2.3 Overview of algorithm selection in this study (adopted from [55])**

### 2.6.2 Model development flow

Developing an ML model to predict predefined labels involves a specific process, which this thesis follows. These steps begin with data gathering and annotation, followed by preprocessing the dataset to prepare it to model and training the algorithm towards evaluating the performance of the trained models utilizing different metrics. These steps are presented in Figure 2.4, which also specifies which processes are evaluated in this thesis and in which chapter.

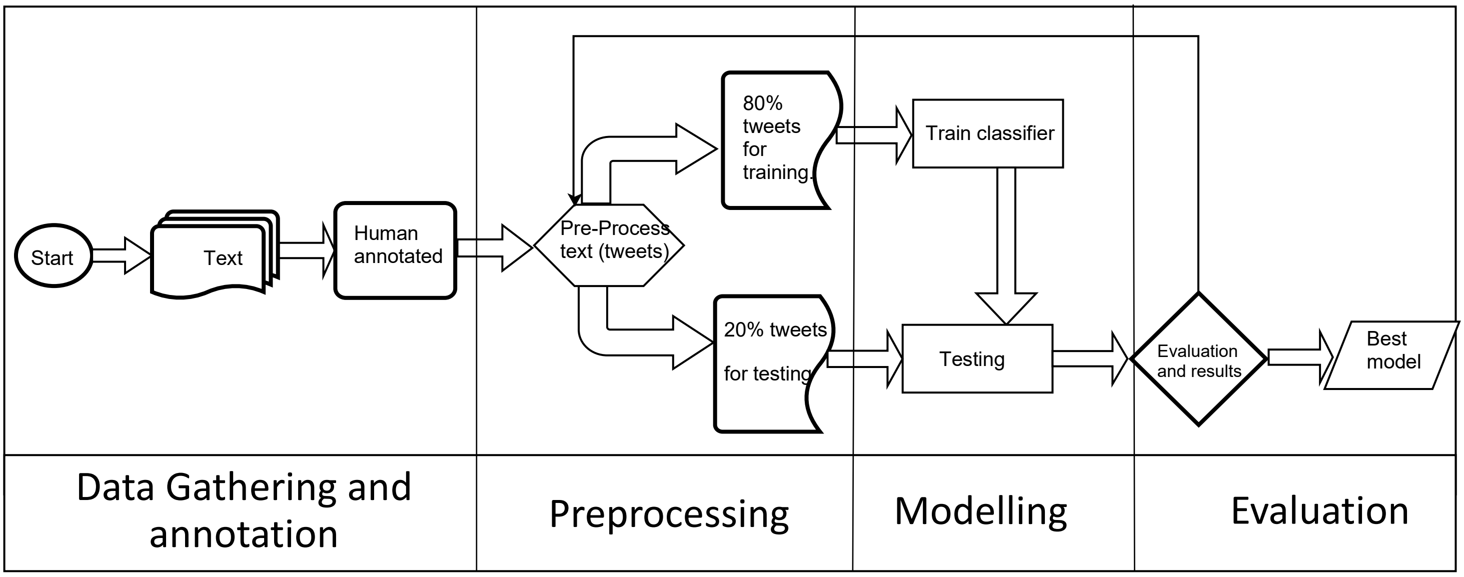


**Figure 2.4 Machine learning model workflow**

#### 2.6.2.1 Data gathering and annotation

The first stage of supervised learning is usually to gather data (i.e., build a corpus of health-related tweets). For Chapter 4, which aimed to develop a model to extract health tweets, this was done by extracting tweets using a health lexicon. These tweets were then annotated as health-related and non-health-related.

Chapter 5 aimed to automatically detect the accuracy of health-related tweets. This was done by labeling health tweets as accurate or inaccurate based on experts' opinions.

#### 2.6.2.2 Pre-processing and splitting the text

After gathering and labelling the data, pre-processing, which is when modification is applied to the data before supplying it to the algorithm, was undertaken [56]. As pre-processing are pre-specified in pre-trained models, they were not evaluated for that model type. However, the literature indicates a contradiction and gaps in assessing the pre-processing of traditional ML algorithms. Thus, the pre-processing techniques were evaluated to justify their selection/omission in Chapter 4. The way the text is pre-processed might affect the algorithm's performance [57], making it crucial to perform evaluation of this step carefully.

In this stage, the pre-processed text was split as follows: 80% to train the models and 20% to test the models.

Notably, in Chapter 4, the processing techniques were subject to evaluation, as indicated in RQ 2.1.

#### 2.6.2.3 Modeling and Experiment

A range of models were trained and evaluated on the obtained dataset using experiments to demonstrate how well they recognized health-related tweets (Chapter 4) and also in detecting the accuracy of health-related tweets (Chapter 5).

#### 2.6.2.4 Evaluation

To evaluate the models, we used the following metrics: precision, recall, F1 score and accuracy. These are the most-used metrics to evaluate ML model performance [58, 59]. Table 2.3 explains the formula used to calculate these metrics, which were used to assess the efficacy of the suggested model. In addition, comparisons with other baselines and past studies were made. To compare these models, we used the F1 score as a decisive metric. This score is the recommended metric when a dataset is imbalanced [60].

**Table 2.3 Metrics used to evaluate the ML model**

| Metric | Formula |
|---|---|
| Recall | True Positives/(True Positives + False Negatives) |
| Precision | True Positives/(True Positives + False Positives) |
| $F_1$ score | (2 * Precision * Recall)/(Precision + Recall) |
| Accuracy (ACC) | (True Positives + True Negatives)/Total Samples |
| Macro averaged $F_1$ | $\frac{1}{N} \times \sum_{i=1}^{n} F_1(i)$, where $N$ is the number of classes. |

## 2.7 Research process and phases

***Research Phase one (Chapter 3):***

The initial goal of this research was to develop an ML model that can predict the accuracy of health-related tweets. A feasibility study was conducted to evaluate the methods used in extracting and evaluating the health-related tweets. It also aimed to analyze the prevalence of accuracy of health-related tweets regardless of their sources. We use quantitative methods (a descriptive statistic and statistical tests) to test the significance of the association between dependent variables.

To achieve this goal, an initial health lexicon was developed, and tweets were extracted during a period of 2 weeks in 2017. Three hundred health-related tweets were evaluated by two doctors as accurate or inaccurate, in which the doctors agreed on 109 tweets. The researchers then analyzed the accuracy of the health-related tweets, and the factors that correlate with their accuracy. More than 100 tweet metrics were analyzed to find any relation between the accuracy and the linguistics, the time and meta factors.

***Research Phase two (Chapter 4):***

After testing the methods in extracting and evaluating the health-related tweets, there was a need to automate these methods towards scaling. The first step for automation was to develop an ML model to extract health tweets. However, as previously explained, in machine learning, there was no best method, and there was a need to use multiple methods.

This chapter aims to develop a model to detect health-related tweets; however, this was not possible without comparing different approaches in machine learning. To meet the aim of this chapter, the literature review identified two different gaps for the practice of developing the Arabic ML model. First, there are no agreements on the steps required to pre-process the data. Second, researchers use different ML models in traditional machine learning. At the time of the study, researchers had only recently started to use deep-learning models for Arabic NLP tasks. Thus, the research identified 26 pre-processing techniques, which the researchers evaluated using four different traditional ML models. The best-of-the-breed models were compared with the best deep-learning ML models.

***Research phase three (Chapter 5):***

The last stage of this research was to develop an ML model to filter the health tweets as accurate and inaccurate. To achieve this goal, the research compares pre-trained language models with the best deep-learning archives used in Chapter 5.

# Preface to Chapter 3

*This chapter was published in the Journal of Medical Internet Research (JMIR – 2019), which has an impact factor of 5.05 and a SNIP of 2.070. It is classified as a Q1 journal in health informatics. Approximately 60% of the work in this paper was done by the author of this thesis, and it currently has 25 citations.*

The first paper reports on a study to assess meta-data and tweet characteristics in terms of their ability to predict the trustworthiness of health tweets in Arabic. Previous studies in this area have judged the trustworthiness of the information in tweets based on the sources of information [61-63]. However, it is uncertain whether assessing the trustworthiness of tweets solely based on source is accurate: It has been reported that even typically trusted health information users are not always reliable in their tweets [12]. Other research has evaluated tweets based on the hashtags and links in the tweets, but the scope of this information is still quite limited. Establishing a feasibility study to examine a wider range of elements that correspond to accurate and inaccurate health-related tweets is critical for understanding the prevalence of inaccurate tweets, regardless of source.

Thus, in this chapter we expand on the metrics used for assessing the trustworthiness of a health-related tweet. More than 100 metrics can be recorded for each tweet [64] and their relation to the trustworthiness of health-related tweets can be investigated. Here we assess these metrics in terms of their ability to inform on trustworthiness and dissemination of tweets.

This pilot study was also essential to hone the design of a more extensive investigation of ML models to detect health and trustworthy health tweets. It helped in the construction of a trustworthy model by first validating the methods used in classifying the tweets as health-related or not health-related. Secondly, it helped in determining the trustworthiness of health-related tweets, again towards providing a valid data set for training and evaluation.

# 3 Trustworthy Health-Related Tweets on Social Media in Saudi Arabia: Tweet Metadata Analysis


## Abstract

### Background

Social media platforms play a vital role in the dissemination of health information. However, evidence suggests that a high proportion of Twitter posts (ie, tweets) are not necessarily accurate, and many studies suggest that tweets do not need to be accurate, or at least evidence based, to receive traction. This is a dangerous combination in the sphere of health information.

### Objective

The first objective of this study is to examine health-related tweets originating from Saudi Arabia in terms of their accuracy. The second objective is to find factors that relate to the accuracy and dissemination of these tweets, thereby enabling the identification of ways to enhance the dissemination of accurate tweets. The initial findings from this study and methodological improvements will then be employed in a larger-scale study that will address these issues in more detail.

### Methods

A health lexicon was used to extract health-related tweets using the Twitter application programming interface and the results were further filtered manually. A total of 300 tweets were each labeled by two medical doctors; the doctors agreed that 109 tweets were either accurate or inaccurate. Other measures were taken from these tweets' metadata to see if there was any relationship between the measures and either the accuracy or the

dissemination of the tweets. The entire range of this metadata was analyzed using Python, version 3.6.5 (Python Software Foundation), to answer the research questions posed.

**Results**

A total of 34 out of 109 tweets (31.2%) in the data set used in this study were classified as untrustworthy health information. These came mainly from users with a non-health care background and social media accounts that had no corresponding physical (ie, organization) manifestation. Unsurprisingly, we found that traditionally trusted health sources were more likely to tweet accurate health information than other users. Likewise, these provisional results suggest that tweets posted in the morning are more trustworthy than tweets posted at night, possibly corresponding to official and casual posts, respectively. Our results also suggest that the crowd was quite good at identifying trustworthy information sources, as evidenced by the number of times a tweet's author was tagged as favorited by the community.

**Conclusions**

The results indicate some initially surprising factors that might correlate with the accuracy of tweets and their dissemination. For example, the time a tweet was posted correlated with its accuracy, which may reflect a difference between professional (ie, morning) and hobbyist (ie, evening) tweets. More surprisingly, tweets containing a kashida—a decorative element in Arabic writing used to justify the text within lines—were more likely to be disseminated through retweets. These findings will be further assessed using data analysis techniques on a much larger data set in future work.

**Keywords:** social media, new media, misinformation, trustworthiness, dissemination, health communication

## 3.1 Introduction

### 3.1.1 Background:

In recent years, there has been significant growth in the uptake of personal communication technologies around the world. This has been largely afforded by the widespread availability of social media (SM) and has been facilitated by the increase in mobile phone ownership. SM has become a valuable tool for communication and it has been utilized in many areas, such as education [65] marketing [66] and health communication [67]. For example, in the field of health communication, the US Centers for Disease Control and Prevention (CDC) [68] and local health departments in the United States [69] have used Twitter to communicate to people during epidemics. Another example is from the United Kingdom and Norway, where health authorities used Twitter to inform their citizens during the West African Ebola outbreak in 2014 and 2015 [70].

The use of SM can improve the nature of health communication as it speeds up the interaction between health care organizations, professionals, and patients [67]. Thus, various SM platforms and apps can play a vital role in health communication and in the promotion of good health [71]. Despite the advantages that SM potentially offers for health communication, it also faces certain challenges. For example, during a health crisis, there is only a limited amount of time for authorities to respond in an efficient way and inform people, while simultaneously helping to eliminate uncertainty on a topic. If this does not occur promptly, it is much more likely that rumors will spread, possibly through SM; when this happens, the negative effects of SM, such as confusion and misinformation, are the probable results [72].

Illustrative examples include the negative consequences experienced by Saudi Arabia and African countries during the Ebola and Middle East respiratory syndrome (MERS) outbreaks. In Saudi Arabia, SM rumors prevented some people from going to emergency departments when they were in an acute condition, resulting in the cancellation of their surgical procedures [25]. In Africa, rumors were found on SM that drinking a huge amount of salt water was a cure for Ebola; it has been reported that this may have caused the deaths of several people [73]. Notably, these misinformation issues seem to affect developing countries more deeply; studies have suggested that 4.5% of Twitter posts (ie, tweets) in the United States are misinformed [74], however, Oyeyemi et al. [73] have found evidence *showing that approximately 58% of retweeted tweets in West Africa are misinformed.* Table 3.1 *summarizes the aforementioned studies.*

Three different studies conducted on health and different types of users reported that Twitter was the preferred platform among health professionals [7] medical students [10] and diabetic patients[27]. In addition, most health-related studies on SM, focus on the English-speaking population and the United States [75, 76] and not on other cultures. Specifically, [35], reported the absence of literature analyzing SM data for health-related purposes in the Middle East. This is particularly surprising because recent statistics from Statista indicate Saudi Arabia has the fourth highest number of Twitter users in the world [13]. Furthermore, when assessing the ratio between Twitter users and populations for each country [77], Saudi Arabia has the highest number of users on Twitter relative to its population. These findings are also supported by other researchers who reported on the elevated prevalence of Twitter usage in Saudi Arabia [14, 78]. That being said, only one study has analyzed Saudi health tweets [12].

Given the likely cultural differences between Twitter use in the West and in the Middle East, it seemed important to assess health-related tweets in a Middle Eastern country. Given the prevalence of Twitter use in Saudi Arabia, that seemed like an appropriate country to choose. Hence, this study focuses exclusively on Saudi Arabia.

While a number of tweet characteristics have been assessed for accuracy of the information on SM, the foremost characteristic of interest in this regard has been the source of the tweet. Intuitively, one would anticipate that tweets from health professionals would be more trustworthy; however, this is an open question. A study by Alnemer et al [12] found that 50% of Saudi health professionals' tweets were not evidence-based. In addition, this study only includes tweets that are posted by accounts with more than 45,000 followers. The relatively high number of followers suggests that these account holders might be considered as opinion leaders in their domain of expertise which, in this case, is health.

These findings question the accuracy of health professionals' tweets, which is a worrying result considering that people are traditionally more likely to trust users who are physicians, health organizations and pharmacists [79-83]. While these sources are trusted, evidence shows that they are not necessarily trustworthy (ie, accurate): Alnemer et al. [12] suggest that, even if sources are traditionally trusted, there is a high possibility that they include inaccurate (ie, untrustworthy) information.

A few methods and tools for detecting misinformation on SM, particularly during a health crisis, have been proposed, usually focused on specific topics and diseases (e.g, Ebola and Zika) [16, 61]. They typically strive at identifying the characteristics of misinformation, while neglecting the factors that indicate trustworthy tweets.

**Table 3.1 Examples of studies that analyze the accuracy of health information on SM**

| | *Number of tweets/ docs* | *Sources* | *Methods to identify misinformation* | *language covered* | *Percentage of the accuracy* | *Topics covered* |
|---|---|---|---|---|---|---|
| *[12]* | *625* | *Twitter* | *Vote if the experts did not agree.* | *Arabic* | *50% inaccurate* | *Only tweets from health professionals* |
| *[73]* | *Not applicable* | *Twitter* | *Not mentioned* | *English* | *58 % of retweeted tweets are inaccurate* | *Ebola* |
| *[74]* | *5,395* | *Twitter* | *Manually check agreement on 125 tweets; whether they were unsubstantiated by the following reference standards: the CDC and Public Health Agency of Canada for scientific claims and a panel of credible online news sources (e.g., CNN, BBC) for news-related claims.* | *English* | *4.5% inaccurate* | *H1N1* |
| *[61]* | *47M* | *Twitter* | *Type of hashtags.* | *English* | *25% of the analyzed tweets were speculative* | *Ebola* |
| *[84]* | *26,728* | *Twitter* | *Define keywords to extracted tweets, based on rumors identified from WHO website. Then Coders labelled tweets* | *English* | *32%* | *Zika* |

In this study, tweets are considered to be trustworthy if they are accurate, and are considered to be untrustworthy if they are inaccurate. This position is similar to Yin et al., [85], who state that a website is trustworthy if it provides correct information, and information is likely to be true if it is provided by a trustworthy website. Likewise Zhao et al [86], who developed a topic model to estimate the trustworthiness of the news on Twitter defined a trustworthy tweet as one that refers to things that really happened. Similarly, this paper considers a tweet as trustworthy if it contains accurate health information and if the process of evaluating the accuracy of tweets is introduced in the methods section.

The perspective we take in our work is to focus on determining the factors that correlate with the trustworthiness of health information tweets as well as the factors that affect the dissemination of those tweets. This work has been undertaken in order to determine how

SM might be effectively oriented toward the dissemination of trustworthy health information.

We assess the trustworthiness of tweets originating from traditionally trusted health sources and examine the relationships between (>100) of a tweet, and its trustworthiness.

Tweets do not need to be accurate or evidence-based to receive traction. In their study, Nastasi et al [87] noted that scientifically inaccurate health tweets were retweeted in the same manner as accurate tweets. That work also indicated the need to study the dissemination metrics of tweets in order to find factors that correlate with high dissemination. Consequently, as a second objective we will assess the factors that correlate with larger dissemination of health tweets.

### 3.1.2 Prior Work

SM data from Facebook, Instagram, YouTube, and Twitter have been used to understand people's attitudes and behaviors in sharing and consuming information related to specific health issues, such as vaccinations, abortions, posttraumatic stress, and cancer [88-93]. Facebook has been used by both private health stakeholders and government agencies to engage with the public [94, 95] ]; studies have provided understanding by analyzing Facebook's timelines [95] and health agencies' accounts on Facebook [94].

Although Facebook is the most popular overall platform, the most popular SM platform to study health is Twitter [6, 96, 97], as evidenced by studies included in different systematic reviews of topics related to health and social media [4, 5, 23, 98]. This focus may be due to the complexity of Facebook data and its unavailability, due to privacy restrictions [96].

Most health studies on Twitter collect their data by using specific keywords [16, 61, 99-103], or from tweets authored by specific health stakeholders such as health organizations [3, 12, 69, 84, 104-107]. However, it appears that most studies that analyze Twitter for health using specific keywords have not analyzed the types of users who post the tweets [16, 61, 99-103]: It is known from other studies that there are different types of users, and they share different types of information or hold different attitudes toward specific health issues [108-111]. For example, a number of studies performed tweet extraction using keywords to identify public concerns during the Zika outbreak [99-103]. Besides being limited to a particular outbreak, these studies did not analyze the interplay between public concerns and the types of users and did not address the factors that make a tweet trustworthy.

A notable study of health-related tweets in Arabic is the one performed by Alnemer et al [12], which manually analyzed the accuracy of tweets authored by preselected health accounts. Their results suggested that governmental institutions are more likely to tweet accurate information than are health professionals or other institutions (ie, physicians, dieticians, and nongovernmental and unofficial health institutions). They reported that 80% of the observed governmental institutes' tweets consisted of accurate health information, followed by physicians (60%). However, the overall accuracy of the tweets, over all observed accounts, was 50%. These findings suggest that even if SM users have health expertise, it cannot be taken for granted that their tweets provide accurate health information. This line of investigation can easily be extended to nonhealth users. Alnemer et al [12] did not examine the characteristics of a tweet that may correlate with its trustworthiness.

In terms of trustworthiness, a number of classifiers for health-related tweets in English have been proposed. For example, Ghenai and Mejova [84] proposed a classifier to detect health rumors on Twitter limited to the Zika virus. A limitation of this study is the fact that their classifier was trained on a limited number of rumors, identified as such by information on external non-SM websites. In addition, annotators who labeled the tweets as misinformed were not health experts. In another study, Ghenai and Mejova [16] focused on the detection of users tweeting or propagating misinformation about cancer, excluding social bots and organizational accounts.

However, social bots have also been considered as a possible source for health misinformation on SM [112, 113]. For example, Allem et al analyzed tweets in regard to e-cigarette discussions and found that social bots may support misinformation on SM in regard to e-cigarette cessation. In their study, they emphasized the importance of distinguishing between social bots and real users. Similar to Allem et al [113] Broniatowski [112] analyzed tweets specifically to understand how bots promote online health content in regard to vaccine-related messages and found that social bots were one of the possible sources for antivaccine advice on SM.

As suggested above, in terms of social bots and misinformation on SM, previous research has emphasized the importance of distinguishing social bots from real users, particularly when the intent is to assess views held by users who are not bots [113, 114]. However, this is not an easy task, as some social bots might mimic user behavior [115]. Social bots might introduce themselves as individual accounts with locations and photos for their profiles [113, 115]. Furthermore, some organizations use social bots to disseminate information, which makes it hard to distinguish between it being the *opinions of the bot* or that of their organization and classifying according to their organization might be difficult [116, 117]. In the work presented here, users were classified as they introduced themselves and it was that classification that was analyzed in terms of the accuracy of information they portrayed.

Kalyanam et al [61] analyzed the association between hashtags and the credibility of tweets related to the Ebola outbreak. They defined credible tweets as those with hashtags that indicated origin from well-known governmental agencies or other authoritative sources (eg,

#cdc or #cnn) and speculative tweets as those with hashtags that indicated the spread of fear, rumor, scam, or humor. It was determined that almost 25% of the analyzed tweets were speculative. Their findings suggested that verified users were more likely to interact with credible hashtags; on average, the number of followers for accounts that posted tweets with credible hashtags was 7000, compared to 2700 for accounts that posted tweets with speculative hashtags in their data set. Kalyanam et al relied-on hashtags, without evaluation of the information carried by the tweets. That is, it is unclear whether tweets classified as credible really contain accurate or trustworthy information.

In terms of identifying influential users on SM, Albalawi and Smith [118] applied six different tools to identify the most influential Twitter users in Saudi Arabia. First, they used the apps Tweepar and SocialBaker, which reveal the most influential users per country on Twitter. With the influential users being identified, they collated four Twitter influence scores via the following: Social Authority by Moz [119], PeerIndex [120], Kred [121], and Klout [122]. However, within the scope of their study, they did not consider health in isolation, and they did not analyze the accuracy of the tweets.

Wong et al [69] analyzed tweets sent by 287 local health departments (LHDs) during the Ebola epidemic in the United States. They found that 70% of the LHDs tweeted at least once about Ebola and that Twitter had become a frequent tool used by LHDs during this particular epidemic. Regarding the dissemination of tweets, one of their findings was that the presence of hashtags and links was highly correlated with the messages being retweeted. Similarly, Suh et al [123] also reported that tweets containing hashtags and links were more likely to be retweeted. They did not consider the impact of the type of users on retweeting. Furthermore, the analyzed tweets were on randomly selected topics, unrelated to the health care domain.

The research results summarized in this section suggest that there is a lack of comprehensive studies on the accuracy of health-related tweets on SM. Specifically, to the best of our knowledge, there are no results that determine the factors that make a health-related tweet trustworthy in general, besides it being authored by a credible institution. In addition, there are no studies on the general factors that affect the dissemination of trustworthy health-related tweets, besides the credibility of the author and the presence of hashtags and links. We believe that the identification of such factors may help health organizations in better disseminating trustworthy information during an outbreak-related health crisis. In particular, there is a lack of studies on health-related tweets in Arabic; this makes such a study a priority, considering the high popularity of Twitter in the Arab world.

Hence, the work presented in this paper addresses the following questions:

1. How can the trustworthiness of health care stakeholders' tweets be identified from the tweets' features?
   1.1. What proportion of trustworthy health-related tweets come from the following sources: health professionals, health organizations, and authorities?
   1.2. What are the other characteristics associated with trustworthy health-related tweets?
2. What are the factors that contribute to the wider dissemination of health care-related tweets that could possibly be used to make accurate health information dominant over other related information on SM?
   2.1. Does the trustworthy nature of health-related tweets increase their dissemination?
   2.2. What other factors contribute to the dissemination of health care-related tweets?

The structure of this paper is as follows. First, the paper introduces the empirical design for identifying the factors affecting the trustworthiness and dissemination of health-related tweets. Second, it presents the results of our work and highlights, in particular, the findings that correspond to the research questions listed above. Finally, this paper ends by discussing the future directions of our research and the possible implications of the results.

## 3.2 Method

### 3.2.1 Overview

This work utilized a standard text analytics methodology, which incorporated the following steps:

1. We developed a health lexicon using two different methods.
2. By means of this health lexicon, we extracted health tweets using the Twitter application programming interface (API).
3. From the remaining tweets, we manually refined tweets related to health using two annotators.
4. Medical professionals manually labeled the remaining tweets as either accurate or inaccurate.
5. We extracted features from the labeled data set. These included attributes of the tweets as well as attributes of the user profiles for the users who authored the tweets. In this paper, only aggregated user data is presented for ethical reasons.
6. We analyzed the labeled data set to provide preliminary answers to the research questions outlined in the previous section.

The outlined methodology is presented Figure 3.1. The sections that follow explain each step in detail.

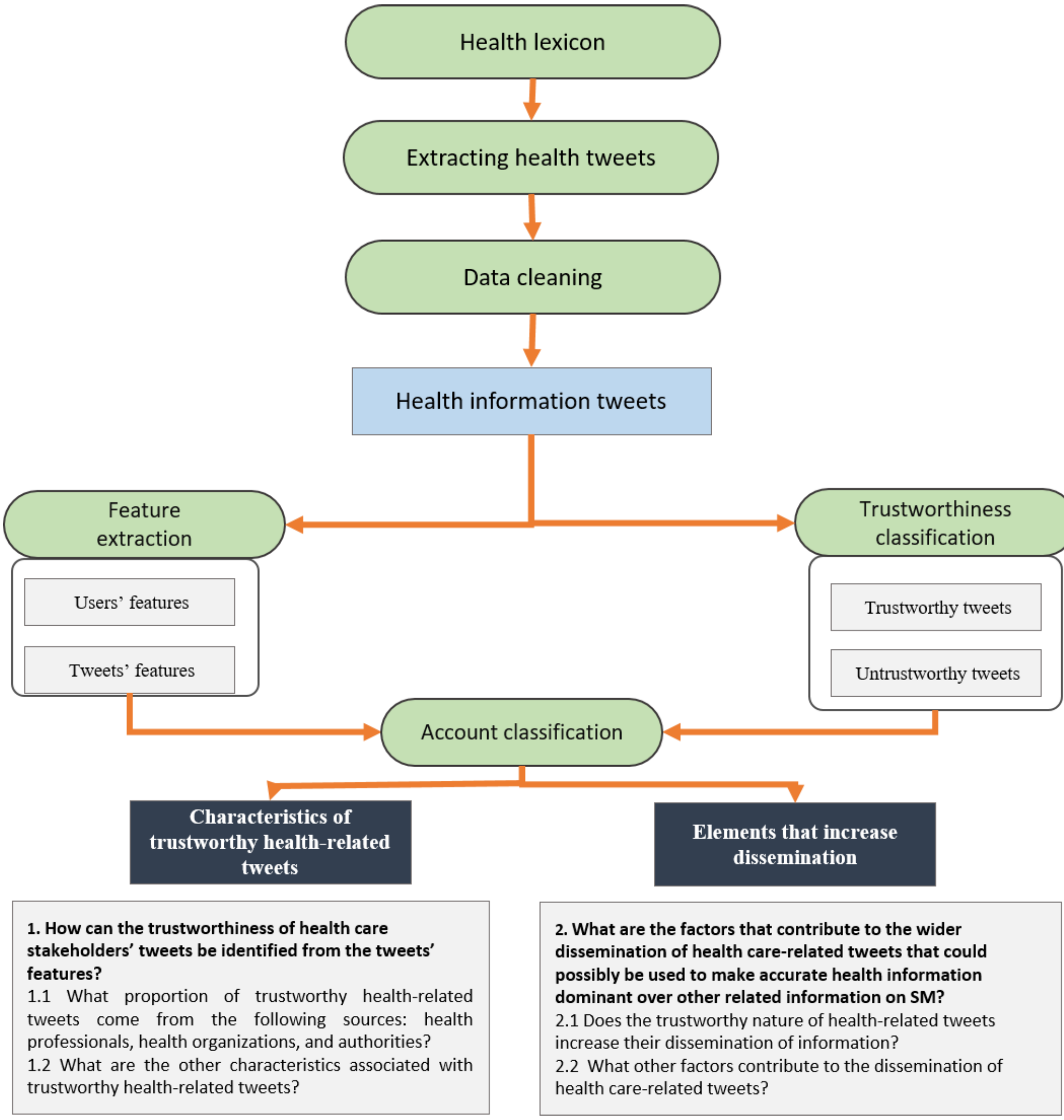


**Figure 3.1. Overview of the research methodology. SM: social media.**

### 3.2.2 Construction of Health Lexicon

In order to identify health-related tweets, a health lexicon was created. It is important to note that the incorrect selection of indicative health keywords could bias the results [124]. Therefore, two separate methods of analysis were utilized to generate this lexicon. The first method consisted of asking three medical doctors with active Twitter accounts to provide 100 health-related words. The doctors we asked are skilled in different disciplines and differ from each other in their age, background, and gender. They were asked to provide an initial list of health-related words that they think would be used in tweets related to health.

The second method involved the usage of 110 health care keywords that were independently identified by an annotator with a college degree in linguistics. As this study concentrates on Saudi Arabia, the annotator identified these keywords by examining a set of tweets with geolocation that indicated a Saudi Arabian origin, although people who enable geolocation are likely to represent a specific demographic group [125, 126]. As such, the words chosen by the annotator might have limited generalizability with respect to the wider demographic group. Thus, the annotator also reviewed health-related accounts and hashtags to identify different health-related words. A complete list of keywords is attached in Appendix A.1.

These two methods were combined in an attempt to construct a health lexicon that is as unbiased as possible.

### 3.2.3 Data Cleaning

Using the lexicon developed as described in the previous section, it was then possible to extract Twitter data for the main part of our work. The Twitter API does not allow users to extract tweets more than a week old [127]. To reduce the impact of this limitation, it was decided that we would extract two data sets. The first data set was extracted on May 18, 2018, and the second data set on August 7, 2018. Table 3.2 describes the characteristics of each data set.

**Table 3.2. Data sets' Characteristics**

| | ***First Data Set*** | ***Second Data Set*** |
|---|---|---|
| ***Total Tweets*** | 209,345 | 196,670 |
| ***Original Tweets*** | 57,793 | 39,454 |
| ***Replying-to Tweets*** | 28,329 | 32,470 |
| ***Re-tweeted Tweets*** | 123,222 | 124,746 |

Reply tweets were difficult to evaluate by annotators, due to a lack of contextual information. As a result, these tweets were removed. It was also necessary to remove all retweeted tweets due to redundancy.

The tweet extraction process resulted in an accumulation of 97,248 tweets: 57,794 from the first data set and 39,454 from the second. Using the random method in Python, version 3.6.5 (Python Software Foundation), a sample of 2800 tweets was selected from the set for use in a prototype study. Even though the lexicon suggested that all of the 97,248 tweets were health related, a manual examination of the tweets showed that this was not necessarily so. Thus, two annotators were employed to filter out tweets unrelated to health from the sample of 2800 tweets. The guidelines for the annotators were as follows:

1. Tweets that describe any function of the body, such as enzymes, organs, or diseases, should be retained.

2. Tweets that give advice or information about supplements, drugs, physical activity, or food and link it to people's health, such as how vitamins, food, and drugs affect people's health, should be retained.

Based on an internal discussion among the three authors of this study, these guidelines were derived from Bobicey and Sokolova's ontology for personal health information [128]. The terms were derived from concepts in their ontology and were considered by all three authors as the most indicative of health-related material.

Each annotator labeled 60% of the tweets, with 10% of the tweets (n=280) labeled by both annotators in order to check the reliability of the analysis. The Cohen kappa statistic for interrater reliability [52] was then calculated, resulting in a value of .872, which indicates excellent agreement between the annotators.

Out of the 2800 tweets, only 552 tweets (19.71%) were labeled as health related. Out of these 552 tweets, 180 tweets (32.6%) originating from the first data set were selected, in addition to 120 tweets (21.7%) originating from the second data set. By doing so, it was possible for us to retain the proportion of tweets in the originally collected data sets.

### 3.2.4 Trustworthiness Classification

Once the previous processes had been completed, 10 medical doctors were asked to manually classify the tweets into the following categories:

1. Accurate health information.
2. Inaccurate health information.
3. Not sure about the accuracy.

The *not sure* option was given to the doctors to avoid forcing them to make a decision on tweets if they did not have enough relevant health knowledge to accurately evaluate them or if the tweets were ambiguous.

A total of 10 Google forms were created, each containing 30 tweets. A link to each form was sent to two doctors by email. In order to achieve high reliability with respect to the accuracy and inaccuracy, we excluded any tweets that a doctor labeled as *not sure*. In addition, we excluded any tweets where the two doctors coding them disagreed on their accuracy. This resulted in 109 labeled tweets in the spreadsheet, 75 (68.8%) of which were labeled as trustworthy (ie, tweets that both doctors labeled as accurate) and 34 of (31.2%) of which were labeled as untrustworthy (ie, tweets that both doctors labeled as inaccurate). The information was then transferred to a spreadsheet and analyzed using Python for descriptive statistics. For the statistical tests, we used the R package, version 3.4.0 (The R Foundation) [129].

The output of this process is illustrated in Figure 3.2.

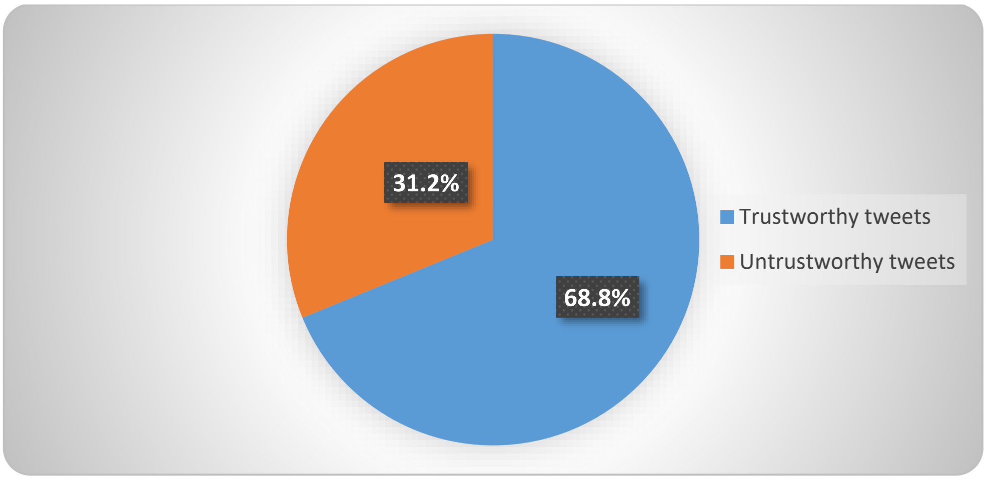


**Figure 3.2 Proportion of trustworthy (n=75) and untrustworthy (n=34) tweets in the sample of 109 labeled, health-related tweets.**

### 3.2.5 Feature Extraction

#### 3.2.5.1 Overview

To answer the research questions, it was necessary to determine the features of the tweets as well as their level of trustworthiness. The features of the tweets were categorized into two types: tweet features and user features.

These features, whether on the user level or the tweet level, may provide data that is useful in classifying the types of users or identifying the credibility of the tweets for different topics [130-134].

#### 3.2.5.2 Tweet Features

Tweet features were extracted directly from the tweets. These included whether or not the tweet was retweeted (ie, as a dissemination measure) as well as various linguistic characteristics of the tweets, such as the number of words and the number of characters in each tweet. Tweet features also included other properties of the tweets, such as URLs contained in the tweet, the time of the tweets, and hashtags. Most of these features have been used by other researchers; however, the linguistic features identified in most other studies were analyzed for Latin-derived words. Thus, more features were added after reviewing literature related to Arabic natural language processing. These features are as follows:

1. Tashkeel: the presence of a tashkeel in the tweet. The tashkeel is a special Arabic character written in the text to represent missing vowels [135].
2. Kashida: the presence of a kashida in the tweet. The kashida is a decorative element in Arabic writing used for justifying the text [136].

In addition, similar to Castillo et al [130], we examined different types of punctuation marks in tweets and their relationship to information credibility. This was based on our insight that people who used punctuation in their tweets appeared more thorough and that this might be associated with greater accuracy and dissemination.

A complete list of tweet features is provided Appendix A.2.

### 3.2.6 **User Features**

#### 3.2.6.1 Overview

Tweets come with metadata that provide basic profile information about the user who posted the tweet, such as screen name, number of friends, number of followers, favorite count, retweet count, and age of the account. In addition, there are cumulative tweeting characteristics for each user. To derive this data, we used the Twitter API to extract another 200 tweets per author by the authors of the 109 labeled tweets. The tweet number of 200 has been suggested as sufficient for extracting user features [131, 137]. The user features were categorized into four groups, which will be described in the following sections. A complete list of these features is provided:

#### 3.2.6.2 Activity and Connectedness Features

These features include metrics that measure how active the user is, such as how often the user replies to other users or how often the user retweets [134].

#### 3.2.6.3 User Linguistic Features

These features include measures of how often the user uses unique hashtags, the average number of hashtags used in tweets, and the mean number of words in the user's tweets. Such linguistic features have been used in other studies to assess information credibility on Twitter [130] and to classify users on SM [131, 138].

#### 3.2.6.4 User Time Features

These features deal with the temporal aspect of the user's tweets. For example, previous research examined the day of the week when tweets were posted to determine if the day is linked to the credibility of news [130]. In the study presented here, more features were added, such as the preferred time of the day—morning, evening, and night—for users to tweet and whether the tweet was posted during weekdays or weekends [139].

#### 3.2.6.5 Users Popularity Features:

These features indicate the popularity of the users, such as the number of followers per user as well as how often users' tweets were retweeted [134].

### 3.2.7 Account Classification

Accounts were classified according to the following criteria [3, 108, 110, 111, 118, 140]:

1. Does the account holder have a health background?
2. Is the account that of an individual or nonindividual?

3. If the account holder is a nonindividual, does the account represent a physical authority or is it an exclusively SM-based account?

Hence, by the end of this process, it was expected that the following categories of user accounts could be analyzed:

1. Individual health accounts.
2. Individual nonhealth accounts.
3. Health organization accounts.
4. Nonhealth organization accounts.
5. Exclusively SM-based accounts.
6. Users whose profiles cannot be extracted and, therefore, remain unknown.

Two annotators classified the types of tweets in the data set based on author accounts and disagreed on seven users. They met to explain their opinions to each other. Finally, after a discussion they agreed on the categories of five of these seven users. An expert in health communication on SM was consulted—a surgeon with a PhD in Health Promotion in New Media—to classify the final two users. In addition, there were about 10 accounts for which profile data could not be extracted; these accounts were classified as *unknown*.

### 3.2.8 Data Analysis

Lancaster et al. [141] recommended that the execution of a pilot study should primarily rely on descriptive and distribution statistics as results. For the continuous variables, it was decided to present the median number, as the median is not affected by the outliers. For the categorical variables, it was decided to employ a statistical test to determine preliminary results. As our data is nonparametric, the Mann-Whitney-Wilcoxon (MWW) test was utilized to establish statistical significance. The MWW test is considered appropriate for nonparametric data and for when the two samples are from different populations (ie, in our case, accurate tweets and inaccurate tweets) and of different sizes [142].

For categorical variables, the Fisher exact test was also used, which is suitable for a small sample of less than 1000 [142]. In the study presented here, there were two categorical variables with more than two values: the type of author and the times at which tweets were posted. Typically, the Fisher exact test does not test for statistical significance for a contingency table larger than 2×2; however, in R it is possible to calculate the P value for larger contingency tables, hence R was used here. The mechanism that allows for the calculation of the P value is based on the work of Mehat and Petal [143] and Clarkson et al [144].

## 3.3 Results

In this section, the most promising results are presented, which indicate factors that may demonstrate the trustworthiness and untrustworthiness of health-related tweets. Overall, more than 100 tweet-level and user-level features in a data set of 109 tweets were explored; these tweets were labeled as either accurate and trustworthy or inaccurate and untrustworthy.

An initial analysis of the tweet-level features indicates that trustworthy health tweets were significantly more likely to have an author that is a member of a list (ie, a curated group of Twitter accounts) ($P=.05$). Although not significant, trustworthy tweets seemed more likely to be favorited by others ($P=.06$).

In contrast, Table 3.3 suggests that untrustworthy health tweets were more likely to have URLs embedded in them (P=.03). Specifically, 24% (8/34) of the untrustworthy tweets had URLs, compared to 8% (6/75) of trustworthy tweets. A total of 4 out of 8 (50%) of the URLs cited in inaccurate tweets referred to news websites, while 2 out of 8 (25%) referred to blogs.

**Table 3.3 Most promising tweet features to help distinguish the accuracy level of tweets.**

| *Metric* | *Description* | *Trustworthy* | *Untrustworthy* | *P value* |
|---|---|---|---|---|
| *URLs* | Does the tweet contain URL | 6/75 (8%) | 8/34 (23.5%) | .03* |
| *Listed* | Is the author listed | 53/75 (70.7%) | 17/34 (50%) | .051* |
| *Fav* | Is the tweet favorited | 46/75 (61.3%) | 14/34 (41.1%) | .06 |
| *Hashtags* | Does the tweet contain hashtag | 20/75 (26.7%) | 4/34 (11.8%) | .13 |
| *Tashkeel* | Does the tweet contain tashkeel | 22/75 (29.3%) | 5/34(14.7%) | .15 |
| *Exclamation_mark* | Does the tweet contain '!' | 3/75 (4%) | 4/34 (11.8%) | .20 |
| *Semi_colm* | Does the tweet contain ';' | 42/75 (56%) | 14/34 (41.1%) | .21 |
| *Retweeted* | Is the tweet accurate or inaccurate | 43/75 (57.3%) | 15/34 (44.1%) | .22 |
| *Kashida* | Does the tweet contain kashida | 6/75 (30.7%) | 8/34 (23.5%) | .49 |

“*P* values were calculated using the Fisher exact test with $P \leq .05$ indicating statistical significance.

The users-level features were also analyzed (see Table 3.4). The analysis revealed the worrying trend that users who had a lower number of followees (F3) were more likely to tweet accurate health tweets (P<.001). More encouragingly, the popularity measure of number of times that author's tweets are favorited (FT2) was associated with trustworthiness, suggesting that accurate tweets were recognized as such. Interestingly, authors who tended to retweet tweets that had hashtags (RMH5) also tended to tweet trustworthy tweets, although the P value is not quite significant in that case (P=.06).

**Table 3.4. Metrics for users who tweet accurate information versus users who tweet inaccurate information.**

| *Metric* | *Description* | *Trustworthy* | | *Untrustworthy* | | |
|---|---|---|---|---|---|---|
| | | Median | Std | Median | Std | *P* value |
| F3 | Followees count | 75.5 | 2517 | 891 | 12952 | *P* < .001* |
| FT2 | Number of times author's tweets are favorited (favorite-author tags) | 281.5 | 76054 | 46 | 2454 | .01* |
| RMH5 | Unique hashtag count in tweets that were retweeted by the author | 6 | 24.2 | 1 | 26.6 | 0.06 |
| OT3 | Number of hashtags in the author's tweets | 18 | 144.4 | 6 | 64.5 | 0.09 |
| RM1 | Number of retweeted tweets by the author where the user mentioned other users | 20 | 76.3 | 13 | 49.7 | 0.09 |
| FT6 | Number of original tweets posted by the author that are favorited | 55 | 60.5 | 26 | 58.7 | 0.11 |
| SSI | Ratio of original tweets posted by the author to tweets retweeted by the author | 4.51 | 61.6 | 8.94 | 93.8 | 0.11 |
| MH5 | Unique keyword count in hashtags set in original tweets posted by the author | 7 | 46.1 | 3 | 22.9 | 0.14 |
| RP1 | Number of reply-to tweets posted by the author | 21 | 47 | 3 | 54.1 | 0.20 |
| RMM5 | Unique mentions in retweeted tweets by the author | 7 | 26.1 | 4 | 24 | 0.43 |
| M1 | Number of tweets where the author mentioned other users | 18.5 | 61.6 | 10 | 66.6 | 0.51 |
| FT1 | Number of tweets favorited by the author | 179 | 1202.6 | 83 | 14073 | 0.84 |

[a]*P* values were calculated using the Mann-Whitney-Wilcoxon (MWW) test with $P \leq .05$ indicating statistical significance.

Figure 3.3 shows the boxplots for the user features, with outliers beyond 90% of the data excluded. This figure suggests that there may be other metrics associated with trustworthiness given a larger data set, such as the number of tweets favorited by the author (FT1), the number of original tweets posted by the author that are favorited (FT6), and the ratio of original tweets posted by the author to tweets retweeted by the author (SSI).

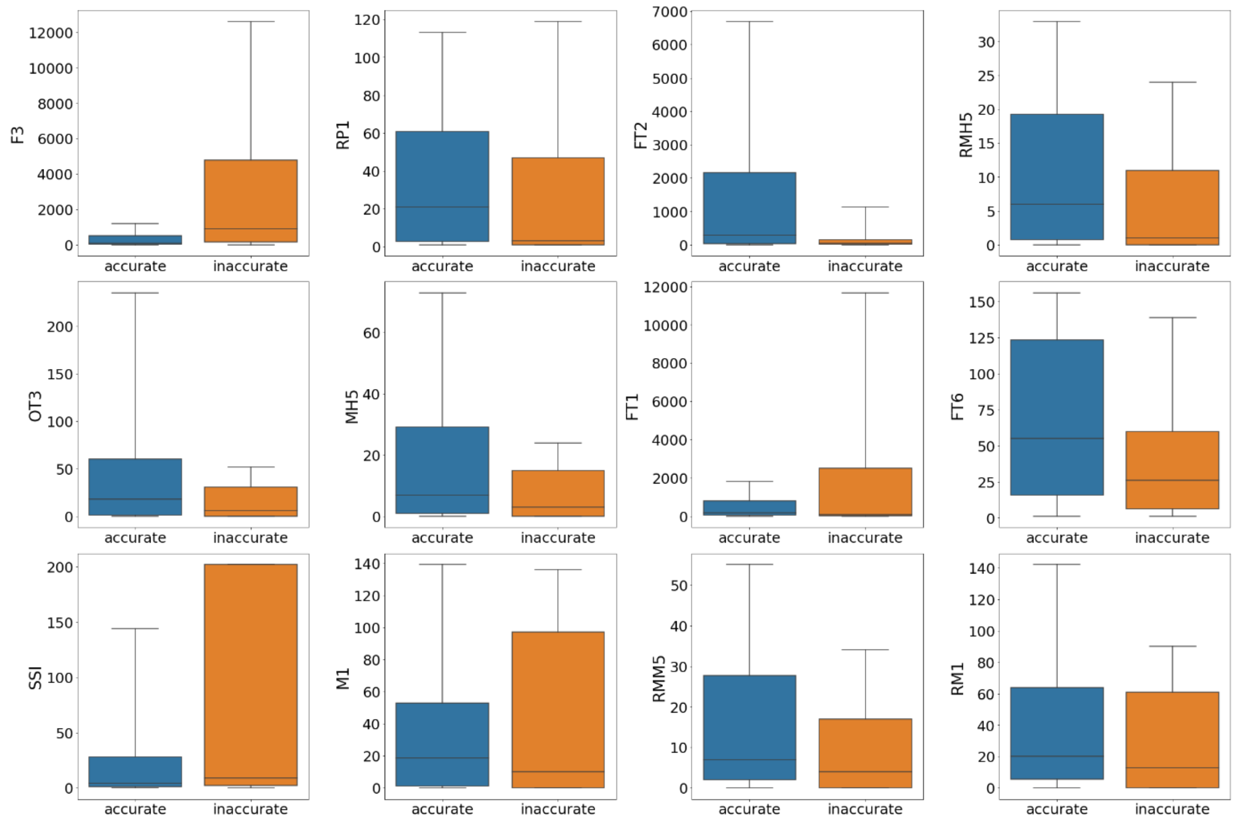


**Figure 3.3 Boxplots for the features most closely correlating with the trustworthiness of tweets.**

As per Figure 3.4, regarding the timing metrics, it appears that the majority of the trustworthy tweets were posted in the morning, while most untrustworthy tweets were posted either in the evening or at night. The associated Fisher test is borderline, with a P value of .06. However, when comparing only two groups—the morning tweets and the night tweets—statistical significance is achieved with a P value of .04. This result aligns with our insight that professional tweets are posted during the daytime and more informal tweets are posted at nighttime.

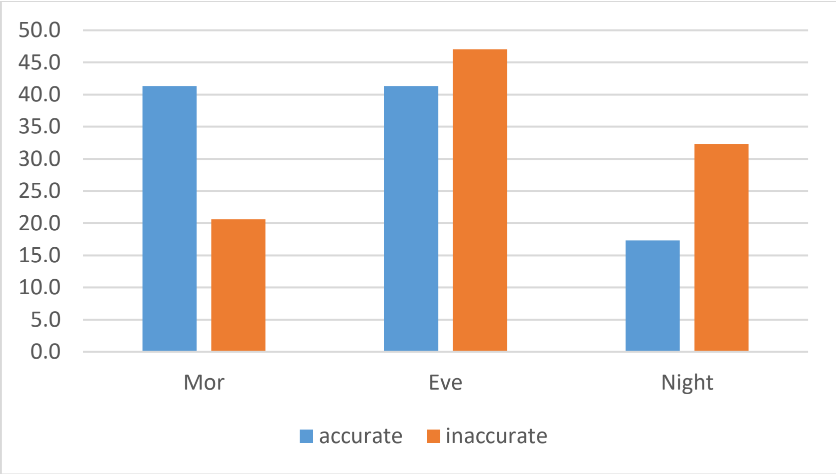


**Figure 3.4 . Distribution of accurate and inaccurate tweets per time of day in three categories: morning (6 am-2 pm), evening (2 pm-10 pm), and night (10 pm-6 am).**

As it can be seen in Figure 3.5, sources classified as organization accounts were more likely to tweet accurate information, followed by sources traditionally considered as trusted users (ie, health professionals, health authorities, and health organization accounts). However, there were only 10 tweets from organization accounts, one of which was from an organization unrelated to health care. Overall, traditionally trusted users are considered the most trustworthy source in the data set, as there were 34 tweets from them and only 4 (12%) were considered inaccurate. The least trustworthy category in the data set includes the users with exclusively SM-based accounts.

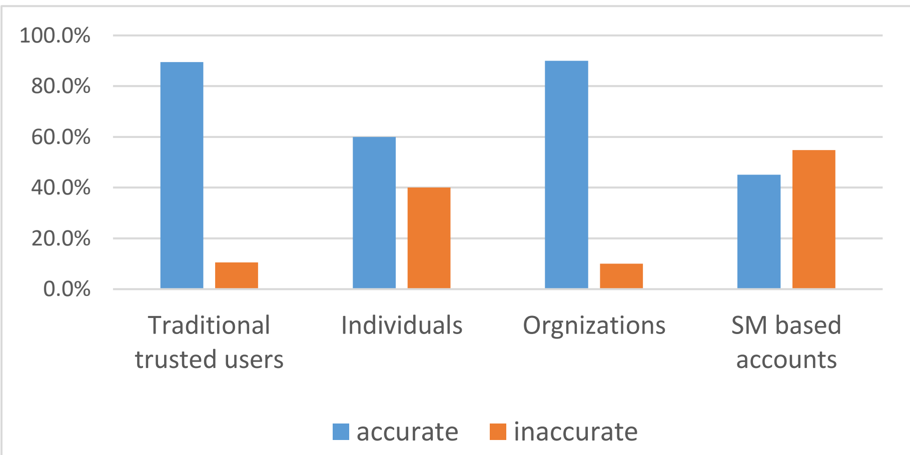


**Figure 3.5 Accuracy of the tweets posted by each author type. SM: social media.**

When the backgrounds of individual users were considered, those with a background in health care (ie, health professionals) seemed to tweet accurate health information and were less likely to tweet inaccurate health information (see Figure 3.6 and Figure 3.7); they posted 29 out of 109 tweets (26.6%). Individuals from a non-health care background were core players in terms of the volume of tweets, with 18 out of 30 (60%) of their tweets labeled as trustworthy. Overall, however, they authored 12 out of 34 (35%) of the untrustworthy tweets and only 18 out of 75 (24%) of the trustworthy tweets.

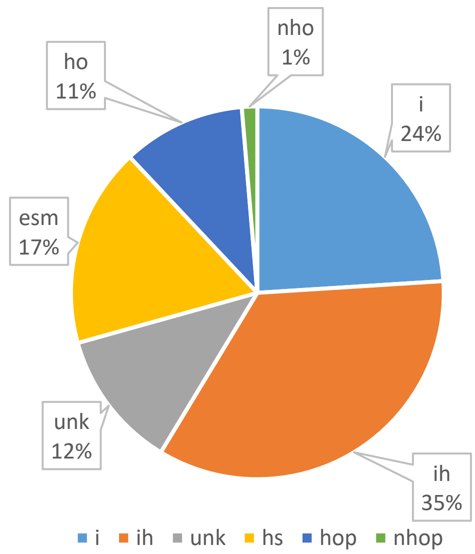


**Figure 3.6 Distribution of the authors of accurate (ie, trustworthy) health-related tweets. esm: exclusively social media-based accounts; ho: health organization accounts; ih: individual health accounts; inh: individual nonhealth accounts; nho: nonhealth organization**

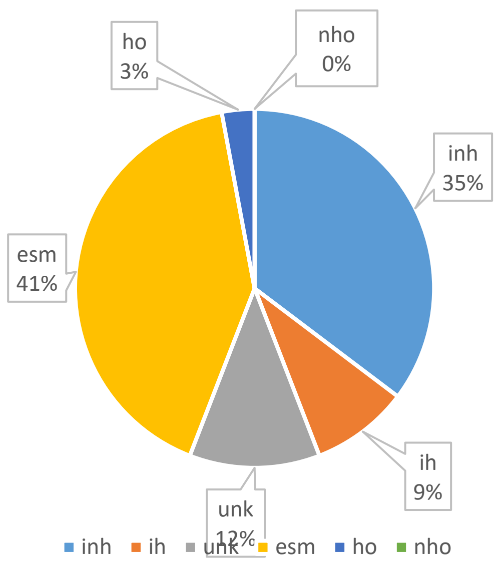


**Figure 3.7 Distribution of the authors of inaccurate (ie, untrustworthy) health-related tweets. esm: exclusively social media-based accounts; ho: health organization accounts; ih: individual health accounts; inh: individual nonhealth accounts; nho: nonhealth organization**

Another interesting finding is that exclusively SM-based accounts seem to tweet much more inaccurate health information than other account types. Therefore, to summarize, trusted health accounts (healthcare organization and professional healthcare individuals) are more likely to tweet trustworthy health information than other types of accounts. Individual users are the main players in terms of the volume of health information on SM, but the high volume does not correlate with trustworthiness. The Fisher exact test indicated statistical significance between the type of the user who posted a tweet and the accuracy of the tweet with $P$ value = .04.

To address the question of whether the trustworthiness of tweets increases their dissemination, a retweeting metric was examined. According to Suh et al [123], retweeting is the key mechanism for information dissemination on Twitter. The results of our study indicate that trustworthy health information was slightly more likely to be retweeted than inaccurate health information, but the difference was not significant. Specifically, 43 out of 75 (57%) of the accurate tweets were retweeted compared to 15 out of 34 (44%) of the inaccurate tweets being retweeted. This is in line with the findings in other studies, which found that health information does not need to be accurate in order to be disseminated [87]. However, when accurate health-related tweets were retweeted, they were more likely to be further retweeted than inaccurate health-related tweets, as shown in Figure 3.8.

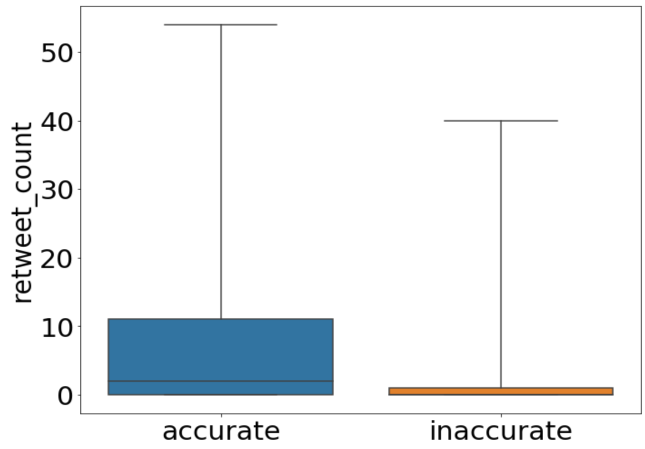


**Figure 3.8 Retweeted counts of accurate and trustworthy versus inaccurate tweets in our data set; outliers outside the 90th percentile were excluded (P=.043).**

The provisional findings suggest that tweets with embedded commas, listed authors, or marked as favorite were also associated with dissemination (P<.001), as indicated in Table 3.5, which also lists a few other factors we examined. The only factor that was clearly counter-indicative of dissemination was the presence of a URL in the tweet, with only 5.2% of these tweets being retweeted.

**Table 3.5. Features for retweeted tweets versus unretweeted tweets.**

| *Metric* | *Description* | *Retweeted* | *Unretweeted* | *P value* |
|---|---|---|---|---|
| *Listed* | Is the author listed | 50/ /58 (86.2%) | 20/51 (39.2%) | *P* < .001* |
| *Fav* | Is the tweet favorited | 51 /58 (87.9%) | 9/51 (17.6%) | *P* < .001* |
| *Comma* | Does the tweet contain comma | 25/58 (43.1%) | 9/51 (17.6%) | .01* |
| *Kashida* | Does the tweet contain kashida | 22/58 (37.9%) | 9/51 (17.6%) | .02* |
| *URLs* | Does the tweet contain URL | 3/58 (5.2%) | 11/51 (21.6%) | .02* |
| *Tashkeel* | Does the tweet contain 'tashkeel' | 19 /58 (32.8%) | 8/51 (15.7%) | .04* |
| *Tweet_accuracy* | Is the tweet accurate | 43/58 (74.1%) | 32/51 (62.7%) | .22 |
| *Hashtags* | Does the tweet contain hashtag | 14/58 (24.1%) | 10/51 (19.6%) | .64 |

[a]*P* values were calculated using the Fisher exact test with *P*<.05 indicating statistical significance.

**Table 3.6. Metrics for users who tweeted accurate information versus users whose tweets were retweeted.**

| *Metric* | *Description* | *Retweeted* | | *Unretweeted* | | |
|---|---|---|---|---|---|---|
| | | Med | Std | Med | Std | *P* value |
| FT2 | Number of times author's tweets are favorited | 857.5 | 8366 | 25 | 121 | *P* < .001* |
| TFF | F1/F3 (rate of followers to followees) | 81.6 | 18950 | 1.09 | 22728 | *P* < .001* |
| Listed_count | Number of lists where the user is member | 29.5 | 912.1 | 1 | 99 | *P* < .001* |
| RT2 | Number of author's tweets re-tweeted by other users | 79.5 | 59.8 | 9 | 25.5 | *P* < .001* |
| FT6 | Number of original tweets posted by the author that are favorited | 77 | 61.5 | 17 | 42.3 | *P* < .001* |
| M2 | Number of unique users mentioned by the author | 16 | 43.1 | 3 | 23.5 | .052* |
| RP1 | Number of reply-to tweets posted by the author | 24.5 | 53 | 4 | 42.2 | .08 |
| SSI | Rate of original tweets posted by the author to tweets retweeted by the author | 3.83 | 63.4 | 21.11 | 87.6 | .11 |
| M1 | Number of tweets where the author mentioned other users | 18.5 | 67.4 | 3 | 47.6 | .12 |
| RM1 | Number of retweeted tweets by the author where the user mentioned other users | 20 | 78.3 | 7 | 56.2 | .12 |
| RT1 | Number of tweets that author retweeted | 19.5 | 49.1 | 7 | 54.4 | .14 |
| OT3 | Number of hashtags in the author's tweets | 20 | 142.5 | 9 | 103 | 26 |

The P value was calculated using the Mann-Whitney-Wilcoxon (MWW) test with P<.05 indicating statistical significance.

The user features TFF, FT2, RT2, FT6 and the listed-count feature appear to have strong association with dissemination as it is shown in Figure 3.9. This comes as no surprise, as these features are typically considered to be popularity metrics. The statistical test confirms statistical significance for most of the popularity metrics as indicated in Table 3.5. Furthermore, the results, as indicated in Table 3.6, show that users who mention other users/posters of tweets (M2) are more likely to have their tweets retweeted.

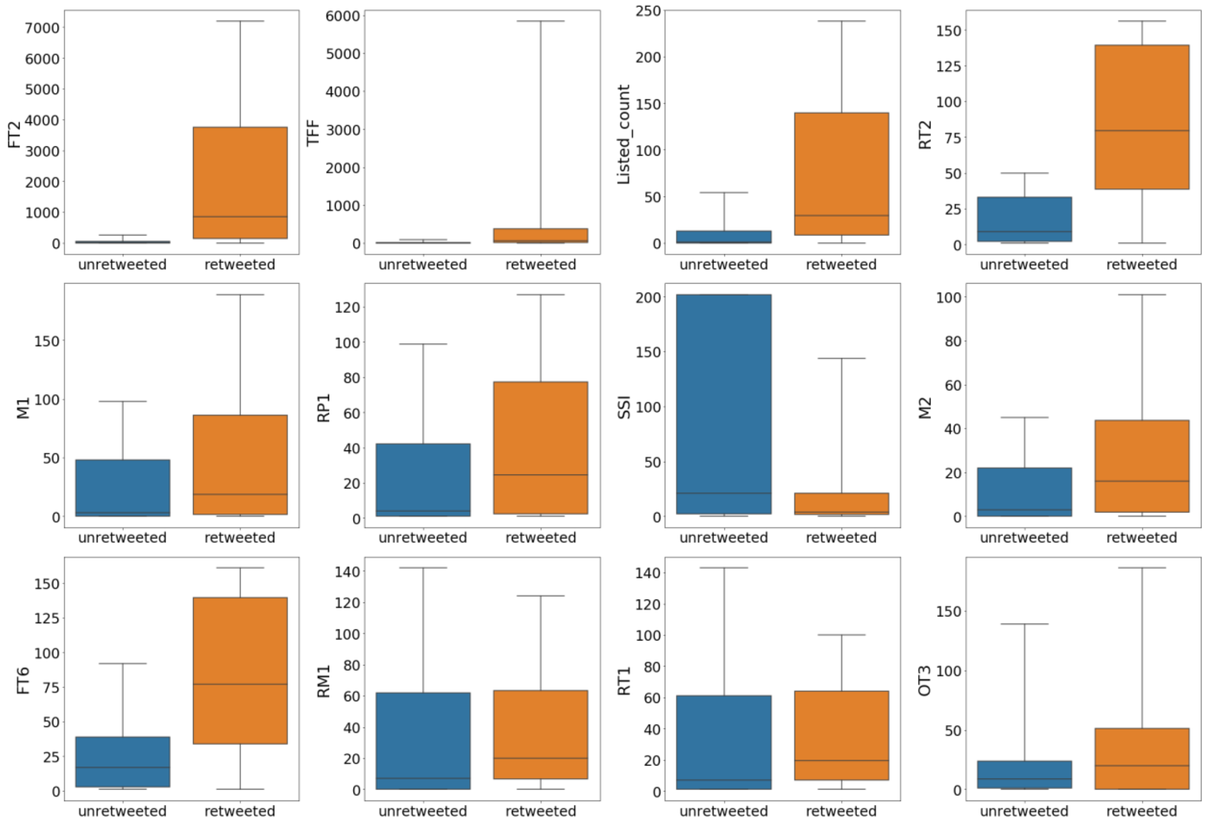


**Figure 3.9 Boxplots for the features most closely correlated to dissemination. outliers outside the 90th percentile were excluded.** FT2: number of times that author's tweets are favorited; FT6: number of original tweets posted by the author that are favorited; M1: number of tweets where the author mentioned other users; M2: number of unique users mentioned by the author; OT3: number of hashtags in the author's tweets; RM1: number of retweeted tweets by the author where the user mentioned other users; RP1: number of reply-to tweets posted by the author; RT1: number of tweets that the author retweeted; RT2: number of author's tweets retweeted by other users; SSI: ratio of original tweets posted by the author to tweets retweeted by the author; TFF: ratio of followers to followees.

## 3.4 Discussion

### 3.4.1 Principal Findings

In terms of individual health tweets, the results of this study do not agree with those of Alnemer et al [12], in that they suggested that 50% of the tweets were not evidence based. However, Alnemer et al labeled every tweet in their data set as either accurate or inaccurate, while in this study, tweets for which not all annotators agreed were excluded. This difference in annotation may explain the differences in the results. Nevertheless, this contradiction indicates the importance of conducting future research to explain these distinctions.

In this preliminary analysis, a group of users linked to the accuracy of the health information was identified, indicating that trusted health users are more likely to tweet trustworthy health information than inaccurate health information. This association is supported by the findings of Medlock et al [82].

Nevertheless, a high proportion of tweets from individuals with no health background were also found to be accurate. This observation suggests the existence of a subgroup of trustworthy SM accounts. The isolation of such a subgroup might be possible through the identification of other characteristics.

But both Wong et al [69] and Suh et al [123] reported that interacting with hashtags was linked to dissemination, while this study provided no clear evidence of such a relationship. Instead, we found that the more a user interacted with other users (ie, number of times that author's tweets are favorited [FT2]), the more likely it was that their tweets were accurate. This finding suggests that trustworthy users have more influence than other users, as FT2 is considered to be an influence metric [134].

Interestingly, the data revealed that most of the accurate health tweets were posted in the morning, while most of the inaccurate tweets were posted at night. This disparity may occur because health professionals may tweet accurate health information while they are at work, possibly as part of their job, while less-trustworthy tweets are more likely posted at night when nonprofessionals are more likely to give an opinion.

In addition, there is no clear answer as to whether trustworthiness is linked to dissemination, because trustworthy tweets were only slightly more likely to be retweeted. However, when considering the retweet count, accurate tweets were more likely to be retweeted more frequently, as shown in Figure 3.8. These preliminary results suggest that there is an association between trustworthiness and the ultimate dissemination of the tweets.

Similar findings were also noted by Kalyanam et al [61]; it raises the question as to why trustworthy tweets are more likely to be retweeted more frequently once they are retweeted. One interpretation might be that there are thresholds for followers that can be exceeded and once they are exceeded, the author might have a certain leverage for their tweets to be retweeted more [145]. However, neither this study nor that of Kalyanam et al looked at self-retweeting specifically. This practice is known to be common in microblogs, such as in circumstances where users retweet to win prizes. Surprisingly, tweets with embedded commas and kashidas were retweeted more, suggesting that correct punctuation may be perceived as a sign of accuracy.

Moreover, some tweet metrics appeared to be linked to both dissemination and trustworthiness; for instance, tweets that embedded the tashkeel were more likely to be retweeted and indicated a trend of possibly being more trustworthy, while tweets that embedded URLs were less likely to be retweeted and trustworthy. These findings contradict those of Wong et al [69], who analyzed specific health accounts and found that URLs in tweets were associated with dissemination.

In regard to the source of the URLs cited in inaccurate tweets, our findings indicated that news websites were the most cited (50%). This is in line with Ghenai and Mejova [84] who found news websites were the most cited source in inaccurate tweets (39% of the URLs).

Our findings, as shown in Table 3.3 and Table 3.5, suggested that the language characteristics of the tweets might be associated with both dissemination and trustworthiness. At a high level, this may suggest that the style in which tweets are written is also linked to dissemination and trustworthiness. At a low level, some of these features are language specific; for example, the tashkeel is used in Arabic but does not exist in Latin languages. This specificity indicates the need to take language type into account when designing any future study. The tashkeel was not tested for significance in trustworthy tweets; as P was equal to .10, these results should be considered for future studies.

In future work, we will seek to develop a machine learning model for classifying health-related tweets as either trustworthy or not trustworthy. To do this, we aim to employ a larger data set and to evaluate the usefulness of a larger set of features as predictors. Some of these features may include measures for the linguistic ability of a user. The extraction of additional features may also require the development of additional machine learning models, such as models for topic detection in order to measure, for example, how often a user tweets about health.

### Internal Validity

Due to the limited data size, this study provided preliminary findings on the relationship between the variables studied. In addition, this study did not establish any causal relationship between variables, only correlations.

### External Validity

The selection of data in this study was limited to health-related tweets in the Arabic language on Twitter. In addition, we collected tweets from a limited time period: two periods of 7 days. Ultimately, we analyzed a small number of tweets and, as a result, we cannot presume generalization for the findings of this study.

We did not include any reply tweets in the analysis; therefore, our results cannot be generalized to interaction-type communication on SM (eg, if the user posts a direct health question to another user). Our results only refer to initial tweets.

Although we used two methods in developing a lexicon, we cannot claim that the lexicon is totally representative of the population. Secondly, in this study we only studied 109 of the 300 tweets labeled by doctors. We excluded tweets where one of the doctors was unsure of the accuracy of the tweet or where there was disagreement between the two doctors regarding the accuracy of the tweet. These measures certainly excluded some health tweets and, more worryingly, may have thus excluded a class of health tweets that were not studied. However, the protocol did heighten the quality of the data in terms of its accuracy. In addition, all 20 doctors who participated in this study as annotators were from the same country, Saudi Arabia, the target of the study.

**Construction Validity**

Although there was a high degree of agreement between the annotators who filtered out tweets not related to health, they did not have health backgrounds. This means that nonhealth tweets may have gotten through this phase. However, this was addressed when the doctors assessed the tweets for health accuracy; they did not identify any tweet as nonhealth related.

This study included the categorization of the authors of tweets into various groups; however, individual health accounts, health organization accounts, and individuals were not externally checked in order to test whether the classification was correct.

In addition, this study intended to examine tweets from Saudi users, specifically during the development of the health lexicon, which was noted in the study's design. However, we cannot guarantee that all tweets had a Saudi origin.

## 3.5 Conclusions

The purpose of this work was to validate the method used to ensure that it was practical for determining the accuracy and the factors associated with the trustworthiness and dissemination of health-related tweets; this was done to provisionally assess factors that may impact on trustworthy tweets and dissemination of tweets. Our results indicate that there may be some clear differences between tweets labeled as trustworthy health information and tweets labeled as untrustworthy health information. They also showed that trusted health professionals were more likely to tweet accurate health information, while exclusively SM-based accounts were more likely to produce untrustworthy tweets. Interestingly, most of the trustworthy tweets were tweeted in the morning, while more of the untrustworthy tweets were tweeted at night. Regarding the dissemination of tweets, there were some features that appeared to be associated with a high dissemination of the tweets. These features appeared at both the tweet-level and user-level analyses.

Due to the limited quantity of data, we cannot have confidence in statistical predictive modelling. The results illustrate that future studies using a large data set may produce a predictive model for classifying tweets as either trustworthy or untrustworthy.

Acknowledgments

We would like to thank Dr. Yousef Albalawi and Dr. Khaled Shahat for their valuable advice. In addition, we thank all the doctors who participated in the process of evaluating health tweets in this study. This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094 and by Taibah University, Al-Ula, Saudi Arabia.

# Preface to Chapter 4

*This chapter was published in the Journal of Big Data (2021). which has a SNIP of 3.478. It is classified as a Q1 journal in information systems, information systems and management, and hardware and architecture. Approximately 65% of the work in this paper was done by the author of this thesis and it currently has three citations.*

To judge the accuracy of health-related tweets, it is first necessary to distinguish between health-related and non-health-related tweets. Since detecting health-related tweets is a classification problem, a literature review is necessary to determine the best method for producing a successful model. Such a review reveals that certain algorithms and techniques have been used to develop models for tweet classification, including traditional machine learning [146-149] and deep learning models [150, 151].

A mandatory step in developing a traditional machine learning model for text document classification is pre-processing the text before it is input into the model. Our review in the area of traditional machine learning models suggests that they lack pre-specified, pre-processing procedures to apply to the text. In contrast, in pre-trained word embedding models, the steps are usually specified in advance by the developers of the pre-trained model [152-154]. Hence, in order to fairly compare deep learning models with traditional machine learning models, the best pre-processing approaches for traditional ML should first be specified.

Chapter 4 sets out to assess and compare several pre-processing strategies with traditional machine learning algorithms. Furthermore, the best types of traditional machine learning algorithms (with optimized pre-processing) are compared to deep learning models in this chapter to identify the best-performing model for detecting health-related tweets.

# 4 Investigating the Impact of Pre-processing Techniques and Pre-trained Word Embeddings in Detecting Arabic Health Information on Social Media


## Abstract

This paper presents a comprehensive evaluation of data pre-processing and word embedding techniques in the context of Arabic document classification in the domain of health-related communication on social media. We evaluate 26 text pre-processings applied to Arabic tweets within the process of training a classifier to identify health-related tweets. For this task we use the (traditional) ML classifiers KNN, SVM, Multinomial NB and Logistic Regression. Furthermore, we report experimental results with the deep learning architectures BLSTM and CNN for the same text classification problem. Since word embeddings are more typically used as the input layer in deep networks, in the deep learning experiments we evaluate several state-of-the-art pre-trained word embeddings with the same text pre-processing applied.

To achieve these goals, we use two data sets: one for both training and testing, and another for testing the generality of our models only. Our results point to the conclusion that only four out of the 26 pre-processings improve the classification accuracy significantly. For the first data set of Arabic tweets, we found that Mazajak CBOW pre-trained word embeddings as the input to a BLSTM deep network led to the most accurate classifier with $F_1$ score of 89.7%. For the second data set, Mazajak Skip-Gram pre-trained word embeddings as the input to BLSTM led to the most accurate model with $F_1$ score of 75.2% and accuracy of 90.7% compared to accuracy of 90.8% achieved by Mazajak CBOW for the same

architecture but with lower $F_1$ score of 70.89%. Our results also show that the performance of the best of the traditional classifier we trained is comparable to the deep learning methods on the first data set, but significantly worse on the second data set.

**Keywords:** Deep learning, Health information, Pre-trained word embeddings, Social media, Machine learning, Natural language processing, Twitter

## 4.1 Introduction

Due to the increased amount of data from user-generated content on social media, text classification has become an important area of research in the last 10 years. This has led researchers to apply text classification methods for analyzing sentiments and topics [33, 58, 155], predicting gender [156-158], and detecting false news [159, 160]. Studies on social media have indicated that, as a wide variety of people use this medium to share health information [26, 28], the information provided is not always accurate [12, 161] and this is a huge issue of concern. However, a precursor for studying the trustworthiness of health-related tweets is the development of a model to detect health-related information posts on social media.

Additional important reasons for devising a high-quality method for identifying health-information posted on social media could include building and/or studying health communication theories, evaluating health communication, and understanding public concerns on social media during an outbreak [3, 162, 163]. Studies that built models to detect (English) health information tweets were conducted by Paul et al. and Tuarob et al.

[164, 165], who developed machine-learning models to detect health-related information on social media platforms.

Unfortunately, these models are highly language-dependent and, as they were not created for the Arabic language, they cannot be directly applied to this language, an important consideration given the prevalence of social media usage in Arabic countries [161]. For example, text normalization is one of the important steps in text classification. In English, this might include normalizing capital letters to lowercase letters, yet there are no lowercase and capital letters in Arabic; normalizing letters in Arabic involves normalizing different forms of alefs (أ إ ا.) to (ا) or removing diacritics that are not used in English. Thus, Maw et al. [166] pointed out that even if some algorithms perform well for a particular language, they might yield worse results when applied to another language.

There have been many studies of text classification regarding Arabic natural language processing on social media. Most of them are focused on sentiment analysis, and a number of literature surveys and systematic literature reviews have been conducted on this Arabic-language-classification-specific task [33, 58, 155]. More specifically, Al-Rubaiee et al. [167], Alayba et al. [149], and Alabbas et al. [148] conducted targeted sentiment-analysis studies. Al-Rubaiee et al. [167] used sentiment analysis to evaluate a bank application. They collected tweets about the bank service and labeled them as either positive or negative. They then pre-processed the tweets using various techniques and compared the performance of the Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers. The best results were for SVM with an accuracy of 89.68%.

Similarly, Alayba et al. [149] collected tweets about health services in Saudi Arabia and labeled them as positive or negative. The best results were achieved using stochastic gradient descent with an accuracy of 91.87%. Moreover, Alabbas et al. [148] trained a classifier to detect natural disasters by labeling tweets, some of which contained information about a flood whereas others did not. They trained different classifiers, namely, SVM with K-Nearest Neighbors (KNN), NB, and compared their performance. The best model was SVM with an accuracy of 90.7%. Alayba and Alabbas studies are expanded on in the next section.

Other Arabic-text classification work used social media data to detect hate speech [168-170] and analyze crisis responses, such as in the event of a flood [171]. However, there is a lack of studies based on detecting Arabic-language health-related tweets. In this paper, we aim to derive a model to accurately detect Arabic language health data on Twitter and test these models on data sets to evaluate the generality thereof.

Statistics show that Twitter is very popular with Arabic speakers, and that it is widely used for sharing health-related information [26, 28]. As such, one of the goals of this paper is to enrich the literature by providing technical details for the development of a model to detect Arabic health-related tweets. Devising such a model can help researchers from many disciplines study health-related tweets in a more comprehensive manner and will provide the foundation for empirical studies that are not conducted with a focus on tweets with a specific origin only (where the origin serves as a means of determining their health-information focus by, for example, only considering tweets emanating from specific health-tweet authors/organizations). For example, while Alnemer et al. [12] extracted tweets from specific health-related Twitter accounts in order to study health-related information on social media, Albalawi et al. [161] pointed out that there are other users who also (more informally) tweet about health and that those should not be ignored in an analysis of health tweets. A model that can automatically extract health-related tweets can further the holistic study of health-related tweets without requiring that specific health-related accounts are followed. Furthermore, providing the technical details for the development of such a model will enrich the literature, not only for this specific text classification task (i.e., extracting health-information tweets), but also for other Arabic-text classification tasks.

This paper is structured as follows. First, we discuss related work in section 2. In section 3, we describe the general methods used in this study, focusing especially on the data sets and evaluation metrics employed. Section 4 reports on the study that assesses the impact of various pre-processings on traditional machine learning techniques, when classifying health-related tweets. Subsequently, section 5 describes a second study which looks at the impact of different word embeddings on deep learning algorithms for the same purpose. Finally, sections 6 and 7 discuss and compare the results, drawing out conclusions from this work.

## 4.2 Related Works

There is a vast body of literature on Arabic text classification for social media. Alayba et al. [149] analyzed tweets to detect sentiment about services in Saudi Arabia. They collected tweets using trending hashtags related to health services, and then they divided their data sets into two categories: negative and positive. When processing the tweets, they removed diacritics and Kashida and normalized three additional letters: اأإ to ا., ة to ه, and ئ to ى; and they used unigram and bi-gram text extraction techniques with Term Frequency–Inverse Document Frequency, hereafter TF-IDF, for feature selection. They then compared the performance of seven algorithms and experimented with a Convolutional Neural Network (CNN). The best results were achieved with a stochastic gradient analysis and SVM, with an accuracy of 91.87. They did not use any stemming methods during pre-processing.

Alabbas et al. [148] developed a model to detect a natural disaster in tweets, specifically a high-risk flood. To achieve this, they trained a classifier on labeled tweets; some containing information about a flood and others that did not. They removed diacritics from the text based on the assumption that most text is written without diacritics. In a manner similar to that of Alayba, they used TF-IDF for feature selection. During their study, Alabbas et al. investigated the performance of different classifiers, specifically the NNET, SVM, KNN, Decision Tree (C4.5–J48), and NB algorithms. Unlike Alayba, they also compared different stemming techniques for the Arabic language: no stemming, light stemming, and prefix/suffix removal. They also normalized one letter, أاإ to ا. The authors concluded that SVM performs better than the other algorithms, and that most of the algorithms included in the study perform better without stemming.

Boudad et al. [146] compared the performance of KNN, SVM and NB in sentiment analysis for Arabic tweets. Moreover, they compared the impact of different types of stemming, specifically light stemming and root stemming; and they also compared TF-IDF to Binary Term Occurrence (BTO) for feature selection. They found that the best accuracy is achieved with light stemming, the SVM classifier, and TF-IDF for feature selection. During the normalization process, they normalized ى and ه in addition to removing hashtags. It is not obvious whether their findings contradict those of Alabbas et al. however, as the model in the earlier study was not trained without stemming.

Duwairi et al. [57] and Oussous et al. [172] studied the impact of root stemming and light stemming in addition to stop word removal on sentiment analysis. While Oussous et al. found that light stemming improves the accuracy, Duwairi et al. stated that stemming and stop word removal do not improve the accuracy of their model. Furthermore, these studies have not investigated the impact of the other pre-processing techniques discussed above. Although, Oussous et al. removed tashkeel, duplicate letters and Kashida, they did not report the impact of such steps on the results of their model.

Abdulla et al. [173] built a model to detect the sentiment of tweets. They found that light stemming decreases model accuracy, which supports the findings of Alabbas et al. In comparison to Boudad et al., however, they only normalized two letters, ه and ا. Like the studies mentioned above, they did not investigate the impact of normalizing letters on the accuracy of their model.

Alakrot et al. [170] developed a model to detect hate speech in YouTube comments, which they trained on 15,000 comments labeled as either positive or negative. They normalized the same letters as Alabbas et al. [148] along with two additional letters, because of the similar morphological sounds thereof. Their best model achieved an $F_1$ score of 82% , and they reported the usefulness of stemming and normalization, which contradicts Alabbas et al. [148] and Abdulla et al. [173].

As the studies described above suggest, there is no agreement on pre-processing steps for the Arabic language as the researchers used different techniques when normalizing the text. Alabbas et al. [148] only normalized one letter, أاإ; Boudad et al. [146] and Abdulla et al. [173] only normalized ة ه; while Alayba et al. [149] and Alakrot et al. [170] normalized other letters. Furthermore, both Boudad et al. [146] and Alakrot et al. [170] reported the usefulness of stemming, while Alabbas et al. [148] and Abdulla et al. [173] found that stemming decreased the accuracy of their models. These conflicting results lead to questions as to which methods are the best for normalizing Arabic data sets, particularly for specific classification tasks.

In addition to traditional machine-learning algorithms, there has been a dramatic increase in the number of studies that apply different deep-learning methods for tackling the Arabic text classification task in the last few years. Some of these studies compared deep-learning models, such as CNN and Long Short-Term Memory (LSTM), to traditional machine-learning models. For example, Oussous et al. [174] compared four models (NB, SVM, CNN and LSTM) to detect the sentiment of tweets. They also investigated the impact of pre-processing techniques, specifically normalizing, stop-word removal, and stemming. They used traditional BTO as feature extraction for NB and SVM, and they used Word2Vec for the word-embedding layer of the CNN and LSTM models. They concluded that normalizing with light stemming improves the accuracy of their model and that the CNN and LSTM classifiers perform better than the SVM and NB ones. They only considered normalizing three letters: ي, ة, and ا.

It is worth noting that word embedding is a learning technique in natural language processing that represents words with vectors [175], the dimensions of which are usually set prior to the word-embedding training. A high dimension vector offers a better opportunity to represent the word semantics [168]. This technique uses geometric word encoding based on how frequently words appear together [160]; thus, words with similar meanings are represented with similar numbers. Yet, to be efficient, word embedding need to be trained on large data sets [176]. Thus, researchers often use already existing pre-trained word embedding as demonstrated by Mohaouchane et al. [150].

They [150] used the same data set that was used by Alakrot et al. [170], and they followed similar pre-processing steps to Alakrot et al.. Mohaouchane et al. [150] used AraVec pre-trained words [152] that were embedded as the input layer for a CNN, and they improved the accuracy of detecting hate speech in this data set from an $F_1$ score of 82 to a score of 84.05.

In contrast to the studies by Oussous et al. [174] and Mohaouchane et al. [150], Abdullah et al. [151] developed a CNN-LSTM model to detect the emotion of tweets. Unlike Oussous et al., Abdullah et al. [151] used AraVec pre-trained words embedding for their input layer.

They claimed that the normalizing and stemming steps did not improve the performance of their model.

Similar to Abdullah et al. [151], Heikal et al. [177] developed a model that uses AraVec pre-trained word embedding in the input layer. They also used different pre-processing techniques by removing diacritics, repeated characters and punctuation. They assembled a model that consisted of a CNN and LSTM architecture. The authors achieved an $F_1$ score of 64%, which they claimed outperforms a state-of-the-art algorithm.

The reason the above-mentioned studies [150, 151, 177] utilized customized pre-processing techniques when using pre-trained word embeddings is unclear. According to Li et al. [178], the ideal method to achieve the most improvement when using pre-trained word embedding is to follow the same steps that were used for the corpus when creating the embeddings vectors unless they are not well-documented. The pre-processing steps to normalize the data sets when using AraVec pre-trained word embeddings are documented and were provided by the models of Soliman et al. [152].

Abuzayed and Elsayed [179] investigated the performance of classical and deep-learning models when detecting hate speech in Arabic tweets. Their results showed that the classical TF-IDF word representation performs better than word embedding with classical algorithms, but the combined CNN-LSTM deep-learning architecture performs better than the classical algorithm. This observation might help to answer the question posed by Guellil et al. [180]: "Are deep-learning approaches really more efficient than traditional approaches, such as SVM, NB, etc., for Arabic natural processing?" (p. 9). This is a core research agenda for this work, but in the context of classifying/identifying health tweets in particular.

While Mohaouchane et al. [150], Abdullah et al. [151], and Heikal et al. [177] used AraVec pre-trained words embedding, there are additional pre-trained Arabic word embedding models that have been investigated. Alwehaibi and Roy [181] asserted that pre-trained models require millions of words to be effectively trained; consequently, they investigated the usefulness of the AraVec, fastText, and the 'Altowayan and Tao' [182] pre-trained word-embedding techniques for text classification. To compare these classification approaches, they developed a CNN-LSTM deep neural network model to predict the sentiment of tweets, and they found that the Altowayan and Tao [182] pre-trained word-embedding method outperforms AraVec and fastText as the authors' best model achieved 93.5% accuracy when classifying texts into positive, negative and neutral sentiment.

Utilizing a collection of 55 million tweets, Fouad et al. [153] developed their own pre-trained word-embedding model by combining three popular techniques—Word2Vec Skip-Gram; Word2Vec Continuous Bag-of-Words (CBOW); and Global Vectors (GloVe). Using the CNN architecture, they compared the performance of their pre-trained word embeddings (ArWordVec) to that of AraVec pre-trained word-embedding methods and found their pre-trained model outperformed the AraVec model.

Based on the literature identified above, Table 4.1 presents the pre-trained word-embedding models that have been applied to the classification of Arabic texts.

It is worth noting that the majority of the studies reviewed above, which used Arabic social media for text classification tasks, used SVM followed by NB. There is also a recent trend of using deep-learning methods for Arabic text classification, where CNN and LSTM architectures were primarily used as deep learning methods. This observation is consistent with the findings of Oueslati et al. [147], who conducted a review on the techniques used for sentiment analysis of Arabic-language tweets.

**Table 4.1 Pre-trained word embedding models**

| Pre-trained word embedding | Number of documents | Sources | Techniques | Availability | Pre-processing |
|---|---|---|---|---|---|
| **fastText [183]** | 400 millions tokens from Wikipedia", i.e. 400 million Wikipedia articles + "24 terabytes of raw text data" from Common Crawl | Common Crawl and Wikipedia | CBOW with sub-wording techniques applied to the methods | Open | Only tokenization. |
| **AraVec [152]** | 66.9 million tweets and 320,636 documents from Wikipedia | Twitter and Wikipedia | CBOW and Skip-Gram with different n-gram and unigram features | Open | Remove non-Arabic letters. replace ة with ه . Normalize alef. remove duplicates, Normalize mentions, URLs emojis. |
| **Mazajak [154]** | 250 million tweets | Twitter | CBOW and Skip-Gram with different n-gram | Open | Removal URLs, Tashkeel, emojis and punctuation |
| **ArWordVec [153]** | 55 million tweets | Twitter | CBOW and Skip-Gram | Open | Normalize mentions, URLs. Remove tashkeel, punctuation, Normalize bare alef. Replace "ى with ي"", Replace "ؤ with ء“", Replace "ئ with "ء, Replace " ة with ه " |

While several recent studies reported the effectiveness of using pre-trained words as the embedding layer for deep-learning models, there have been only few comparative studies of word-embedding techniques in the context of Arabic text mining. For example, four different studies [150, 151, 177, 179] used AraVec, only, one study used fastText [182], and no studies were found that used ArWordVec.

As for traditional methods, the majority of Arabic works have emphasized some pre-processing techniques, such as stemming, but none of the studies discussed determined the impact of normalizing Arabic letters or removing diacritics. Some claimed these techniques negatively affect the classifier performance [184, 185] but did not elaborate on or provide evidence for their assertions. Furthermore, there have been no studies to date on the detection of Arabic health-related tweets on Twitter.

Thus, this paper aimed to investigate the impact of different pre-processing techniques on model accuracy. An additional aim was to employ deep-learning methods to compare the performances of pre-trained word-embedding techniques. This will be carried out through a text-classification task focused on detecting Arabic-language health-related tweets. Using these studies as pre-requisites, this study aimed to compare the best classifiers developed using deep learning methods to best classifiers developed using traditional machine learning (ML) methods to identify the overall best-of-breed classification approach available for health tweet identification.

## 4.3 Methods

We derived two different approaches to achieve the aims of the study.

For the first objective, which concerned testing the impact of pre-processing techniques on the accuracy of the predictive models, we tested 14 variants of normalizing Arabic letters in addition to 12 pre-processing techniques (explained in section 4.1) on four different algorithms. These algorithms are among the most widely used algorithms for text classification [186].

In the second experiment, which was performed to answer the second objective of this study (comparing the performance of the CBOW/Skip-Gram variants of the four pre-trained word embeddings presented in Table 4.1 using a deep learning approach) we specifically used CNN and the BLSTM (Bidirectional LSTM) architecture to compare these pre-trained word embedding models. BLSTM and CNN are among the most used deep learning architecture that have been applied to text classification problem [187, 188].

Lastly, we compared the accuracy of classifier models developed using traditional ML methods to classifier models developed using deep learning methods. Figure 4.1 presents an overview figure for this study.

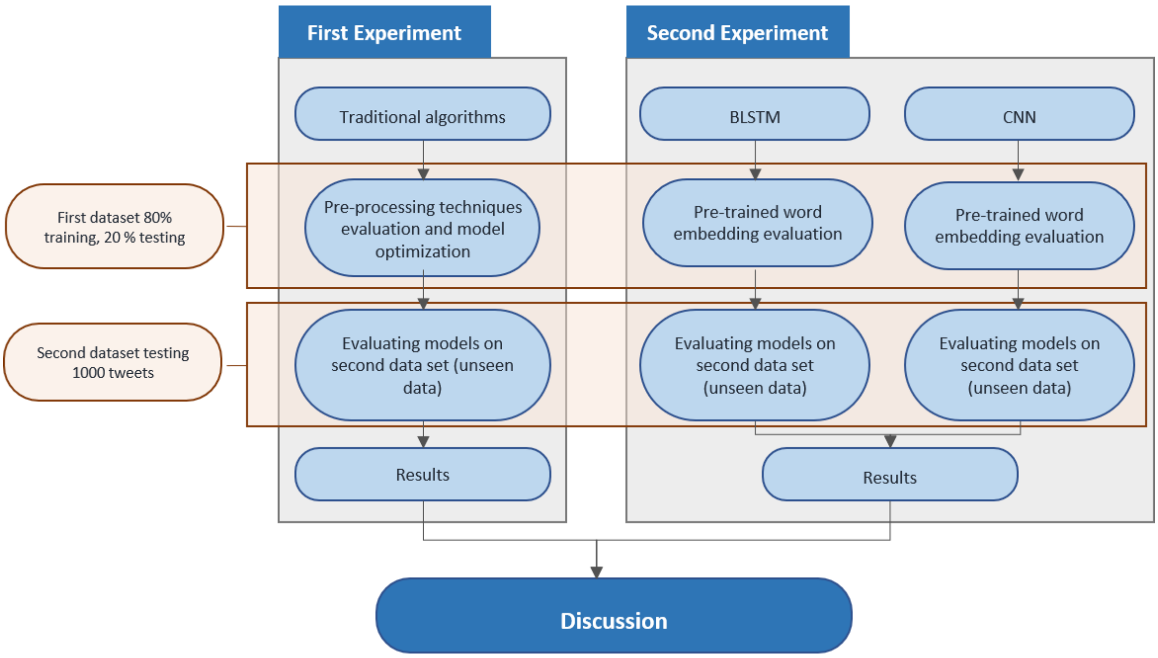


**Figure 4.1 Study overview.**

### 4.3.1 Data Sets and Model Evaluation Metrics

Generalization is the ability of a trained model to accurately categorize new data for which it has not been previously seen/exposed [189]. Chung et al. [190] state that even though most ML-development data is divided into testing and training examples it is questionable whether a ML model would hold in a more general sense as both the training and the test data sets are usually derived from the same environment. Thus, in addition to a first data set, on which each model was trained and tested, we further tested each model on another data set on which the models had not been trained. This data set included words related to COVID-19 and was extracted between March 2020 and April 2020. It differs from the first data set in two ways. Firstly, it was created at a different time-point and secondly, it was extracted during a pandemic, which allowed the model to be tested more for generally.

Next, we describe the process of creating the health lexicon used in extracting these data sets, and then provide more details about each data set.

### 4.3.2 Health Lexicon

The health lexicon, used for extracting health-related tweets, combines keywords from three different sources in order to minimize bias [191]. These sources include are:

- an annotator - a graduate linguist and native Arabic speaker who reviewed health-related accounts and hashtags to identify 110 health-related words.
- field experts - three medical doctors who are active on Twitter; they suggested 100 health-specific words that typically occur in health-related tweets.
- an existing health dictionary – we took 232 words from the Arabic health dictionary proposed by Collier et al. [192]. These 232 words are the only words out of all 968 words in the dictionary, that occur in the tweets we have collected.

We then combined all the words. However, we found that there were still some words not *specific* to health in the lexicon, resulting in a high number of false positive tweets. Thus, similar to Hicks et al. [193], Prus et al., [194] and Zhang and Ahmed [195], we removed these words. Our final lexicon consists of 263 Arabic health-related terms created from the sources above. It is available at http://tiny.cc/ArabicHealthLexicon.

*Note that we chose not to determine whether the words in the health lexicon were MSA or an Arabic dialect. As the words used were found on social media, it's reasonable to expect that the health lexicon has a mix of words from MSA and Arabic dialects.*

*But we did not consider any Arabized tweets for two reasons.*

*First, they were not suggested by the physicians or by the linguists, who suggested keywords based on the language used on Twitter.*

*Second, Arabized words on Twitter are uncommon. According to different studies, the usage varies from 0% to 5% [194, 195]. For Gulf Countries, which include Saudi Arabia—the focus of this study—the use of Arabized words on Twitter has been reported to be 0% [195]. This may explain why the physicians and linguists did not suggest any Arabized keywords.*

### 4.3.3 The First Data Set

Using the health lexicon described above, 297,928 tweets were collected (by employing the Twitter Premium API). These are tweets posted between the 15th July and the 31st August 2019. 5,000 tweets were randomly sampled from the data set. These tweets were independently classified by two annotators as either "health-related" or "not health-related". By following Shoukry and Rafael's [196] procedure, a third annotator was brought in whenever there was a disagreement between the two annotators.

Cohen's kappa statistic for interrater reliability [52] demonstrated excellent agreement between the two annotators independent coding (k = 0.84). As a result, 1,415 of the 5,000 tweets (28.3%) were labeled as health related. Both models were trained on 80% of this data set and were tested on the remaining 20%. This data set is available at http://tiny.cc/AlbalawiDS1.

Data set imbalance typically needs management [197]. At the algorithms level for traditional ML, we tried different models, as explained in section 4.3. For example, one of the algorithms we used, SVM, is known to be less impacted by imbalanced data [198].

Another solution to handle an imbalanced data set is to rework the data sets by re-sampling. However, reworking the data sets in this fashion would increase complexity, and is not guaranteed to increase the model's performance [199, 200]. In addition, the data set that we used is only slightly imbalanced, with a ratio of 1.2:3. Sun et al. [198] states that an imbalanced data set is one having "many more instances of certain classes than others" with Somasundaram et al. [201] going further to suggest that "a data set is considered to be imbalanced if one of its classes plays a huge dominance over the rest of the classes." The imbalance in our data set is not of that scale and Brownlee [60] states that slightly imbalanced data sets should not be a concern: that typically such a classification problem should be treated as classification problem with a balanced data set.

### 4.3.4 The Second Data Set

The second data set used in this study consists of tweets posted between the 20$^{th}$ February and the 31$^{st}$ March 2020. First, we extracted 4,548,839 Arabic tweets using COVID-19-related keywords and then applied our health lexicon to reduce the number of tweets. Finally, we sampled and manually labeled 1,000 tweets from this data set, which is the same number of tweets used for test from the first data set. In this sample, 188 tweets are labeled as health related. We refer to this data set as *unseen data* in this chapter as it was not used in any way for training or evaluation of classifiers and it is collected from a different time period compared to the first data set. This data set is available at http://tiny.cc/AlbalawiDS2.

Please note we only share tweet IDs and the labels as the Twitter policy prevents the content of the tweets to be redistributed, apart from tweets IDs, that can be used to obtain the text of the tweets with the Twitter API [202].

#### 4.3.5 Evaluation Metrics

To evaluate the traditional algorithms, we used the $F_1$ score. The $F_1$ score is a recommended metric for imbalanced data sets, while accuracy is the recommended metric for a balanced data set [60]. To evaluate the final model in both experiments and make the comparison between them, we used four metrics - recall, precision, $F_1$ score, and accuracy. These are the most-used metrics to evaluate ML model performance [58, 59]. Yet, as per the first experiment and the recommendation of Brownlee [60], we used $F_1$ score as the decisive metric to select the best-of-breed model.

## 4.4 First Experiment

The first experiment concerned traditional ML algorithms. It evaluated the importance of different pre-processing techniques and their impact on classification.

### 4.4.1 Common Pre-processing Techniques

By reviewing the literature, we were able to identify more than 26 pre-processings for potential analysis:14 variants of normalizing Arabic letters in addition to 12 techniques have been applied in the pre-processing steps on Arabic-language social media data:

#### 4.4.1.1 Tokenization

In the tokenization process, text is divided into units, and typically here, those units are words. They are usually delimited by spaces or punctuation, and the results are referred to as tokens [203].

#### 4.4.1.2 Noise removal:

Noise removal aims to eliminate unwanted characters from the text. In the literature, we found the following techniques used:

##### 4.4.1.2.1 Removal of non-Arabic letters

Several of the aforementioned studies [148, 151, 204-206] removed non-Arabic data from the text examples.

##### 4.4.1.2.2 Removing numbers

Numbers do not always contribute additional information about the text. We found three studies [207-209] that emphasized number removal from the sampled texts.

##### 4.4.1.2.3 Removing usernames, external links, and hashtags

Usernames, external links, and hashtags are found in many tweets. Three of the cited studies [210-212] removed these from the text.

#### 4.4.1.3 Normalization:

Normalization is a process that converts a list of words to a more uniform sequence [168]. In the literature, we found 5 techniques used for this:

##### 4.4.1.3.1 Removal of punctuation

Punctuations marks typically do not add extra meaning to the text, although punctuation sometimes has a useful meaning, especially when analysing sentiment [213]. Eight of the previous studies [151, 204, 205, 214-218] removed the punctuation from the examined text.

#### 4.4.1.3.2 Removal of tashkeel (diacritics)

Diacritics are often used to represent short vowels, gemination or nunation [219]. In Arabic there are 8 basic diacritics and if combined they can form a total of 13 different diacritics [220]. Four of the previously described studies [181, 212, 221, 222] removed diacritics.

#### 4.4.1.3.3 Removing repeated characters

Because some users use repeated characters when they want to emphasize something, researchers refer to this as the speech effect. Several of the cited studies [204, 206, 209, 223, 224] removed these characters.

#### 4.4.1.3.4 Removal of duplicate letters

The rationale for these removals is similar to that for removing repeated characters. However, some argue that many Arabic words originally contains repeated letters, so they only deleted characters if they occur more than twice. An example of this is the work of Alqarafi et al. [217], who deleted duplicate letters if they occurred more than twice.

#### 4.4.1.3.5 Removing Kashida

Kashida, also known as *tatweel*, is a decorative element in Arabic writing used to justify or stretch the text with a phonetic value [225]. We found two studies [177, 185] that removed Kashida.

### 4.4.1.4 Arabic-specific Normalization

Arabic is considered a Semitic language, with script written from right to left. The Arabic language has 28 letters. However, as some Arabic letters are phonetically similar, users on social media frequently misspell words by using the wrong but phonetically similar letters [170]. In addition to some phonetically similar letters, some letters can be written in more than one form. This might be more apparent in the case of the alef variances “أاإآ” , which are often written as a bare alef “ا”; possibly due to their similarity in appearance [226, 227]. For example, the word “أنت”, which means 'you' in modern standard Arabic, is commonly written as “انت”, i.e. without the hamza “ء”, and some people might even misspell it and write it as “إنت” , with the hamza under the alef [228]. Thus, different forms of alef are unified as a bare alef.

Hence, Arabic-specific normalization indicates that the normalization is specific to the Arabic language as it directly deals with Arabic letters; therefore, it is not possible to apply these Arabic-specific normalization techniques to other languages. In the literature, some researchers have normalized two letters, while others have normalized five or six letters. Furthermore, the same letters are sometimes normalized in different ways. For example, “ي” and ئ” have been replaced with “ى” [229], and “ىء” and “ئ” have been replaced with “ي” [215]. Table 4.2 summarizes the most-used techniques for normalizing Arabic letters that were presented in the literature.

Please note that this study does not aim to be conclusive regarding all possible pre-processing techniques. It focuses on pre-processing techniques commonly identified in our literature review, and only those techniques. While this is not entirely systematic, it does provide high coverage of work-to-date and it improves on current research practice where the basis for selecting pre-processing techniques is not presented explicitly [57, 172, 213]. However, future work should consider this issue carefully, to incorporate additional, relevant pre-processing techniques not yet considered, or only tangentially considered, by the community.

**Table 4.2 Normalization techniques used by different researchers**

| Replace | With | Relevant Studies |
|---|---|---|
| أ, إ, and آ | Bare-alif ا | [146, 148, 170, 208, 211, 229-234] |
| ى | ي | [146, 169, 215, 221, 230-235] |
| ي and ئ | ى | [229] |
| ىء and ئ | ي | [215] |
| ؤ and ئ | ء | [214, 231, 233, 236, 237] |
| ئ | ى | [222] |
| ة | ه | [149, 211, 222, 231-234, 236-238] |
| چ | ج | [237] |
| ڤ | ف | [237] |
| ءي and ءى | ئ | [208] |
| ص | س | [170] |
| ض | ظ | [170] |
| ؤ | و | [208, 215, 236] |
| كـ | ك | [179, 214] |

#### 4.4.1.5 Removing stop words

Many studies removed stop words. There are several methods of removal for Arabic stop words. Examples of studies that removed stop words are [151, 167, 230].

### 4.4.2 Stemming

Stemming is the process used to get the stem from the word. To achieve this, three different techniques are used in the literature:

#### 4.4.2.1 Light stemming

Light stemming is the process of removing the prefixes, infixes and clitics from words. For light stemming, we used the Tashaphyne Python library [239]. This method was used in three of the cited studies [167, 237, 240].

#### 4.4.2.2 Root stemming

Root stemming, which is also called also heavy stemming, aims to transform a word to its root [220]. It is usually faster to perform than lemmatization (see Section 4.1.6.3). In

Arabic, most word roots consist of three letters [241]. Thus, the results of root-stemmed words will be mostly words made from three letters. For the root stemming, we used the Tashaphyne Python library [239].

#### 4.4.2.3 Lemmatization

Lemmatization has a similar aim as root stemming in that the aim is to return a word to its origin; however, unlike root stemming, lemmatization uses a lexicon or dictionary to map a word to its root. Thus, in the present study, to get the roots of Arabic words, we mapped a word to its roots using the dictionary Qalsadi [242].

### 4.4.3 Feature Extraction

The feature-extraction process transforms text into vectors [243]. Bags of words (BOW) and TF-IDF are the two most-used methods for extracting features from the text. In BOW, words frequencies are counted, and word position is ignored. TF-IDF is considered to be a statistical approach that is more sensitive for less-general words as TF measures term frequencies in the text, and IDF is a proxy for the importance of a term [244].

### 4.4.4 Classification Algorithms Used

#### 4.4.4.1 Multinomial NB

NB is a probabilistic model, and in its basic version it is one the most-used algorithms in text classification [245], including sentiment analysis [217, 246] and spam filtering [247]. In this work, we used the variation of NB knows as Multinominal NB (MNB) [248].

#### 4.4.4.2 SVM

SVM, which is grounded in statistical learning theory, is one of the most popular ML classification methods.

SVC and NuSVC are implementations of support vector machine classifiers. They are quite similar and are both based on LIBSVM (Library for SVMs), which was devised by Chang and Lin [249]. LinearSVC is based on the work of Fan et al. [250] and is more flexible than SVC because it provides more options for penalties and choices of loss functions [251]. In this study, we used LinearSVC.

#### 4.4.4.3 Logistic regression

Logistic regression is a linear classifier that uses a hyperplane to separate two classes. This algorithm was used in the present study to differentiate between health-related and non-health-related tweets in accordance with the work of Dressel and Farid [252].

#### 4.4.4.4 KNN

KNN, is fundamentally different from other algorithms discussed in this paper because this algorithm memorizes the training data set rather than learning discriminative functions, and it is thus classified as a memory-based approach [253].

### 4.4.5 Experiment Setup and Results

The setup of the first experiment consists of three phases. The baselines for each algorithm were first developed. Each pre-processing technique was then individually tested on each of the four algorithms, and the results were compared against the baseline for each algorithm. It would be computationally expensive to apply all combinations of pre-processings for each algorithm. Thus, the approach followed in this study is to evaluate each pre-processing technique with the four selected algorithms. We then apply the combination of pre-processing techniques that best enhance the model performance in the second phase, using brute force to combine the pre-processing techniques and find the best combination. Lastly, we choose the best model with the best combination to evaluate on the second data set. Figure 4.2 illustrates these three phases and the flow of the first experiment.

#### 4.4.5.1 Developing baselines

We trained four baselines models without applying any of 26 pre-processings. Moreover, we use Python Grid Search and Pipeline to tune hyperparameters as well as to apply other pre-processing techniques that this paper does not assess, which are outlined in Table 4.3. Each algorithm has a number of hyperparameters that must be tuned, so they are also "brute-forced". The hyperparameters for each algorithm are outlined in Tables 1–4 in Appendix B.1.

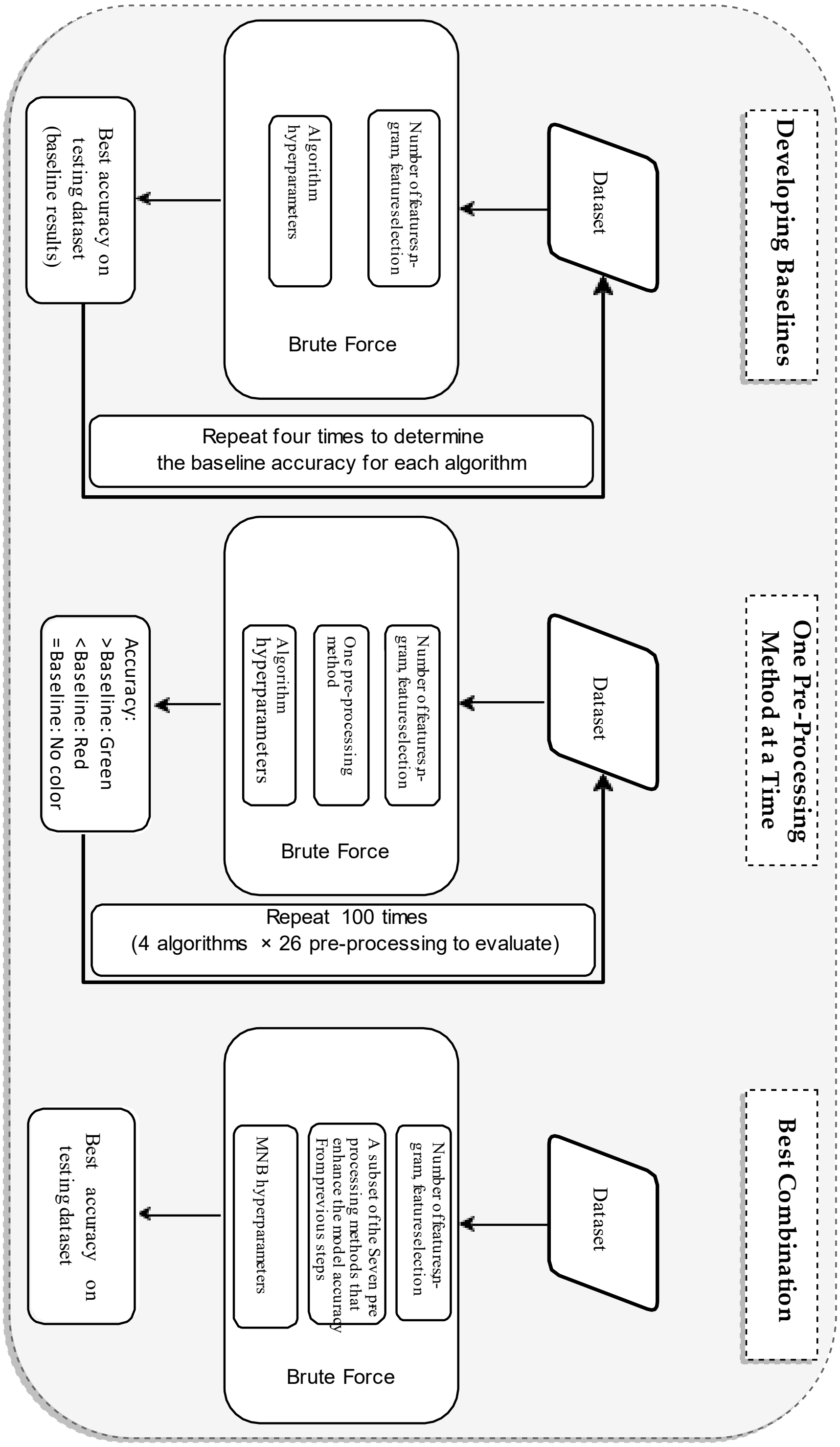


**Figure 4.2 Overview flowchart for the first experiment. (In the best combination, we only tried MNB as it was the best algorithm from previous steps.)**

**Table 4.3 Techniques used as brute-force algorithms in each attempt for all algorithms.**

| Techniques used | Parameters or range used |
|---|---|
| Number of features | Ranges from 7,000 to 18,000 |
| N-gram | (1, 1), (1, 2), (1, 3), (1, 4) |
| Feature selection | Count vectorizer and TF-IDF |

Table 4.4 outlines the accuracy results achieved for the testing data set without applying any pre-processing methods that this paper aimed to investigate. Hence, these models were used as baselines to compare to the impact of pre-processing. It is important to note that these were not used as a standard for further development; instead, we used the best achieved accuracy as the baseline to judge whether other pre-processing methods improve the results or not.

**Table 4.4 Baseline results for four algorithms used in this study**

| Algorithm | N-gram | Feature selection | $F_1$ score |
|---|---|---|---|
| LinearSVC | 1 | Term frequency | 84 .0 |
| Logistic regression | 1, 2 | TF-IDF | 84.0 |
| Multinomial NB | 1, 2 | TF-IDF | 86.0 |
| KNN | 1, 2 | TF-IDF | 77.6 |

*Finally, to reduce the overfitting and enhance the prediction performance of our models, we utilized cross-validation. To perform cross-validation, the dataset was partitioned into k subsets/folds of equal size. Then the model was trained k times with one of the folds used for testing, and the other k-1 folds used for training. There is no formal rule in choosing the number of folds [254]. We experimented with several k values ranging from 2 to 10, and four folds produced the best results, thus it was used in this work.* We used Python and the scikit-learn Version 0.22 library to conduct these experiments [255].

#### 4.4.5.2 Using one pre-processing method at a time

The accuracy of each algorithm without applying pre-processing techniques is used as a baseline to compare with the performance of each of the pre-processing methods discussed in Section 3.1. We employed methods that are similar to those used by Symeonidis et al. [213], who compared the impact of these pre-processing techniques on classifiers trained for the sentimental analysis of English language. We applied one technique at a time, applied to each of the four classification algorithms. The model that achieved the most accurate results was then selected for further refinement, with all the pre-processing combinations that were shown to enhance the accuracy of the model presented in Table 4.5. Take, for example, the seven techniques that enhance the MNB classifier. All the possible combinations of those seven were calculated by the following equation $2^n$, where n is the number of pre-processing techniques. Therefore, we tried 128 variations as $2^7 = 128$. The results of these experiment are found in Appendix B.2.

As mentioned above, the results show that 7 pre-processing and normalization techniques improved the MNB and logistic regression performance in terms of $F_1$ score, 15 techniques improved LinearSVC and 13 improved KNN. It is worth noting that Light Stemming, Lemmatization and Remove repeated characters improved the $F_1$ score in all the models we tried, whereas Remove non-Arabic letters reduced the $F_1$ score in all the models, as shown in Table 4.5.

**Table 4.5 Accuracy (in percentages) of each of the pre-processing techniques used for the extracted tweets**

| | Techniques used | MNB | Logistic regression | LinearSVC | KNN |
|---|---|---|---|---|---|
| | Baseline models | 86.0 | 84.0 | 84.0 | 77.6 |
| 1. | Remove non-Arabic letters | 85.4 - | 83.6 - | 82.8 - | 76.7 - |
| 2. | Remove numbers | 85.5 - | 82.9 - | 84.3 + | 77.4 - |
| 3. | Remove usernames, external links, and hashtags | 85.2 - | 83.2 - | 83.4 - | 78.1 + |
| 4. | Remove punctuation | 86.0 | 84.0 | 84.0 | 77.6 |
| 5. | Remove diacritics | 86.0 | 83.6 | 83.8 - | 76.6 - |
| 6. | Remove repeated characters | 86.4 + | 84.3 + | 84.9 + | 79.2 + |
| 7. | Remove duplicate letters | 86.0 | 83.2 - | 84.1 + | 78.7 + |
| 8. | Remove Kashida | 86.3 + | 83.8 - | 84.6 + | 78.0 + |
| 9. | Replace إ, أ, and آ with ا | 85.8 - | 83.6 - | 84.1 + | 77.4 - |
| 10. | Replace ى with ي | 86.7 + | 84.0 | 84.6 + | 77.9 + |
| 11. | Replace ي and ئ with ى | 86.8 + | 84.0 | 84.2 + | 78.0 + |
| 12. | Replace ىء and ئ with ي | 86.0 | 83.0 - | 84.3 + | 77.8 + |
| 13. | Replace ؤ and ئ with ء | 85.8 - | 83.8 - | 83.9 | 77.7 + |
| 14. | Replace ئ with ى | 86.0 | 84.0 | 84.3 + | 77.6 |
| 15. | Replace ة with ه | 86.7 + | 83.8 - | 84.8 + | 77.1 - |
| 16. | Replace چ with ج | 86.0 | 84.0 | 84.0 | 77.6 |
| 17. | Replace ڤ with ف | 86.0 | 82.8 + | 84.0 | 77.6 |
| 18. | Replace ءي and ءى with ئ | 86.0 | 84.0 | 84.0 | 77.6 |
| 19. | Replace ص with س | 85.7 - | 83.7 - | 83.6 - | 78.0 + |
| 20. | Replace ض with ظ | 86.0 | 83.6 - | 84.0 | 77.9 + |
| 21. | Replace ؤ with و | 85.8 - | 82.8 - | 84.2 + | 77.6 |
| 22. | Replace ک with ك | 86.0 | 84.0 | 84.2 + | 77.6 |
| 23. | Remove stop words | 85.2 - | 84.4 + | 83.4 - | 76.6 - |
| 24. | Light Stemming | 86.6 + | 85.3 + | 86.2 + | 79.1 + |
| 25. | Root stemming | 84.4 - | 85.2 + | 85.1 + | 77.8 + |
| 26. | Lemmatization | 86.7 + | 86.2 + | 86.5 + | 80.1 + |

**Note: Plus sign (+) indicate the technique improved the $F_1$-score of the baseline model; negative sign(- ) indicate the technique decreased the $F_1$-score; and cells without sign indicate the technique had no impact on the $F_1$-score of the algorithm.**

#### 4.4.5.3 Best combination

In the third phase, we used a brute-force algorithm to determine the best combinations of the favourable pre-processing techniques discussed above. This phase focused on MNB, as it achieved the best performance for all but one variant in the previous phases.

It is worth mentioning that, for the MNB model, not all the pre-processing techniques listed in Table 4.5 as favourable were shown to be the most effective in combination. For example, out of the seven pre-processing techniques that improved the MNB classifier, only four contributed to the best combination. In other words, after experimenting with all the combinations, we found that MNB achieved the best $F_1$ score with a combination of remove duplicate, remove Kashida, replacing ة with ه and replacing ى with ي. This combination improved the $F_1$ score from 86.0% to 87.9% on the first data set. In terms of generalization, when we applied the best model on the second data set, the accuracy of the algorithm sharply decreased to 60.54%, which might be due to the fact there were words included that the algorithms had not seen before. These results are shown in Table 4.6.

**Table 4.6 Results for MNB classifier with the best combinations**

| Model | **Baseline Classifier** | | | | **Optimized classifier** | | | |
|---|---|---|---|---|---|---|---|---|
| Metrics | **Precision** | **Recall** | **$F_1$ Score** | **Accuracy** | **Precision** | **Recall** | **$F_1$ Score** | **Accuracy** |
| **First Data set** | 87.5 | 84.6 | 86.0 | 91.6 | 89.1 | 86.6 | 87.9 | 92.7 |
| **Second Data set** | 61.5 | 55.0 | 58.1 | 85.1 | 66.5 | 55.6 | 60.5 | 86.4 |

## 4.5 Second Experiment

In the second experiment, we aimed to investigate four pre-trained word embedding models for Arabic found in the literature using deep learning methods. These pre-trained models were summarized in Table 4.1 above. We also aimed to compare the best classifier model produced in this experiment to the best classifier model produced using tradition ML methods, as a result of the first experiment.

In the second experiment, we trained a classifier using a deep-learning approach. As this work sought to generalize a model for new data, we use pre-trained words as the input layer for the model. According to the literature and, as described in an earlier section, there are four pre-trained word embedding models, all of which are found in Table 4.1.

Using trained word-embedding models provides an opportunity for the classifier to correctly classify words that were not seen in the training data set [256], which solves the problem in traditional text classification that occurs when the classifier fails upon encountering an unseen word [257]. For pre-processing text in the second experiment, we employed the same steps provided by the authors of pre-trained word embeddings models. According to Li et al. [178], the ideal method to achieve the most improvement when using pre-trained word embedding is to follow the same steps that were used for the corpus when creating the embeddings vectors.

### 4.5.1 Models

We experimented with BLSTM and CNN architectures for the classification task in order to compare the different pre-trained word-embedding techniques.

### 4.5.2 BLSTM

Assuming that the input to a neural network is a sequence of data, LSTM is a type of recurrent neural network that is designed to learn and take advantage of dependencies between parts of the input sequence. Text is a sequence of words, and the LSTM architecture has been found to give good results in text classification tasks, specifically in its BLSTM variation, which learns dependencies on both past and future elements in the input sequence [258]. In this work, we experimented with an BLSTM architecture similar to the one proposed by Soufan [259]. The BLSTM model begins with an input and embedding layers to which a dropout layer is added, and this is followed by the BLSTM layer with an added dropout layer. To reduce the dimension of output from this model, a global max-pooling layer is used, as shown in Figure 4.3.

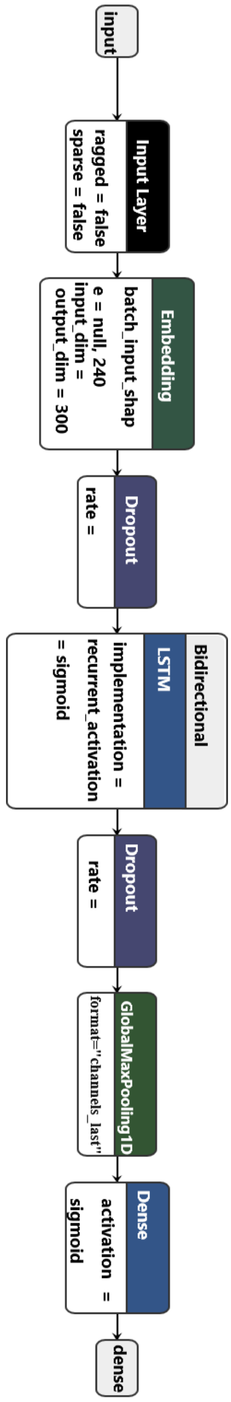


**Figure 4.3 BLSTM architecture.**

### 4.5.3 CNN

While CNN was originally proposed for image analysis, this deep learning architecture was recently proven to perform effectively on many text classification problems; in fact, it sometimes performs better than other approaches, including BLSTM [150, 260]. In this work, we proposed an architecture that is similar to that of Mohaouchane et al. [150], Heikal et al. [177] and Fouad et al. [153]. The proposed model begins with input and embedding layers followed by three CNN layers, each of which contains input and embeddings from the previous layer. Max-pooling layers are used after each of the CNN layers to reduce the output dimensions, and all output from these layers is concatenated and flattened before including a fully connected layer. Figure 4.4 illustrates the architecture of the CNN model used in this study.

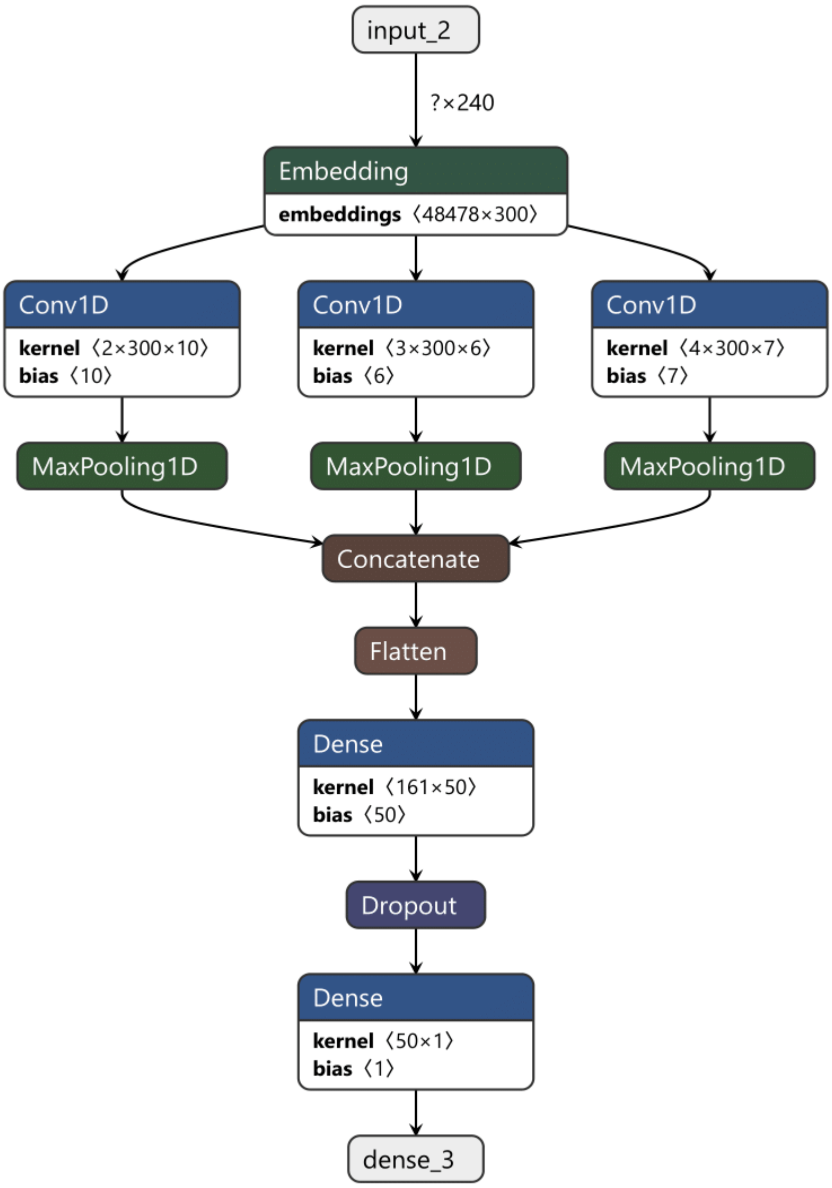


**Figure 4.4 CNN architecture.**

### 4.5.4 Hyperparameter Tuning

There are different hyperparameters that must be tuned to optimize the performance of the model. Several methods are suggested in the literature, including random search, grid search, and the Bayes method [261, 262]. According to Hutter et al. [261] and Feurer and Hutter [262], the Bayes method outperforms other tuning methods. We therefore used the implantation of this algorithm in the Keras Tuner Python library [263]. We limited each experiment to 200 rounds, and the model was terminated if the rounds did not achieve optimal results, with only the best results then used. Table 4.7 outlines the best hyperparameters for the CNN model, and Table 4.8 outlines the best hyperparameters for the BLSTM model.

**Table 4.7 optimal hyperparameters of the CNN**

| | **CNN_Mazajak sg** | **CNN_Mazajak cbow** | **CNN_fasttext** | **CNN_arwordvec sg** | **CNN_arwordvec cbow** | **Cnn arvec sg** | **Cnn arvec cbow** |
|---|---|---|---|---|---|---|---|
| **cov_filter** | 32 | 32 | 1 | 1 | 32 | 1 | 32 |
| **cov_filter1** | 32 | 32 | 1 | 32 | 32 | 32 | 32 |
| **cov_filter2** | 1 | 32 | 1 | 32 | 32 | 1 | 32 |
| **cov_kernel** | 32 | 32 | 32 | 32 | 32 | 1 | 1 |
| **cov_kernel1** | 1 | 1 | 32 | 1 | 32 | 32 | 32 |
| **cov_kernel2** | 1 | 1 | 1 | 1 | 32 | 1 | 1 |
| **pool_filter** | 32 | 32 | 1 | 32 | 1 | 1 | 1 |
| **cov1_activation** | relu | sigmoid | relu | relu | relu | sigmoid | relu |
| **cov1_activation1** | relu | relu | relu | relu | sigmoid | relu | relu |
| **cov1_activation2** | sigmoid | relu | relu | relu | relu | relu | relu |
| **dropout_1** | 0.0 | 0.0 | 0.6 | 0.600 | 0.600 | 0.300 | 0.000 |
| **dense_units** | 380 | 20 | 20. | 20.000 | 380.000 | 20.000 | 380.000 |
| **Dense activation** | relu | relu | sigmoid | relu | relu | sigmoid | sigmoid |
| **dropout_2'** | 0.0 | 0.000 | 0.5 | 0.0 | 0.0 | 0.5 | 0.0 |
| **learning_rate** | 0.01 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| **Batch size** | 40 | 40 | 40 | 40 | 40 | 40 | 300 |

**Table 4.8 optimal hyperparameters of the BLSTM**

| | Mazajak sg | Mazajak cbow | Fasttext | ArWordVec sg | ArWordVec cbow | arvec sg | AraVec cbow |
|---|---|---|---|---|---|---|---|
| **Neurons** | 280 | 360 | 20 | 500 | 500 | 500 | 500 |
| **Drop rate** | 0.2 | 0 | 0.5 | 0.3 | 0.2 | 0.5 | 0.2 |
| **Drop rate** | 0.2 | 0 | 0 | 0.2 | 0 | 0 | 0 |
| **Learning rate** | 0.001 | 0.001 | 0.01 | 0.001 | 0.001 | 0.01 | 0.01 |
| **Batch size** | 120 | 40 | 40 | 300 | 300 | 40 | 300 |

### 4.5.5 Results

Here we present the results of seven pre-trained word embeddings that were used as input layers for two architectures BLSTM and CNN, as shown in Table 4.1.

#### 4.5.5.1 First Model: BLSTM

It can be observed from Table 4.9 that, for the first data set, most of the pre-trained word-embedding models caused the BLSTM to perform in a similar manner. The highest accuracy and $F_1$ score achieved by BLSTM, with Mazajak CBOW, were 93.8% and 89.7%, respectively. Mazajak CBOW also achieved the second-best recall and precision at 90.9% and 88.52% respectively. The highest precision was achieved by ArWordVec CBOW at 91.87%, while the highest recall was achieved by AraVec Skip-Gram at 88.85%. It is also noted that the performance of AraVec CBOW was the worst in terms of precision, $F_1$ score and accuracy with results at 87.09%, 86.66% and 91.9%, respectively.

Similarly, on the second data set, it is shown in Table 4.10 that Mazajak CBOW had the best precision and accuracy at 88.16% and 90.8%, respectively. Mazajak Skip-Gram performed similarly to Mazajak CBOW on the second data set and achieved the best recall at 74.6% and the best $F_1$ score at 75.2%. Mazajak Skip-Gram achieved the second-best accuracy at 90.7% as compared to 90.8% achieved by Mazajak CBOW, a difference of only 0.1%. Overall, it is noted from Table 4.9 that the best pre-trained word embedding model using BLSTM architectures for both data sets is Mazajak Skip-Gram as it is had the second best $F_1$ score on the first data set and the best $F_1$ score on the second data set. As explained in section 2.2, the $F_1$ score was used as a judgment metric as $F_1$ is more optimal for imbalanced data set [60].

**Table 4.9 Results of BLSTM using different pre-trained word embeddings on the first and second data sets.**

| | First Data set | | | | Second Data Set | | | |
|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **$F_1$ score** | **Accuracy** | **Precision** | **Recall** | **$F_1$ score** | **Accuracy** |
| **AraVec Skip-Gram** | 89.14 | **88.85** | 89 | 93.3 | 82.19 | 63.49 | 71.64 | 90.5 |
| **AraVec CBOW** | 87.09 | 86.23 | 86.66 | 91.9 | 79.35 | 65.08 | 71.51 | 90.27 |
| **Mazajak Skip-Gram** | 90.27 | 88.2 | <u>89.22</u> | <u>93.5</u> | 75.81 | **74.6** | **75.2** | <u>90.7</u> |
| **Mazajak CBOW** | <u>90.9</u> | <u>88.52</u> | **89.7** | **93.8** | **88.19** | 59.26 | 70.89 | **90.8** |
| **fastText** | 89 | 87.54 | 88.26 | 92.9 | <u>83.2</u> | 57.67 | 68.13 | 89.8 |
| **ArWordVec Skip-Gram** | 89.6 | 87.54 | 88.56 | 93.1 | 71.76 | 64.55 | 67.97 | 88.5 |
| **ArWordVec CBOW** | **91.87** | 85.25 | 88.44 | 93.2 | 76.92 | <u>68.78</u> | <u>72.63</u> | 90.2 |

**Note: Bold numbers indicate the best value while underlined numbers represent the second-best value.**

#### 4.5.5.2 Second model: CNN

In contrast, as shown in Table 4.10, for the first data set, AraVec Skip-Gram had the best CNN performance with an accuracy of 92.7%, and $F_1$ score of 88.01% and recall at 88.87%, as shown in Table 4.10. The best precision was achieved by ArWordVec CBOW at 89.67%. The second-best model performance was Mazajak Skip-Gram for recall, $F_1$ score and accuracy at 85.25%, 87.1% and 92.3%, respectively.

For the second data set, the best performance model was again AraVec Skip-Gram, with 71.96% for recall, 74.32% for $F_1$ score and 90.6% for the accuracy, while fastText achieved the best precision at 84.12% but had the worst recall and $F_1$ score at 56.08% and 67.3%, respectively.

When comparing pre-trained embedding models performance using the two architectures, the AraVec performance with either Skip-Gram or CBOW did not change significantly between the two architectures, while the other pre-trained word embeddings Mazajak and ArWordVec both decreased. This caused AraVec Skip-Gram to perform better using CNN architecture. Thus, to choose the overall best model for the CNN architecture, it was found that CNN with AraVec Skip-Gram performed the best in terms of the $F_1$ score on the first and the second data sets, as shown in Table 4.10. In addition, most of pre-trained word-embedding models performed better with BLSTM architecture; therefore, BLSTM generally appears to perform better when detecting Arabic-language health-related tweets in this study. This is particularly true with its best embedding (Mazajak Skip-Gram) for both first and second data sets.

**Table 4.10 Results of CNN model using different pre-trained word embeddings in the first and second data sets.**

| | First Data set | | | | Second Data set | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$ score | Accuracy | Precision | Recall | $F_1$ score | Accuracy |
| AraVec Skip-Gram | 88.16 | **87.87** | **88.01** | **92.7** | 76.84 | **71.96** | **74.32** | **90.6** |
| AraVec CBOW | 87.46 | 84.59 | 86 | 91.6 | 78.31 | 63.49 | 70.18 | 89.8 |
| Mazajak Skip-Gram | 89.04 | <u>85.25</u> | <u>87.1</u> | <u>92.3</u> | 78.26 | 66.67 | 72 | 90.2 |
| Mazajak CBOW | <u>89.44</u> | 83.28 | 86.25 | 91.9 | <u>81.05</u> | 65.61 | <u>72.51</u> | <u>90.6</u> |
| fastText | 87.46 | 84.59 | 86 | 91.6 | **84.12** | 56.08 | 67.3 | 89.7 |
| ArWordVec Skip-Gram | 85.76 | 82.95 | 84.33 | 90.6 | 70.56 | <u>67.2</u> | 68.83 | 88.5 |
| ArWordVec CBOW | **89.67** | 79.67 | 84.38 | 91 | 78.57 | 64.02 | 70.55 | 89.9 |

**Note: Bold numbers indicate the best value while underlined numbers represent the second-best value.**

## 4.6 Discussion

For the first experiment, which was concerned with pre-processing techniques, the best algorithm performance was achieved with 4 pre-processings out of a possible 26. Some of the popular techniques presented in Table 4.5 used by other researchers, such as normalizing *alef* and *different types of stemming* and removing numbers were not pre-processing methods that improved the accuracy of our final model in the first experiment.

In the literature, there was a focus on studying the impact of stemming on algorithm performance [146, 170, 172, 221]. Most studies found stemming increased the accuracy of the baseline model [146, 170, 221], and this study is in agreement with theses previous studies. Having said that, the best combination of the pre-processing techniques for our final model outperformed any combination of pre-processing that included any type of stemming, as shown in Appendix B.2. It also important to note that, out the four pre-processing techniques that were used in the final model, only one can be considered as not being an Arabic specific pre-processing technique, which is the removal of the repeated character, Normalizing the letters ي and ه are Arabic specific. Likewise, the fourth pre-processing technique removed Kashida, which is widely used by Arabic writers. This might suggest that in text classification for the Arabic language, Arabic specific normalization techniques might play a bigger role in improving the model performance compared to the other general pre-processing techniques. This possibility also highlights the importance of this study and the need for more studies to systemically assess the impact of normalizing Arabic specific techniques on the model performance of more data sets.

Nevertheless, we found that rarely used pre-processing techniques performed well in improving the classifier model. For example, lemmatization was only used in one study [186] in the literature reviewed in this paper. Yet, as it can be seen in Table 4.5, lemmatization performed well with all four classifier models. Notwithstanding, it was not one of the four techniques that improved the accuracy of the final best MNB model in the first experiment. It is also worth noting that whereas the MNB classifier achieved an 87.7 $F_1$ score on the first data set, its performance decreased on the second data set.

In the second experiment, we noted two observations. Firstly, there was no big difference between Mazajak Skip-Gram and Mazajak CBOW in their performance on the first data set using BLSTM. This also applied for Mazajak Skip-Gram and CBOW with BLSTM on the second data set. Furthermore, this also applied for Mazajak and ArWordVec using the CNN architecture. In contrast, there was a noted difference when we compared the performance of AraVec CBOW to Skip-Gram: AraVec Skip-Gram performed better than AraVec CBOW in both architectures. The second observation is the AraVec performance slightly decreased between the two architectures, whereas the Mazajak, ArWordVec and fasText had a more notable decrease. This caused AraVec to perform better using CNN architecture than other pre-trained word embedding models on the first data set. Furthermore, on the second data set using the CNN architecture, AraVec Skip-Gram performance had a negligible increase compared to BLSTM architecture.

When comparing deep-learning methods to traditional algorithms, the results for the first data set indicated that the BLSTM architecture with all pre-trained model embeddings performed better than the MNB classifier except for AraVec CBOW, where the MNB classifier performed better. When models using the CNN architecture, were compared to the MNB classifier, it is found that the MNB classifier performed better than most CNN classifiers, except for the CNN classifier that used AraVec Skip-Gram as an input layer, as is reported in Table 4.9 and Table 4.10. The CNN classifier that used AraVec as an input layer performed identically in terms of accuracy at 92.7% and only marginally different for $F_1$ score at 88.01% compared to 87.9%, where AraVec Skip-Gram performed better than the MNB classifier.

In the second data set, however, the CNN and BLSTM models both performed better with all the pre-trained word-embedding models than did the MNB classifier. The results suggest that the MNB classifier for the first data set is comparative to some deep learning methods, but all the deep learning methods outperformed the MNB classifier on the second data set, this data set representing more generalized, unseen data. This might contribute to answering the question in the literature that Guellil et al. posed [180]: “Are deep-learning approaches really more efficient than traditional approaches?”. The answer, as determined in this experiment seems to be “yes” with regard to generality.

**Limitations and strengths:**

Previous researchers concerned with evaluating pre-processing tasks considered only different types of stemming and stop words removal [57, 172]. Even in English, a recent study only compared 16 pre-processing techniques [213], here we have reviewed the literature and identified 14 variants of normalizing Arabic letters in addition to 12 pre-processing techniques that have been used for Arabic classification tasks. Future study should focus on investigating the impact of pre-processing on more than one data set. It should also focus on the impact of Arabic-specific normalization.

It should be noted that other newer deep learning techniques, such as autoencoders to learn features, or transformer-based language model such as BERT (Bidirectional Encoder Representations from Transformers) [264] and, for Arabic, AraBERT [265] may well outperform the models used here. However, the focus of this study is to use common deep learning architectures with pre-trained word embeddings and compare them with common traditional machine learning models frequently used in the related literature.

Another limitation is in Phase 3 in the first experiment (Figure 4.2). Although we used brute force in combing the pre-processing techniques and carried out 128 experiments for this task, we did not consider the order in which the pre-processing techniques were applied, which might have had an impact on the results. It is also possible that different combinations applied to different traditional approaches may have yielded more significant gains and thus led to other approaches overtaking MNB.

## 4.7 Conclusion

The goal of this paper was to evaluate the impact of pre-processing techniques on traditional algorithms, and we discovered that most of the techniques did not improve the accuracy of the baseline model. In addition, three out of four pre-processing techniques used in the final model for the first experiment are language specific. For the deep learning methods, we found that the BLSTM architecture performed better than the CNN architecture and the MNB classifier. BLSTM with Mazajak CBOW pre-trained word embedding performed the best on the first data set, while BLSTM with Mazajak Skip-Gram performed the best with unseen data. Overall, it was found that BLSTM with Mazajak Skip-Gram pre-trained word embedding was the best model with an $F_1$ score of 89.22% for the first data set and 75.2% for the second data set.

# Preface to Chapter 5

*This chapter was published in the Journal of Formative Research. which has a SNIP of 0.87. It is classified as a Q3 medicine and computer science application. Approximately 70% of the work in this paper was done by the author of this thesis and, as yet, has no citations.*

Automatic classification of the trustworthiness of health-related tweets is the next intuitive step after detecting health tweets. A key component in doing this is the development of appropriate data sets. Thus, this chapter describes the process of developing such a data set, where the accuracy of each tweet is labeled by two health experts. While other studies invite a third annotator to resolve disagreements, our approach is stricter in reducing uncertainty in the data set by excluding tweets on which there is a disagreement between the two annotators [266].

In this work, to develop a classifier, we attempted to capitalize on advances in ML made over the last two years. There have been significant improvements in pre-trained language models that yield good results in many NLP tasks. For example, the superiority of pre-trained transformer language models compared to other text classification methods is well documented, especially in the recent literature [31, 265, 267-270].

However, little is known about how pre-trained language models would perform in detecting the trustworthiness of health-related tweets. For example, previous works on the trustworthiness of health-related tweets in Arabic have investigated the performance of only traditional ML models [271].

The main goal of this chapter is the development of a model for detecting the trustworthiness of health-related tweets. Thus, the chapter examines the performance of pre-trained word embedding and compares it to the performance of pre-trained language models.

# 5 Pretrained Transformer Language Models vs Pretrained Word Embeddings for the Detection of Accurate Health Information on Arabic Social Media


## Abstract

### Background

In recent years, social media has become a major channel for health-related information in Saudi Arabia. While social media makes accurate health-related information easily accessible to many people, it has also become a channel for spreading health-related misinformation. Prior health informatics studies suggest that a large portion of health-related posts on social media are inaccurate. Given the subject matter and scale of such information, it is important to be able to automatically discriminate between accurate and inaccurate Arabic health-related posts.

### Objective

The first goal of this study is to generate a data set of generic health-related tweets in Arabic, labeled as either accurate or inaccurate health information. The second objective is to leverage this data set for training a state-of-the-art deep learning model for detecting the accuracy of Arabic health-related tweets. In particular, this study aims at training and comparing the performance of multiple deep-learning models that utilize pretrained word embeddings and transformer language models.

### Methods

We used 900 health-related tweets from a previously published data set extracted between the 15th of July and the 31st of August 2019. Also, we applied a pretrained model to extract an additional 900 health-related tweets from a second data set collected specifically for this study between the 1st of March and the 15th of April 2019. The 1800 tweets were labeled by two doctors as “accurate,” “inaccurate,” or “unsure”. The doctors agreed on 779 tweets, which were thus labeled as either “accurate” or “inaccurate”. Nine variations of pretrained transformer language models were then trained and validated on 623 tweets (80% of the data set) and tested on 156 tweets (20% of the data set). For comparison, we also trained a bidirectional long short-term memory (BLSTM) model with seven different pretrained word embedding as the input layer on the same data set. The models were compared in terms of their accuracy, precision, recall, $F_1$ score, and the macro average of the $F_1$ score.

**Results**

We constructed a data set of labeled tweets, 38% of which were labeled inaccurate health information, and 62% of which were labeled accurate health information. We suggest this is of high efficacy as we did not include any tweets where the doctor-annotators were unsure or in disagreement. Of the deep learning models investigated, the AraBERTv0.2-Large model was the most accurate with an $F_1$ score of 87%, followed by AraBERTv2-Large, and by AraBERTv0.2-Base.

**Conclusions**

Our results indicate that the pretrained language model AraBERTv0.2 is the best for classifying tweets as either inaccurate or accurate health information. Future studies should consider applying ensemble learning to combine the best models, since it may produce better results.

## Keywords

social media; machine learning; pretrained language models; BERT; deep learning; health information; infodemiology; tweets; language model; health informatics; misinformation

## 5.1 Introduction

### 5.1.1 Background

In the last two decades, there has been a dramatic increase in the number of people who utilize social media (SM) to participate in discussions on various topics, such as politics [272], health information [28], and education [273]. Regarding health-related information, several recent studies from Saudi Arabia have found that Twitter is the preferred SM platform for communicating and accessing medical information. For example, it was preferred by orthopedic surgeons to reply to (personal and professional) medical questions [274]; by dental practitioners for medical consultations [26]; by diabetic patients searching for health information [27]; by female students at a university in Saudi Arabia for reading about systemic lupus erythematosus [10] and by adolescents searching for oral health information [28].

A significant problem in this form of communication is that there is no quality control over the medium, and the majority of the health information presented on Twitter seems inaccurate, as illustrated by the various studies summarized in Table 5.1. Indeed, multiple data science studies utilize data sets of health-related communication on SM to study this phenomenon and other studies [63, 271, 275] go further to design frameworks to detect the accuracy of health information on SM.

Previous studies focus on specific health issues and sometimes on very specific types of rumors [17, 62, 84, 271, 276]. This suggests the need for a more general framework that can detect the accuracy of health information across known and previously unknown health conditions, for example, during an outbreak of a previously unknown infectious disease.

Given the prevalent use of Twitter for health information spread in Saudi Arabia [7, 10, 26-28, 274, 277], we aim to inform the development of a new and more generic framework that is not bound to a specific disease/rumor type, to detect the accuracy of a broad-base of Arabic health-related tweets.

### 5.1.2 Related work

In this section, we review methods used to label health-related tweets as either accurate or inaccurate in order to create labelled data sets. We also review previously proposed machine learning (ML) models, including deep learning (DL), for detecting the accuracy of health-related tweets.

#### 5.1.2.1 Methods used to label health-related tweets

Studies that address the accuracy of health-related tweets can be classified into three groups. The first group comprises studies that label health-related tweets according to the information they contain, regardless of the source of the information. The second group is composed of studies that rely on external (fact-checking or very reputable) websites. The last group is studies that rely on various characteristics of the tweets or on the source of the information only to judge the accuracy of the tweets.

Regarding the concept of accuracy and misinformation, Chou et al. [278], define *misinformation* as information that lacks scientific evidence. A more precise definition can be found in [279], where the authors define inaccurate information or misinformation as "explicitly false," according to what would be deemed incorrect by expert consensus. In [280], the authors combine these definitions to describe misinformation or inaccurate health information as information that is not supported by clear evidence and expert opinion.

Studies relying on the opinions of experts seem to indirectly/directly employ these definitions to assess accuracy, although it should be noted that, while misinformation is inaccurate, it is not necessarily intended to be so. In contrast, disinformation is information that is intentionally deceptive [281]. Examples of ‘opinions of experts’ studies are included in Table 5.1 ([12, 271, 275, 282]). They involve labeling health-related tweets based on the opinion of a health expert. The tweets were labeled as either inaccurate or accurate by at least two experts. A third expert is typically involved when there is a disagreement between the original two experts: this expert casts the deciding vote for controversial tweets.

Vraga and Bode [266] criticized the above definition of misinformation, raising the point that there are many issues on which experts do not agree. However, they state that as long as there is more evidence that supports the information, the agreement rate between the experts increases. Taking a stricter approach, Albalawi et al. [161] and Sharma et al. [283] excluded tweets on which experts disagreed in an attempt to exclude uncertainty from their data sets. Table 5.1 summarizes the aforementioned studies.

Unsurprisingly, studies that rely on expert opinion use relatively small data sets (ranging from 109 tweets to 625 tweets) compared to studies that employ other labeling methods (see Table 5.1). Even those that use non-experts but use manual coding (performed by non-expert annotators) tend to work on a small sample of the data set [63, 284].

The second group is composed of studies that rely on an external website, such as a fact-checking website, to label the tweets. One such example is the study by Elhaddad et al. [62], which relies on a fact-checking website to identify misleading information. A similar method is employed by Ghenai et al. [84], who rely on the website of the World Health Organization (WHO) to identify six rumors. From these rumors they derived keywords to extract tweets relevant to these rumors. The drawback of this method is that only tweets that were relevant to specific rumors are extracted and thus, the model was trained only on this limited number of rumors. Furthermore, these methods are highly language restricted:

Both studies referred to in Table 5.1 are performed in English, as mandated by the WHO website and the fact-checking website.

Other methods rely on various characteristics of the tweets or on the source of the information only, without judging the actual information. For example, in [61], the authors identified tweets as credible if they included hashtags that indicated that they originated from noted agencies or other reliable sources, and tweets are identified as speculative if they include hashtags that imply the increase of fear, rumors or scams.

Similarly, the authors in [63, 74, 276] defined credible tweets as tweets that have information from a confirmed, reliable source, such as the WHO, the Centers for Disease Control (CDC) or another official health agency. This method differs from the method used by the second group mentioned above because it first identifies a tweet and then examines its source. By contrast, the methods in the second group first identify a trustworthy website and then use the information the website for identifying tweets of interest.

More generally, Yin et al. [85] stated that a website is *trustworthy* if it provides correct information and suggests that information is likely to be true if it is provided by a trustworthy website. Studies that rely on trustworthy websites to identify rumors [63, 84, 276] seem to follow this definition, even if they do not explicitly state it.

It should be noted that, based on the data in Table 5.1, all Arabic studies that rely only on expert opinion [12, 161, 271] are of a small scale and qualitative, so it would be impossible to scale those up. Notably, the percentage of inaccurate tweets for English studies that relies on expert opinions is in the range 10%–25%, while the corresponding range for Arabic is 31%–50%. This suggests a greater occurrence of inaccurate health-related tweets in the Arabic language than in English.

**Table 5.1. Summary of studies that analyze the accuracy of health information on SM**

| | Number of tweets/ docs | Sources | Methods to label | Language covered | Percentage of the accuracy | Topics covered | Type of study |
|---|---|---|---|---|---|---|---|
| [282] | 358 | Twitter | Expert votes; relabeling in cases of disagreement. | English | 25.4% inaccurate | COVID-19 | Exploratory |
| [161] | 109 | Twitter | Two doctors; delete if there is a disagreement. | Arabic | 31% inaccurate | General | Quantitative pilot study |
| [271] | 208 | Twitter | Expert votes; relabeling in cases of disagreement. | Arabic | 38% inaccurate | Cancer | ML |
| [283] | 183 | Facebook | Two doctors; delete if there is a disagreement. | English | 12% inaccurate | Zika | Quantitative |
| [12] | 625 | Twitter | Vote if the experts did not agree. | Arabic | 50% inaccurate | Only tweets from health professionals | Quantitative and exploratory study |
| [275] | 5000 | Health forum | Annotator voting. Also, consulted an expert to validate information labelled as misleading | Chinese | 11.4% misinformation | Autism | ML |
| [285] | 2,460 | Twitter | Coders. Check the inter-agreement on 200 tweets | English | 10% inaccurate | Ebola | Quantitative |
| [74] | 5,395 | Twitter | Coder: Check agreement on 125 tweets; unsubstantiated by the following reference standards: the CDC and Public Health Agency of Canada for scientific claims and a panel of credible online news sources (e.g., CNN, BBC) for news-related claims. | English | 4.5% inaccurate | H1N1 | Exploratory |
| [63] | 800 | Twitter | Annotator's agreement; relabeling in cases of disagreement. Here, the definition for misinformation is 'news items without a source'. | English | Unknown | Zika | ML |
| [61] | 47M | Twitter | Type of hashtags. | English | 25% of the analyzed tweets were speculative | Ebola | Quantitative |
| [276] | 409,484 | Twitter; keywords | Although they use coders, their definition for a rumor includes lack of a source, so unconfirmed information will automatically be classified as uncredible. Also, tweets classified only by one coder. They checked inter-agreement on 20 tweets. | Not noted, but the keywords are English | 70% uncredible | Covid-19 | ML |
| [62] | 7,486 | Various websites | Fact-checking websites and official websites | English | 21% | Covid-19 | ML |
| [284] | 500 | Instagram | Coders' agreement | English | 23% | Zika | Exploratory |
| [84] | 26,728 | Twitter | Define keywords to extracted tweets based on rumors identified from WHO website then Coders labelled tweets | English | 32% | Zika | ML |

#### 5.1.2.2 Machine learning approaches

Out of the 14 studies reported in Table 5.1 that analyze the accuracy of health-related tweets in general, 6 proceed to train an ML model to detect the accuracy of health information, as shown in Table 5.2.

Studies that report on training ML models include Elhadad et al. [62] and Al-Rakhami et al. [276], who used ensemble learning on an English data set. Elhadad et al. [62] used ensemble learning that involves multiple DL architectures, and Al-Rakhami [276] train ensemble models consisting of traditional ML algorithms, such as support vector machine (SVM) and Random Forest (RF). Another similarity between these studies is the method used to identify misleading information. Elhadad et al. [62] build their data set by extracting ground truth data and rumors from a fact-checking website. Al-Rakhami et al. [276] considered tweets credible if they have a reliable source and misleading otherwise. Both models report a high level of accuracy (more than 97%), as shown in Table 5.2.

From Table 5.1 and Table 5.2, it is clear that studies that rely on a fact-checking website [62, 84] and studies that determine the accuracy of a tweet based on its source [276] obtain high level of accuracy, possibly because, these models are trained on relatively large data sets. For example, Al-Rakhami [276] train their model on 409,484 tweets. But the automated labelling leaves open the possibility of incorrect labeling and all these studies are done on English.

**Table 5.2. Summary of studies that developed ML models to detect the accuracy of health-related information**

| Study | ML approach | Results | Labelling type |
|---|---|---|---|
| Elhadad et al., [62] | Deep learning multi model GRU LSTM CNN | 99.99% $F_1$ | Ground truth data from websites |
| Ghenai et al. [84] | Random forest | 94.5 % Weighted average for $F_1$ score | Crowdsource agreement but key words based on 4 WHO-website identified rumors |
| Al-Rakhami et al. [276] | Ensemble learning, random forest + SVM | 97.8% accuracy | Single annotator only after confirming source |
| Zhao et al. [275] | Random forest | 84.4% $F_1$ score | Annotator vote; Also, consulted an expert to validate misleading information |
| Sicilia et al. [63] | Random forest | 69.9% $F_1$ score | Agreement of a health expert |
| Saeed et al. [271] | Random forest | 83.5% accuracy | Agreement of a health expert |

Most of the studies that develop ML models focus on outbreaks (four out of six studies). Studies that develop ML models for non-outbreak conditions [271, 275] obtain less accurate results compared to outbreak conditions. This might be because these non-outbreak condition models are trained on a limited number of documents compared to the outbreak models. We also find that the level of accuracy obtained for non-outbreak data sets were about 84% (see Table 5.2). It is also notable that all of these studies train an RF model.

Table 5.2 (and our associated literature review) suggests that recent advancements in DL have not been applied to the detection of misleading Arabic health information sufficiently. With our previous work we have shown that DL architectures utilizing word embedding as an input layer outperform other traditional ML models, such as SVM and Naive Bayes in the detection of Arabic health-related information on SM [286], but now we move past that to classification of Arabic health-related tweets based on their accuracy.

Word embedding is a learned representation for words in natural language processing (NLP) [175]. Words with similar meanings usually have similar numbers in their vectors. The closer the words are in meaning, the shorter is the distance between the two vectors representing them. One of the main criticisms of the word embedding approach is that it is considered to be contextual free, i.e., the embedding of a word is not affected by its position in the sentence [287]. Hence, it is also called static word embedding. However, in practice, the meaning of a word may depend on the position it has in a sentence.

In recent years, pretrained language models have been proven to work well for many NLP tasks, including entity recognition, language translation, and text classification [288]. Unlike static word embedding techniques, such as Skip-Gram and CBOW, lanugage models can learn the context of the words and thus, can assign different values for the words depending on their context [287]. There are different types of lanugage models, including Contextual word Vectors (CoVe) and the embeddings from language model (ELMo) [287]. One of the most popular langauge models models is Bidirectional Encoder Representations from Transformers (BERT), which has been proven to perform well on text classification tasks.

The superiority of the transformer models compared to other text classification methods is well documented, especially in recent literature. Multiple studies compare transformer models to other DL models [31, 267-269, 289], and the results show that transformers outperform otheir ML models, including different DL architectures and traditional ML models, such as SVMs and RF. This indicates the potential capability of transformers to better detect the accuracy of Arabic health information on SM.

Therefore, in this study, we aim at contributing to this field by developing a data set of certified accurate/inaccurate Arabic health-related tweets and investigating the ability of the BERT model/pretrained word embedding models to detect the accuracy of Arabic health-related tweets across a wide range of health-related issues.

## 5.2 Methodology

The empirical method is composed of two parts. The first part addresses the extraction of health-related tweets with a model proposed in our previous study [286]. In that study we utilized a health lexion that focused more on general health keywords rather than specific outbreaks, as a recent study suggests that general health msinformation is more likely to spread when compared to COVID-19, for example [290]. In contrast, Table 5.1 illustrates that most of the studies in this area focus on a specific domain or disease outbreak.

The extracted health-related tweets are labelled by health experts as either accurate or inaccurate. Figure 5.1 presents an overview of this part of the study.

In the second part, we propose two types of trustworthiness-detecting models to automatically classify health-related tweets as either accurate or inaccurate, and evaluate them: BLSTM DL models, and pretrained transformer language models.

### 5.2.1 Building data sets of trustworthy health-related tweets

In this study, we utilize two data sets containing health-related tweets. The first data set is the result of our previous work [286].

The first data set was extracted from 297,928 tweets posted between the 15th of July and the 31st of August 2019. 5,000 of them were randomly sampled and labelled by two annotators as either "health-related" or "not health-related". A third annotator resolved the disagreements between the two annotators.

The first data set was extracted during the summer holidays in Saudi Arabia for 45 connective days. To assess generality, we extracted the second data set for a different timeframe: it was during Hajj AND Eid al Adha (Muslim holy days) and during school days between 1st of March and 15th of April 2019. The second set of 900 tweets using the ML methodology proposed in that same study [286], because the availability of health professionals was constrained by the ongoing COVID-19 pandemic and because the ML model derived in that study achieved a high-quality result (= 93% accuracy ).

The methodology proposed in [286], consists of extracting tweets from a set of collected tweets with the help of a health lexicon, and then further filtering out tweets not related to health with the help of a ML model. Based on the health lexicon, 217,702 tweets were extracted. We sampled 5000 tweets and applied the ML model to extract 900 health-related tweets.

Finally, we added the 900 tweets from the second data set to 900 tweets sampled from the first data set and had those 1800 tweets labeled as either accurate or inaccurate heath information by two medical doctors.

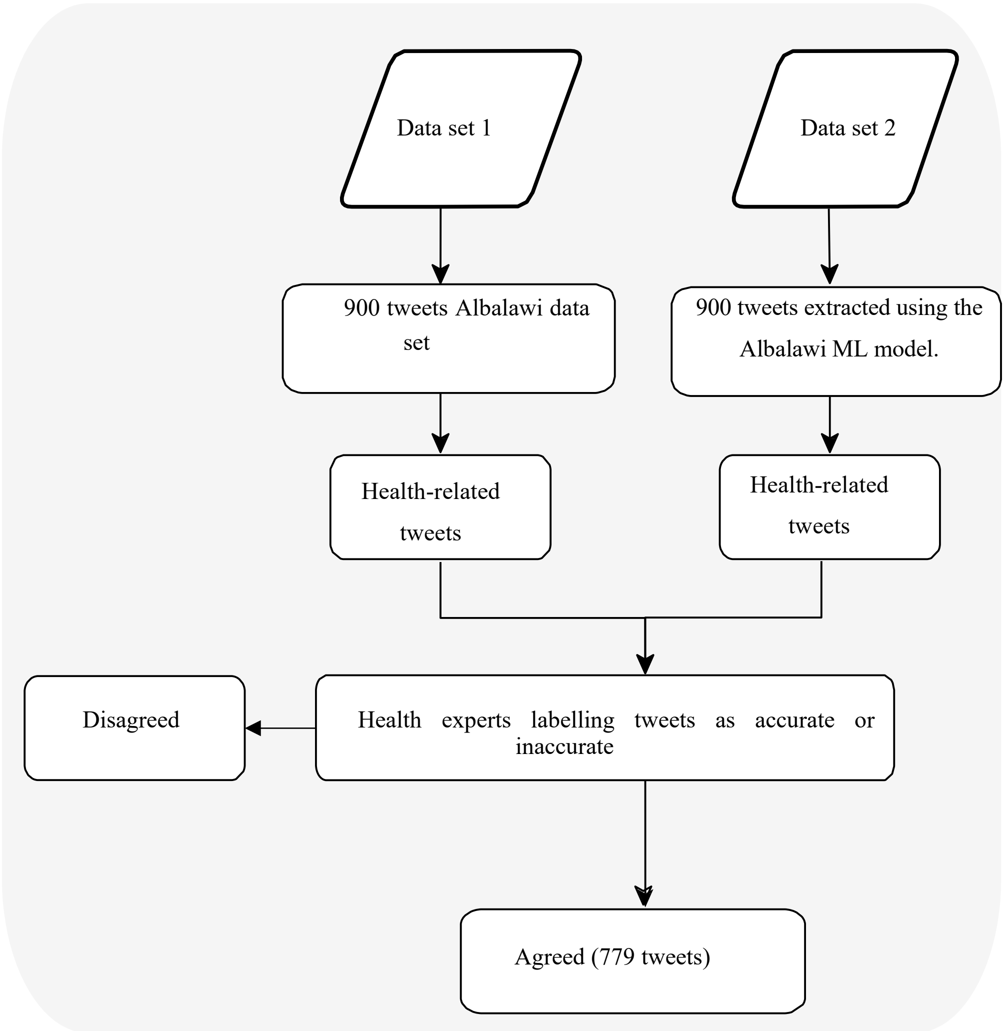


**Figure 5.1 Overview of the process followed in labelling tweets as either accurate or inaccurate.**

### 5.2.2 Labeling accurate/inaccurate tweets

They were asked to manually label each of the 1800 health-related tweets into one of the following categories:

- Accurate health information.
- Inaccurate health information.
- Not sure about the accuracy.

That is, we follow the protocol of relying on the opinions of experts to define the accuracy of the information collected. Taking into account the points made by Vraga and Bode [266], every tweet is assessed by two experts and a tweet is included in the final data set for this study only if both experts agree on its accuracy; that is, we reduce uncertainty by excluding information that is not sanctioned by all experts (Indeed, later we will show that between-doctor reliability in this coding was limited, buttressing Vraga and Bode's stated need for increased certainty when using human classification). The 'not sure' option was offered to the doctors to avoid forcing them to evaluate the tweets if they do not have enough relevant health knowledge to accurately evaluate them, or if the tweets are ambiguous.

While other studies invite a third annotator to resolve disagreements, our approach is stricter in reducing uncertainty in the data set by excluding tweets on which there is a disagreement between the two annotators. Out of 1800 tweets, the two doctors agreed on 779, which were labelled as containing either accurate or inaccurate health information. The doctors disagreed on 163 tweets. The remaining 858 tweets were labelled as "unsure" by at least by one of the doctors. We dropped the tweets on which at least one of the doctors was unsure and used the remaining 779 tweets in our experiments.

While 779 tweets constitute a relatively small data set, most of the data sets constructed in literature based on agreements between health experts were relatively small. As shown in Table 5.1, the highest number of health-related tweets judged by health experts in other studies was 625, in the study of Alnemar et al [12].

These 779 tweets labelled as either accurate or inaccurate can be found in Multimedia Appendix 1. Please note that we only share tweet IDs and the labels as the Twitter policy prevents the content of tweets being redistributed. These tweet IDs, can be used to obtain the text of the tweets with the Twitter API [291].

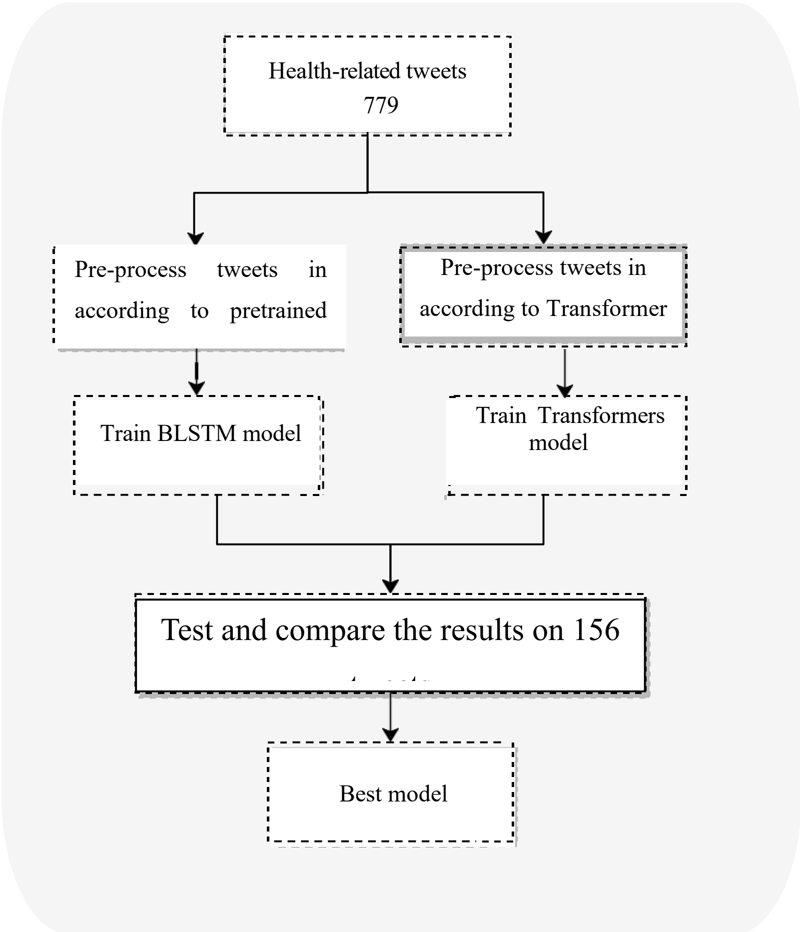


**Figure 5.2 Overview of the process used to train and select ML**

### 5.2.3 Deep learning models considered

After completing the annotation of the health-related tweets as either accurate or inaccurate, we trained 16 classification models, 7 of which used a BLSTM architecture with pretrained word embeddings as their input layers, and 9 of which utilized a pretrained transformer language model. Figure 5.2 illustrates the steps implemented during this stage. We provide further details below.

#### 5.2.3.1 The BLSTM architecture

For seven of the 16 models we trained, we utilized a BLSTM architecture with pretrained word embeddings as the input layer. LSTM is a type of a recurrent neural network that takes advantage of dependencies between parts of the input sequence, and it can learn these dependencies. LSTM also preserves the information of past input. The BLSTM variation differs from LSTM by its ability to learn the dependencies between past and future elements [292]. BLSTM have been found to perform well for many NLP tasks, including text classification [258]. The BLSTM model begins with input and embedding layers to which a dropout layer is added, followed by the BLSTM layer with another added dropout layer [286]. BLSTM has been shown to work better than traditional ML models (Support Vector Machine, Naive Bayes, K-Nearest Neighbors, and Logistic Regression) and conventional neural networks in a previous study on detecting Arabic health-related tweets [286]. For the input layer, we use seven pretrained word embedding models for Arabic [152-154, 183]. Please note, each of AraVec, Mazajak and ArWordVec come in two variations: CBOW and Skip-Gram and the final one is BLSTM fastText.

#### 5.2.3.2 Transformers models

BERT is a transformer language model that has shown superiority in many NLP tasks. There are different Arabic pretrained language models, based on transformers that have been developed recently by the Arabic NLP community. Most of these pretrained language models built on top of the BERT-base model. Some of them provide a version based on BERT-large too.

The difference between BERT-base and BERT-large is that BERT-base uses 12-layers, 768-hidden layers, 12-heads and approximately 136 million parameters, whereas the BERT-large model has 24-layers, 1024-hidden layers, 16-heads, and approximately 370 million parameters [264]. All models may not leverage BERT-large because it is more difficult to train and comes with a higher computational cost compared to BERT-base [293].

Examples of pretrained language representation models for Arabic that offer both base and large variants are ArabicBERT [294] and AraBERT [265]. AraBERT is considered the first Arabic-specific transformer language model introduced in 2020 by Antoun et al. [265]. In 2021, an updated version of AraBERT was released [295]. AraBERT is considered one of the best transformer language models for NLP, outperforming other models in Arabic sentiment analysis [296]. AraBERTv2 pre-processes the text using Farasa segmentation. Farasa segmentation involves breaking the words based on prefix and suffix [297], while AraBERTv0.2 pre-process the text without using Farasa segmentation. In this work, we experiment with these six models: AraBERTv2, AraBERTv0.2 and ArabicBERT in both variants of BERT: base and large.

In addition to these six models, we also investigate three other state-of-the-art pretrained language models which are QARiB [298], MARBERT, and ARBERT [299], which are based only on BERT-base. These models reportedly perform well on text classification tasks [294, 298-300]. Table 5.3 summarizes the characteristics of the pretrained language models used in this study.

**Table 5.3 Pretrained language models.**

| Name | Based on | Size | Corpus |
|---|---|---|---|
| ARBERT [299] | Bert-base | 61 GB of MSA text (6.5B tokens) | Books, News (news articles, Wikipedia articles) |
| MARBERT [299] | Bert-base | 128 GB of text (15.6 B tokens | 1B Arabic tweets |
| QARiB [298] | Bert-base | 14 B tokens<br>Vocabulary: 64 k | 420 million tweets and ~ 180 million sentences of text Arabic Giga Word, Abulkhair Arabic Corpus and OPUS. |
| ArabicBERT [294] | Bert-base and Bert-large | 95 GB of text,<br>8.2 B words | Arabic Oscar version, Wikipedia and other resources |
| AraBERTv0.2 [295] | Bert-base and Bert-large | 77 GB or 200,095,961 lines or 8,655,948,860 words or 82,232,988,358 chars | OSCAR unshuffled and filtered.<br>Arabic Wikipedia articles<br>The 1.5 B words Arabic Corpus<br>The OSIAN Corpus<br>Assafir news articles. |
| AraBERTv2 [295] | Bert-base and Bert-large | 77 GB or 200,095,961 lines or 8,655,948,860 words or 82,232,988,358 chars | OSCAR unshuffled and filtered.<br>Arabic Wikipedia articles<br>The 1.5B words Arabic corpus<br>The OSIAN corpus<br>Assafir news articles. |

### 5.2.4 Evaluation metrics

$F_1$ score, recall, precision, and macro average (macro avg) of $F_1$ are used to evaluate the ML models, as detailed in Table 5.4. The macro averaged $F_1$ is the averaged $F_1$ across all classes which are accurate and inaccurate health-related tweets [301].

**Table 5.4 Metrics used to evaluate the ML models.**

| Metric | Formula |
|---|---|
| Recall | True Positives/(True Positives + False Negatives) |
| Precision | True Positives/(True Positives + False Positives) |
| $F_1$ | (2 * Precision * Recall)/(Precision + Recall) |
| Macro averaged $F_1$ | $\frac{1}{N} \times \sum_{i=1}^{n} F_1(i)$ where $N$ is the number of classes. |

### 5.2.5 Pre-processing data

In this study, the text is pre-processed following the procedure outlined by the authors of the corresponding pretrained word embeddings models. Li et al. found that this is the best text pre-processing practice when working with pretrained word embeddings [178]. Similarly, for all pretrained word embedding models [152-154, 183] and pretrained language models [294, 295, 298, 299], we followed the same steps provided by the original studies.

We split the data set into a training, validation and test data sets in the ratio 507 tweets (65%) for training, 116 tweets (15%) for validating the model and 156 tweets (20%) for testing.

## 5.3 Results

The kappa coefficient for all categories was 0.377, which is in fair agreement, according to Cohen [52]. However, the benchmark scale proposed by Fleiss et al. [53] to evaluate agreement indicates that such a coefficient is poor (< .40, poor; .40 to .75, intermediate–good; > .75, excellent). Given the low kappa coefficients across the three categories, we considered only cases where both doctors were explicitly in agreement, as they were on 779 tweets from the original data sets.

Of the 1021 tweets that were excluded, 874 (48.6%) were labeled “not sure” by at least one doctor, and in the cases of 147 tweets, the doctors disagreed regarding the accuracy of the tweets.

Out of the 779 tweets doctors agreed on in our data set, 296 (or approximately 38%) were labeled as inaccurate, as shown in Table 5.5. This is similar to the inaccuracies reported in other studies, as shown in Table 5.1.

**Table 5.5 Distribution of trustworthy and untrustworthy health-related tweets.**

| | Number of tweets | Presenting |
|---|---|---|
| Accurate | 483 | 62% |
| Inaccurate | 296 | 38% |
| Total | 779 | 100% |

Table 5.6 presents examples of accurate and inaccurate health-related tweets. As can be seen from the tweets in the table, they cover a wide range of topics including, but not limited to, psychology and cancer. Interestingly, in the third accurate-tweet example, the difficulty for non-experts in discerning accurate from inaccurate health information is illustrated, as advice against taking anti-diarrhea drugs in the event of food poisoning is slightly counter intuitive.

Some tweets claimed the benefits of some traditional foods and spices. For example, some tweets promoted Zamzam (holy water for Muslims), claiming there was scientific research that stated that it could strengthen the human immune system; experts classified the information as inaccurate.

Also, the examples of accurate tweets presented here suggest that accurate health-related tweets tend to be more preventive in nature, a finding supported by a wider sampling of accurate tweets. As shown in Table 6, the accurate tweets gave advice to stop eating too much red meat as it causes gout or increases cholesterol; to stop smoking to prevent lung cancer; and to stop taking anti-inflammatory drugs in the event of food poisoning. In contrast, and as note earlier, inaccurate tweets were promoting natural and alternative medicine such as curbing eating and drinking Zamzam water for their health benefits. An interesting example was in relation to cancer, where accurate tweets advised readers to stop smoking but some of the inaccurate tweets were also preventative: they advised taking a spoonful of cinnamon to prevent all types of cancer.

**Table 5.6 Examples of inaccurate and accurate health-related tweets**

| | **Accurate** | **Inaccurate** |
|---|---|---|
| 1 | Tomorrow enjoys the feast. and get closer to God with your sacrifice<br>And eat but do not extravagant and feed the contented and be merciful as God has commanded you<br>Eating too much red meat might:<br>Raise the level of triglycerides<br>Raise cholesterol<br>Increase uric salt in the blood Increases gout attacks in the joints | Scientific research,<br>The research says that Zamzam water bears the name (water), but it differs radically from water compounds, as all the waters of the world belong to the acidic compound, except for (Zamzam water).<br>It is (alkaline!) Glory be to God. There is no other alkaline water on the face of the earth. So, when you drink it in abundance, the human body has a strong immunity against viruses!! |
| 2 | Symptoms of social phobia<br>Sometimes, social phobia can be accompanied by physical signs and symptoms, which may include:<br>Flashness<br>Rapid heart palpitations<br>Shivering and sweating<br>Upset stomach or nausea<br>Difficulty catching breath<br>Dizziness or lightheadedness<br>Feeling like your mind has gone blank<br>Muscle tension | When Western scholars searched for the causes of mental illness, they found only two reasons (fear and sadness) fear of the future and sadness of the past, both of which are the opposite of happiness. |
| 3 | In the event of food poisoning, please take care not to use anti-diarrheal medicines, as they may worsen the condition | Did you know that a 5-minute tantrum is so stressful that it weakens the immune system for more than 6 hours |
| 4 | Hemoglobin is a group of proteins in red blood cells whose function is to transport oxygen from the lungs to the body, return carbon dioxide from the body, and transport it to the lungs and get rid of it through breathing.<br>Iron is an important element and enters the composition of hemoglobin, so if iron deficiency, hemoglobin decreases, and anemia occurs. | Cupping helps smokers to quit smoking or reduce the negative impact on the body through:<br>Removing excess hemoglobin from the body by excreting aging red blood cells, and thus the disappearance of the pathological symptoms of high hemoglobin caused by smoking |
| 5 | Among the ways to prevent lung cancer:<br>Stay away from smoking<br>Avoid passive smoking<br>Avoid carcinogenic and radioactive materials. | Just a spoonful of cinnamon daily:<br>Rich in anti-inflammatory and antioxidants<br>Prevents all types of cancer<br>Prevents heart disease<br>Anti-diabetes |

*These tweets are translated from Arabic to English. For privacy and ethical issues, original tweets cannot be shared.

In terms of the comparison of models, we observed that overall, BERT models perform better than BLSTM models, based on $F_1$ score for both classes. Overall, AraBERTv0.2 Large performs better than all other models. Specifically, the best model was AraBERTv0.2-Large (macro $F_1$ score = 87%), followed by AraBERTv2-Large (macro $F_1$ score = 86%), and by AraBERTv0.2-Base (macro $F_1$ score = 85%), as shown in Table 5.7. These findings hide larger, but still small, variations in individual techniques precision and recall scores, for inaccurate and accurate tweets. For example, while AraBERTv0.2 Base achieved a recall of 78% for inaccurate tweets, AraBERTv0.2 Large achieved a recall of over 83%.

The results also suggest that, in general, BERT-Large models tended to be better at detecting inaccurate tweets than the BERT-Base models. The AraBERTv.2, AraBERTv0.2, and ArabicBERT Large models performed better than their Base versions at detecting inaccurate health tweets, as Table 5.7 shows. In contrast, the BERT-Base models might be better at detecting accurate tweets, except for the AraBERTv2, whose Large and Base versions performed similarly.

Of the pretrained word embeddings, the results in Table 5.7 show that Mazajak Skip-Gram is the best based on $F_1$ score.

**Table 5.7 Comparison of the performance of ML models for detecting the accuracy of health-related tweets.**

| | Class | Precision | Recall | $F_1$ score | Macro avg |
|---|---|---|---|---|---|
| **AraBERTv2 Base** | Inaccurate | 0.804 | 0.7627 | 0.7826 | 0.8279 |
| | Accurate | 0.86 | 0.8866 | 0.8731 | |
| **AraBERTv2 Large** | Inaccurate | 0.8276 | 0.8136 | 0.8205 | 0.8564 |
| | Accurate | 0.8878 | 0.8969 | 0.8923 | |
| **AraBERTv0.2 Base** | Inaccurate | 0.8519 | 0.7797 | 0.8142 | 0.8543 |
| | Accurate | 0.8725 | 0.9175 | 0.8945 | |
| **AraBERTv0.2 Large** | Inaccurate | 0.8448 | **0.8305** | **0.8376** | **0.8701** |
| | Accurate | **0.898** | 0.9072 | **0.9025** | |
| **MARBERT** | Inaccurate | 0.7759 | 0.7627 | 0.7692 | 0.8154 |
| | Accurate | 0.8571 | 0.866 | 0.8615 | |
| **ARBERT** | Inaccurate | 0.7903 | **0.8305** | 0.8099 | 0.8447 |
| | Accurate | 0.8936 | 0.866 | 0.8796 | |
| **QARiB** | Inaccurate | 0.7797 | 0.7797 | 0.7797 | 0.8228 |
| | Accurate | 0.866 | 0.866 | 0.866 | |
| **ArabicBERT Large** | Inaccurate | 0. 8654 | 0.7627 | 0.8108 | 0.8532 |
| | Accurate | 0.8654 | 0.9278 | 0.8955 | |
| **ArabicBERT Base** | Inaccurate | **0.8913** | 0.6949 | 0.781 | 0.83492 |
| | Accurate | 0.8364 | **0.9485** | 0.8889 | |
| **BLSTM Mazajak CBOW** | Inaccurate | 0.7719 | 0.7458 | 0.7586 | 0.8079 |
| | Accurate | 0.8485 | 0.866 | 0.8571 | |
| **BLSTM Mazajak Skip-Gram** | Inaccurate | 0.8542 | 0.6949 | 0.7664 | 0.8222 |
| | Accurate | 0.8333 | 0.9278 | 0.8780 | |
| **BLSTM ArWordVec Skip-Gram** | Inaccurate | 0.8261 | 0.6441 | 0.7238 | 0.7919 |
| | Accurate | 0.8091 | 0.9175 | 0.8148 | |
| **BLSTM ArWordVec CBOW** | Inaccurate | 0.7925 | 0.7119 | 0.75 | 0.805 |
| | Accurate | 0.835 | 0.8866 | 0.86 | |
| **BLSTM AraVec CBOW** | Inaccurate | 0.6865 | 0.7797 | 0.7302 | 0.7737 |
| | Accurate | 0.8571 | 0.866 | 0.8172 | |
| **BLSTM AraVec Skip-Gram** | Inaccurate | 0.7313 | **0.8305** | 0.7777 | 0.8136 |
| | Accurate | 0.8144 | 0.8144 | 0.8494 | |
| **BLSTM fastText** | Inaccurate | 0.8158 | 0.5254 | 0.6392 | 0.7382 |
| | Accurate | 0.7627 | 0.9278 | 0.8372 | |

**Note: Bold numbers indicate the best value, while underlined numbers represent the second-best value**

## 5.4 Discussion

As noted earlier, the examples given in the Results section showed that the accurate tweets were more focused on preventive medicine while the inaccurate tweets were more focused on alternative and natural medicine. It could be argued, however, that this is due to the keywords used in extracting and filtering the tweets or due to the selected tweet examples. Nevertheless, a previous study mentioned that the prevalence of natural alternatives and alternative medicine compared to medicine provided by the healthcare system [87] may be harmful. To illustrate the importance of this with respect to specific patients, there was a reported case of a cancer patient who took alternative medicine promoted on SM, which caused the hospital to temporarily stop her cancer treatment to repair the damage caused by that medicine.[302]. At a more general level, insights like these could provide additional levers with which to detect inaccurate health tweets going forward.

The results of BLSTM with pretrained word embedding models (AraVec, Skip-Gram, and Mazajak) are comparable to the results of some BERT models—that is, MARBERT, QARiB, and ArabicBERT Large. Indeed this has been previously reported in the literature where MARBERT and QARiB have outperformed some of the other transformer models, for example, ArabicBERT and AraBERT [298, 299]. Again, a takeaway from here is that pretrained word embeddings might outperform pretrained BERT models in this first comparative study directed at Arabic. There is no guaranteed best model between pretrained word embeddings and pretrained transformer models for this language.

But, in general, the results showed the superiority of the BERT models over BLSTM with pretrained word embedding models. Overall, 19 best or second-best results were obtained by the nine BERT-based approaches, while only three best or second-best results were obtained by the seven pretrained word embedding models.

Most models performed better at detecting accurate health tweets than inaccurate ones. The detection rate (recall) for accurate tweets ranged from 0.9485 to 0.8144. This means that the majority of the models missed only about 5% to 19% of the accurate tweets, a result that is promising. In contrast, the detection rate for the inaccurate tweets was lower and had a wider range, from 0.8305 to 0.5254, implying that the best models missed up to 17% of the inaccurate tweets. That is concerning because we would like to successfully identify all inaccurate tweets, and yet even the best model missed 17% of them.

The flip side of this is precision: how many accurate/inaccurate tweets identified by the technique are actually accurate/inaccurate. In terms of the inaccurate tweets, the approaches range from 0.7759 to 0.8913: quite a large span and one which means that, if the wrong technique is chosen, approximately a quarter of the tweets identified as inaccurate are incorrectly classified. Probably more of concern is the number of tweets identified as accurate that are not. Similarly, here, the span is from 0.8913 to 0.7627, again implying that, if the wrong technique is chosen, this could be problematic.

Some models that had high rates of detection for accurate health tweets could have low detection rates for inaccurate ones. For example, the ArabicBERT-Base and BLSTM fastText models were the best and second-best, respectively, at detecting accurate tweets, with success rates of 0.9485 and 0.9278, respectively. Yet, in detecting inaccurate tweets, BLSTM fastText had the lowest detection rate (52%) and the ArabicBERT-Base model had the second-lowest detection rate (69%). In other words, a practitioner who used the best model for identifying accurate health tweets might miss about 30% to 48% of inaccurate tweets.

Likewise, the ARBERT and AraVec Skip-Gram models performed similarly to the AraBERTv0.2 Large model in terms of precision when detecting inaccurate health-related tweets; however, these two models did not perform as well on other metrics. Although the ARBERT model performed well compared to the BSLTM models, with regard to classifying accurate tweets as inaccurate ones, it had the third-lowest rate of amongst the nine BERT models tested in this study. In other words, the ARBERT models incorrectly classified accurate tweets as inaccurate ones at a higher rate than six other BERT models, as Table 5.7 shows.

Ideally, a technique would provide high precision in both identification and recall, but that did occur in the data set for either accurate or inaccurate tweets. AraBERTv0.2 Large came closest in this regard with high accurate-tweet precision and recall, best recall for inaccurate tweets but suboptimal precision for inaccurate tweets. Likewise, AraBERT v2 Large performed quite well across accurate tweets but did not do quite so well on inaccurate tweets.

However, these models (AraBERTv0.2 Large and AraBERTv2 Large) consume relatively more resources, being based on BERT-large. Among the base models, AraBERTv0.2 Base has an $F_1$ score of 0.8543, which is good. If resources are an important consideration, then it can be considered as an alternative.

Regarding the performance of pretrained word embeddings, we found that, Mazajak Skip-Gram is the best. We made the same observation in our previous work on the detection of health-related tweets [286].

Finally, with respect to the accuracy of the best model in our study, i.e., AraBERTv0.2 Large, our results are satisfactory when compared to the results of previous studies [63, 271, 275] which make use of expert opinion. The $F_1$ score of our best model was 87%, while the best $F_1$ score reported in the study of Zhao et al. [275] was 84%, as shown in Table 5.2. Furthermore, while these previous studies target a specific health topic (such as cancer [271] or autism [275]), we use a data set of tweets on a wide range of healthcare topics suggesting that it would be more difficult to classify our data set.

It should be noted that all three studies with accuracy or F1 score higher than 90% do not rely on expert opinion (see Table 5.1 and Table 5.2). Also, two of these three studies [62, 276] targeted a specific outbreak condition (Covid-19) and their models were trained on a larger data set (for example [276] train their model on 409,484 tweets). For the third study [84], the keywords used to extract initial tweets are derived from six pre-identified rumors related to Zika. The size and the nature of data used to train these models might explain why they seem to achieve better accuracy when compared to the model proposed here: In this study, we trained a model to detect the accuracy of generic health-related information, making the approach applicable to tweets that are more or less categorical in their labeling (as illustrated in the samples shown in Table 5.6).

**Limitations**

This study only considers tweets agreed upon by experts. Although this helps us to reduce the uncertainty in our data set, it might be a limitation since the model is not trained or tested on tweets which are more marginal; tweets on which the experts are unsure.

One of the strengths of this model is that it is trained on general health-related tweets. The accuracy of the model for each health condition or health topic may vary, and future studies should evaluate the model for specific health topics.

All the models used here are language dependent and might not be applicable directly to other languages. However, there are BERT alternatives for many languages, and there is evidence that BERT outperforms word embedding-based models. We, therefore, believe that this model could perform similarly in other languages.

Regarding the metrics used to evaluate the models, it should be noted that the F1 measure has been subjected to some criticism. Although we have shown the F1 score for both classes (accurate and inaccurate health tweets), it should be noted that the measure gives equal importance to both classes (accurate and inaccurate health tweets). Also, the F1 score generally does not consider the true negative in its equation [303, 304].

## 5.5 Conclusion

The goal of this study was to develop and evaluate a state-of-the-art ML model for detecting the medical trustworthiness of health-related tweets in Arabic. To achieve this, we first constructed a labeled data set to train classifiers. Then we compared two different deep-learning approaches for training a classification model, namely six pretrained word embedding models as an input model for BLSTM, and eleven pretrained transformer language models. The percentage of health inaccurate tweets in the data is approximately 38.2%,which is comparable to previous studies that utilize data sets with a number of inaccurate health-related tweets in the range of 30%–50%. Our AraBERTv0.2 Large model achieves 87.0% macro average of the $F_1$ score on the test data set, which is satisfactory. Overall, our results clearly indicate that the AraBERTv0.2 Large model outperforms other models in detecting the medical accuracy of health-related tweets.

This study established an ML model for identifying the accuracy of health-related tweets as a response to the proliferation of health misinformation on SM. Although misinformation detection has been researched, only one study was concerned with detecting the accuracy of Arabic health-related tweets, and it was only for a specific topic (cancer). Furthermore, no DL model has been evaluated in prior studies to detect the accuracy of Arabic health-related tweets. In this study, we utilized a more extensive data set to develop a more general model using state-of-the-art ML models that have not been implemented before for this type of problem.

The potential of such work cannot be overstated. If a robust model can be built, it would allow for the detection and dissemination of accurate tweets. Likewise, it would allow for the flagging of inaccurate tweets. Both measures would significantly improve health information dissemination on Twitter. However, it should be noted that while this work will improve the situation it will still inaccurately classify 13% of tweets.

Moreover, the examples in Table 5.6 imply disparities between accurate and inaccurate information in terms of the topics covered across the data set – a trend which informal sampling of that data set supported. Accurate tweets seem to be more preventive, and inaccurate health tweets seem to promote 'natural' and alternative medicine. Thus, it might be more feasible to develop a model for detecting health topics in combination with a model for detecting the accuracy of health information and thus improving the accuracy.

In terms of further improving the accuracy of the model developed, ensemble learning can yield better results by combining models that perform well (ArabicBERT Large, ARBERT and AraVec Skip-Gram). However, ArabicBERT and AraBERTv0.2 were trained on a similar corpus as shown in Table 5.3. Another approach could be, to combine models that are pretrined on a different corpora, for example ArabicBERT Large and MARABER (ArabicBERT pretrained on Wikipedia articles and news articles; MARBERT pretrained on 1 billion tweets).

## Acknowledgments

We thank all the doctors who participated in the process of evaluating health tweets in this study. This work was supported, in part, by Taibah University, Al-Ula, Saudi Arabia, and by a grant from the Science Foundation Ireland (grant number: 13/RC/2094).

# 6 Conclusions and Future Directions

## 6.1 Summary and Overview

According to Statista, there are more than 3.9 billion SM users at the moment of writing this thesis [1], and the number is expected to increase in the next few years. One of the most frequently discussed topics on SM is health, but at the same time, a large percentage of the health information discussed is inaccurate. The harm that this health misinformation can cause to people's health is well documented [25, 73, 305]. This work presented a framework for detecting the trustworthiness of tweets in the Arabic language with an accuracy of 87.8%, albeit when the tweets were already classified as health tweets.

The pilot study conducted as a part of this research suggested that approximately 31% of Arabic health-related tweets may be inaccurate. It was also discovered that some tweets' metrics correlate with the trustworthiness of health-related tweets. After the pilot study, the research directed itself at deriving a holistic framework for identifying trustworthy tweets based on ML approaches. The first component in this framework was one to identify health-related tweets from a corpus of tweets and the second was to evaluate the trustworthiness of these health-related tweets.

In developing these different models, it was necessary to construct a health lexicon and a data set to train the models for detecting the health tweets and then determining their trustworthiness. In addition, there was a need to justify the choice of algorithms and techniques used to develop the models. This was done by reviewing the relevant literature in Chapters 4 and 5 and identifying the best methods and techniques used for both detecting the health-related tweets and determining their trustworthiness. The identified methods and techniques, from the literature, were then employed in an experimental study and the best-performing models were identified via comparisons. The findings from the experiments contribute to Arabic NLP by providing recommendations for developing Arabic NLP models.

The remainder of this chapter is organized as follows: Section 6.2 summarizes the methodology and key steps for this research. Section 6.3 then reviews the contributions of the work. In section 6.3.2, we present the data-set contributions for this thesis. In addition, other core contributions are presented in section 6.3.3. Finally, the limitations of this research, in section 6.4 and recommendations for future studies in section 6.5 are presented.

## 6.2 Thesis: Methodology and Key Steps

The primary research aim of this thesis was to provide a state-of-the-art model for detecting the trustworthiness of health-related tweets. The thesis set out to identify and evaluate health-related tweets by constructing three different data sets and experimenting with various ML models. These experiments allowed us to us identify high-performing classifiers for the NLP task of classifying health related tweets and accurate health related tweets. Toward this objective, the research involved the following steps:

1. Conducting a pilot study to explore the trustworthiness of health-related tweets and their dissemination. More than 100 metrics were identified from the literature and then tested to see whether they had any relationship to the trustworthiness of health-related tweets and their dissemination. In addition, the manual procedures used in constructing the health-related tweet data sets and to detect the trustworthiness of health-related tweets were evaluated.
2. Towards developing a model to detect health-related tweets, this study constructed a health lexicon that was used to extract tweets. Two annotators labeled 5,000 of the resultant tweets obtained, as health-related or not health-related.

3. Performing a literature review to identify gaps in current pre-processing techniques for text and showing that techniques used in the literature cannot be guaranteed to enhance the ML models. Twenty-six pre-processing techniques were evaluated on four traditional ML algorithms, which were trained on the previously mentioned data set (step 2). A brute force approach was used to find the best combination of pre-processing with the best performance traditional ML algorithms.
4. Reviewing the recent advances in Arabic NLP and identifying the present state-of-the-art in Arabic pre-trained word embeddings. These pre-trained word embeddings were used as inputs for DL models. A comparison was made between these models and the best-of-breed of traditional ML models to identify the best model for detecting health-related tweets.
5. Constructing a data set for training a trustworthiness model. 900 out of 5,000 tweets from the data sets developed in step 2 were sampled. The best model trained in step 4 was used to extract another 900 health-related tweets. These 1,800 tweets were labeled as either accurate or inaccurate by two medical doctors to construct the trustworthiness health-related tweet data set.
6. Training a trustworthiness model. As artificial intelligence has been evolving rapidly during our work, transformer language models have become the state-of-the-art for many NLP tasks. Thus, in this step, we compared the pre-trained word embedding approach to the recently published pre-trained transformer language model for Arabic.

The next section discusses how the steps outlined above were embodied in meeting the objectives, as contributions to the field of research.

## 6.3 Contributions

### 6.3.1 Key Findings

**The predictive power of tweets metrics**

The feasibility study identified metrics that can be used to assess the likelihood that health related tweets are disseminated and trustworthy, as described in sections 2.3.5 and 2.3.6. This research focused on more than 100 metrics that can be used to study health-related tweets; they are listed in Appendix A.2.

It was found that health information tweeted in the morning was more likely to be accurate compared to tweets in the evening. In addition, health professionals (traditional trusted users) are more likely to post accurate health-related tweets. SM-exclusive accounts are the least trusted users in the data set in the study. The study also suggests that untrustworthy health tweets were more likely to have embedded URLs compared to trustworthy health tweets. Somewhat surprisingly tweets that are embedded with a comma, kashida or tashkeel are more likely to be re-tweeted.

**Impact of pre-processing on traditional ML in health-tweet classification**

We identified and evaluated 26 text pre-processing techniques for training traditional ML models. To the best of our knowledge, this is a significant improvement to the most comprehensive previous work in this domain, which evaluated 16 pre-processing techniques [213].

The primary takeaway was that most of the text pre-processing techniques used for Arabic NLP fail to improve traditional ML classifiers. In addition, each pre-processing approach may behave differently from algorithm to algorithm. For example, root stemming improves the accuracy of the logistic regression, SVM and KNN models but decreases the accuracy of the MNB models.

The primary concern that researchers considered when developing an ML model is if a single pre-processing approach improves a classifier more than others do, but this may not be relevant if pre-processing techniques are considered in combination. Then the question becomes: ‘what is the best combination of the pre-processing steps for ML models?’: the question addressed in this study. For example, replacing ي and ئ with ى was the best single pre-processing technique with MNB. Nevertheless, when brute force was used to identify the best combination, replacing ي and ئ with ى was not recognized in the best pre-processing combination. The best combination for MNB was to remove duplicates(letters), remove kashida, replace ة with ه, and replace ى with ي. This combination improves the F1 score for MNB classifier from 86.0 to 87.9 and the accuracy from 91.6 to 92.7.

**Evaluation of traditional ML against DL models in health-tweet classification**

The literature from other classification domains suggests that DL approaches outperform traditional ML approaches [306-310]. However, in our experiments, in this domain, MNB outperformed five CNN models on the first (known) data set. Overall, though, on the unknown data set, BLSTM and CNN outperformed the traditional techniques, even when those techniques had optimized pre-processings. And between the two DL approaches BLSTM outperformed CNN substantially.

**Pre-trained language models and word embeddings in accuracy classification**

Our results show that overall, pre-trained language models outperform pre-trained word embeddings. However, some pre-trained word embeddings outperform certain pre-trained language models or preform similarly. The conclusion we can draw is that there is no clear winner between the two in this task vertical.

### 6.3.2 List of data-set Contributions

In terms of its more material contributions, this thesis has provided two Arabic data sets for classifying texts into health-related and non-health-related categories. Likewise, it provides two Arabic data sets for classifying texts into accurate-health-information and inaccurate-health-information. These can be used for training models and evaluating models for these tasks and should prove very valuable for researchers when further probing ML for these tasks. Therefore, the materials from this research contribute to the Arabic NLP community. Table 6.1 summarizes these materials.

**Table 6.1. Summary of the material contributions**

| Data set | Description | Methods of validation | Chapters |
|---|---|---|---|
| 5,000 tweets | 5,000 tweets labeled by annotators as health related or not | Agreement: if there was disagreement between annotators, a third annotator was involved | Chapter 4 |
| 1,000 tweets | 1,000 tweets labeled by annotators as health related or not | Agreement: if there was disagreement between annotators, a third annotator was involved | Chapter 4 |
| 109 tweets | 109 health-related tweets labeled by two doctors as accurate or inaccurate | Agreement: two doctors agreed on the trustworthiness of the tweets | Chapter 3 |
| 779 | 779 health-related tweets labeled by two doctors as accurate or inaccurate | Agreement: two doctors agreed about the trustworthiness of each tweet; if there was disagreement, the tweet was excluded | Chapter 5 |

### 6.3.3 Core Contributions

The core contribution of this thesis is the development of a workflow for detecting the trustworthiness of health-related tweets. Figure 6.1 illustrates the overall workflow that starts with detecting Arabic health-related tweets and then subsequently detects their trustworthiness.

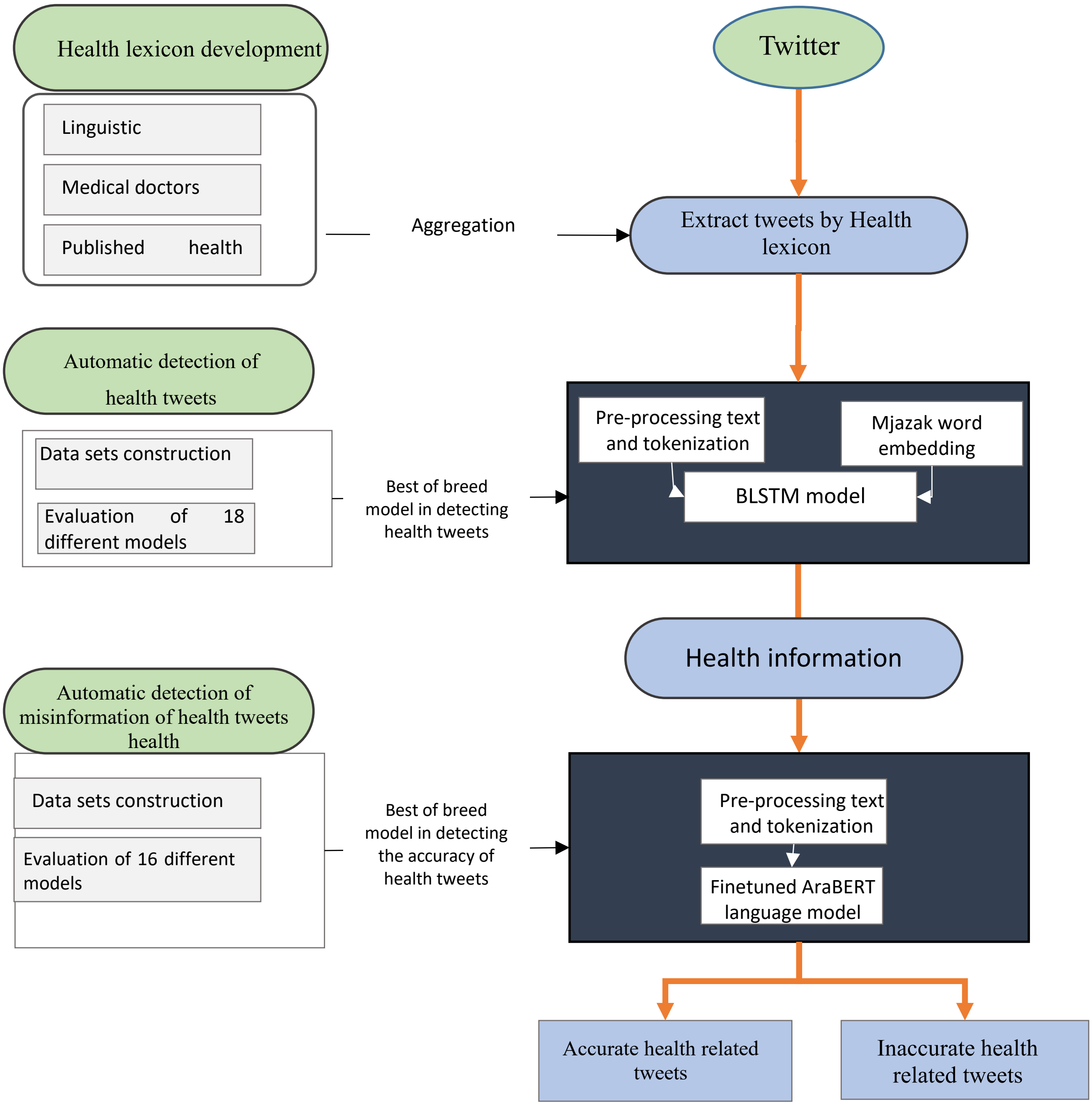


**Figure 6.1 Overall workflow for determining the accuracy of health-related tweets.**

The first component of the workflow is the detection of health-related tweets with an accuracy of 93.8%. This was achieved by developing a health lexicon and then training a ML model to filter tweets. The model provided here can be used not only as an input for the following trustworthy model but also for studying various aspects of health-related tweets, such as what makes a health-related tweet more likely to be disseminated on a large scale.

The last component of the workflow is model for determining the accuracy of health-related tweets. Our best model for this purpose achieved an accuracy of 87.8% but that is a slightly misleading figure in that, if following the workflow, there will also be an inaccuracy in identifying health tweets and that inaccuracy will cumulatively impact overall.

## 6.4 Limitations

In these studies, we did not include any reply tweets in the analysis because reply tweets usually lack contextual information. Therefore, our results cannot be generalized to interaction-type communication on SM (e.g., if a user posts a health question directly to another user). In addition, this work excluded tweets that our expert annotators disagreed on. This may have affected the results of the trustworthy model because tweets that did not achieve explicit agreement between experts may have had something in common such that their exclusion affects the classifier's performance. Likewise, it emphasizes emphatically trustworthy tweets and emphatically untrustworthy tweets, possibly limiting our classifier's ability to classify more marginal tweets.

This study used Twitter to collect data on health-related tweets in the Arabic language. In addition, although tweets were gathered over two periods and for a total of 90 days, as explained in Chapter 5, this still represents a time/generality limitation: Because of how these samples were gathered, it cannot be assumed that the conclusions of this study are generalizable for any other period.

In Chapter 4, three methods were employed to build the lexicon. Although this may be considered a strength, it could also be considered a limitation because we cannot claim that the lexicon is totally representative of the population. In addition, there were tweets for which the physicians either expressed doubt or disagreed about their veracity. This process undoubtedly omitted some health tweets, and more concerningly, may have eliminated an entire class of health tweets that were then left unstudied. However, this procedure was adopted to improve the accuracy of the data. In addition, all the physicians who served as annotators in this study were from the same nation, Saudi Arabia, which was the country the study aimed to consider.

## 6.5 Future Work

In future work, as the natural next step in the research, it would seem beneficial to additionally input tweet metrics into the ML model to help detect the trustworthiness of health-related tweets. This could elevate the already impressive performance of the ML model. In addition, findings from the pilot study also suggest that the type of user who posts a health-related tweet correlates with the trustworthiness of the tweet. Thus, developing a classifier to classify the types of users, or identifying user-type by some other means (address) could possibly improve the accuracy of the trustworthy model.

Future work should also consider developing ML models using ensemble learning to yield better results by combining models that perform well (ArabicBERT large, ARBERT, and AraVec Skip-Gram). Another strategy that could be followed is to combine models that are pre-trained on different corpora, for example, ArabicBERT large and MARABER. (ArabicBERT is pre-trained on Wikipedia articles and news articles; MARBERT is pre-trained on 1 billion tweets).

The examples of accurate and inaccurate tweets provided in Chapter 5 of this thesis suggest discrepancies between accurate and inaccurate health tweets in terms of the topics covered across the data set. Accurate health tweets appear to be more preventive, whereas inaccurate health tweets promote 'natural' and alternative medicine. As a result, developing a model for recognizing health issues in combination with a model for detecting trustworthiness might help in improving the accuracy trustworthiness classifier. Also, it might help health organizations to plan health campaigns, for example, by identifying health topics with a high percentage of inaccurate health tweets.

Lastly, future work should also probe why certain pre-processings and approaches work better than others rather than focusing more exclusively on achieving the best solution.

Indeed, this criticism can be levelled at most research efforts that apply predictive modelling (here, detecting the trustworthiness of health tweets) [311]. This is a much wider issue than the scope of the work described here.

This study has provided a novel framework to evaluate Arabic text pre-processing techniques. For future studies, more data sets can be used to assess these pre-processing techniques even further. In addition, the framework used to assess text pre-processing techniques can be compared to other methods of identifying the best pre-processing techniques [213].

# A. Appendix for Chapter 3

This appendix presents the list of material included in pilot study reported in Chapter 3.

## A.1 Link to health keywords used in Chapter 3

Health keywords

## A.2 List of tweets' features used in the study

| Level | Scope | Feature symbol | | Description |
|---|---|---|---|---|
| Tweet | Activity | hashtags_count | | Number of hashtags used in the tweet |
| Tweet | Activity | urls_count | | Number of URLs used in the tweet |
| Tweet | Activity | URLs | | Does the tweet contain URLs |
| Tweet | Activity | Hashtags | | Does the tweet have hashtag or not |
| Tweet | Linguistic | count_words | | Number of words in the tweet |
| Tweet | Linguistic | *Kashida* | | Does the tweet contain *Kashida* |
| Tweet | Linguistic | charac_count | | Number of characters in the tweet |
| Tweet | Linguistic | Exclamation_mark | | Does the tweet contain '!' |
| Tweet | Linguistic | q_mark | | Does the tweet contain '?' |
| Tweet | Linguistic | dot_statues | | Does the tweet contain '.' |
| Tweet | Linguistic | Comma | | Does the tweet contain '،' |
| Tweet | Linguistic | Semi_colon | | Does the tweet contain ';' |
| Tweet | Linguistic | Tashkeel | | Does the tweet contain 'tashkeel' |
| Tweet | Popularity | Retweeted | | Is the tweet retweeted or not |
| Tweet | Popularity | Fav | | Is the tweet "favorited" |
| Tweet | Time | Time | | Time at which the tweet is posted |
| Tweet | Time | week days | | Is the tweet posted during week days or at the weekend |

| **Leve l** | **Scope** | **Feature symbol** | **Description** |
|---|---|---|---|
| User | Activity | wdeve_device | Most used device during the evening of the week days |
| User | Activity | wdmor_device | Most used device during the morning of the week days |
| User | Activity | wdnight_device | Most used type of device to post tweet during the night of the week days |
| User | Activity | wemor_device | Most used device during the morning of the weekend days |
| User | Activity | RT1 | Number of tweets that author retweeted |
| User | Activity | OT1 | Number of original tweets |
| User | Activity | Tweet Score | RT1 + OT1 |
| User | Activity | RP1 | Number of reply-to tweets posted by the author |
| User | Activity | FT1 | Number of tweets this user has liked |
| User | Activity | GT | OT1+RP1 |
| User | Activity | GT1 | OT1+RP1+FT1 |
| User | Activity | SSI | Rate of original tweets posted by the author to tweets retweeted by the author (OT1/RT1) |
| User | Activity | F3 | Followees count |
| User | Activiety | OT2 | Number of links shared |
| User | Activity | M1 | Number of tweets where the author mentioned other users |
| User | Activity | M2 | Number of unique users mentioned by the author |
| User | Activity | OT3 | Number of hashtags in the author‘s tweets |
| User | Activity | MH6 | Proportion of tweets that have hashtags to tweets that does not |
| User | Activity | GInSc | RT1+RP1 |
| User | Activity | statuses_count | Number of tweets posted by the user |
| User | Activity | verified | Is the user account verified or not |
| User | Activity | location | Location of the user |
| User | Activity | mo_usedlang | Most used language of the user |
| User | Activity | se_mo_usedlang | Second most used language of the user |
| User | Activity | per_of_mo_usedlang | How often does the user use the most used language in their tweets? |
| User | Activity | per_of_se_mo_usedlang | How often does the user use the second most used language in their tweets? |
| User | Activity | n_countries | Number of countries visited by the author (if geo enabled) |
| User | Activity | n_cities | Number of cities visited by the author (if geo enabled) |
| User | Activity | MU1 | Mean URL count in original tweets |

| | | | |
|---|---|---|---|
| User | Activity | MU2 | Median URL count in original tweets |
| User | Activity | MU3 | Max URL count in an original tweet |
| User | Activity | MU4 | Minimum URL count in an original tweet |
| User | Activity | MU5 | Unique URLs count in URLs in original tweets |
| User | Activity | RM1 | Number of retweeted tweets by the author where the user mentioned other users |
| User | Activity | RMU1 | Mean URL count in re-tweeted tweets |
| User | Activity | RMU2 | Median URL count in re-tweeted tweets |
| User | Activity | RMU3 | Max URL count in re-tweeted tweets |
| User | Activity | RMU4 | Minimum URL count in re-tweeted tweets |
| User | Activity | MH1 | Mean hashtags count in original tweets |
| User | Activity | MH2 | Median hashtags count in original tweets |
| User | Activity | MH3 | Max hashtags count in an original tweet |
| User | Activity | MH4 | Minimum hashtags count in an original tweet |
| User | Activity | MH5 | Unique hashtags count in original tweets |
| User | Activity | MM1 | Mean mentions' count in original tweets |
| User | Activity | MM2 | Median mentions count in original tweets |
| User | Activity | MM3 | Max mentions count in an original tweet |
| User | Activity | MM4 | Minimum mentions count in an original tweet |
| User | Activity | MM5 | Unique mentions count in URLs in re-tweeted tweets |
| User | Activity | RMH1 | Mean hashtags count in re-tweeted tweets |
| User | Activity | RMH2 | Median hashtags count in re-tweeted tweets |
| User | Activity | RMH3 | Max hashtags count in re-tweeted tweets |
| User | Activity | RMH4 | Minimum hashtags count in re-tweeted tweets |
| User | Activity | RMH5 | Unique hashtags count in tweets that were retweeted by the user |
| User | Activity | RMM1 | Mean mentions' count in re-tweeted tweets |
| User | Activity | RMM2 | Median mentions count in re-tweeted tweets |
| User | Activity | RMM3 | Max mentions count in re-tweeted tweets |
| User | Activity | RMM4 | Minimum mentions count in re-tweeted tweets |
| User | Activity | RMM5 | Unique mentions in retweeted tweets by the author |
| User | Activity | rate_tweets_count_to_days | Average tweets per day |
| User | Activity | rate_GT_count_to_days | Average of general activity (post tweets, favourites) per day |
| User | **Activity** | trF1 | Log transformed Friends |
| User | Activity | trF3 | Log transformed Followers |
| User | Activity | trFT1 | Log transformed Favourites |
| User | Activity | trstatuses_count | Log transformed Tweet count |
| User | Activity | OT1 | Number of original tweets |

| | | | |
|---|---|---|---|
| User | Behaviour | geo_enabled | Is the geolocation enabled or not? |
| User | linguistic | UNC | Screen name words' count |
| User | linguistic | UDC | Description words' count |
| User | linguistics | lang | Language of the user |
| User | linguistics | n_langauges | Number of languages used by the user in their tweets |
| User | Popularity | F1 | Followers count |
| User | Popularity | FR | F1/F1+F3 |
| User | Popularity | TFF | F1/F3 (rate of followers to followees) |
| User | Popularity | RT5 | Sum of retweet count for OT1 (tweets post by the user) |
| User | Popularity | RT2 | Number of author’s tweets re-tweeted by other users |
| User | Popularity | FT2 | Number of times author’s tweets are favorited |
| User | Popularity | FT6 | Number of original tweets posted by the author that are "favorited" |
| User | Popularity | Listed_count | Number of lists where the user is member |
| User | Time | wd_freq | Rate of tweets posted during week days to total numbers of tweets |
| User | Time | wdeve_rate | Rate of tweets posted during evening of week days to total numbers of tweets |
| User | Time | wdmor_rate | Rate of tweets posted during morning of week days to total numbers of tweets |
| User | Time | wdnight_rate | Rate of tweets posted during night of week days to total numbers of tweets |
| User | Time | we_freq | Rate of tweets posted during weekend to total numbers of tweets |
| User | Time | weeve_rate | Rate of tweets posted during evening of weekend to total numbers of tweets |
| User | Time | wemor_rate | Rate of tweets posted during morning of weekend to total numbers of tweets |
| User | Time | wenight_rate | Rate of tweets posted during night of weekend to total numbers of tweets |
| User | Time | wenight_device | Most used type of device to post tweet during the night of the weekend |
| User | Time | twee_eve_we | Number of tweets posted by the user during evening of weekend |
| User | Time | twee_mor_we | Number of tweets posted by the user during morning of weekend |
| User | Time | twee_ni_we | Number of tweets posted by the user during night of weekend |
| User | Time | created_date | Date in which the user created the account |

# B. Appendix for Chapter 4

## B.1 Hyperparameters for each algorithm

MNB

*Table 1 HYPERPARAMETERS OPTIMIZED FOR NB*

| *Hyperparameters* | 'tfidf__norm' | 'clf__alpha' |
|---|---|---|
| Range | ('l1', 'l2'), | [1, 1e-1, 1e-2] |

SVM:

*Table 2 HYPERPARAMETERS OPTIMIZED FOR SVM*

| *Hyperparameters* | 'clf__C' | 'clf__loss' |
|---|---|---|
| Range | [.05, .12, .25, .5, 1, 2, 4] | ['squared_hinge','hinge'] |

LG

*Table 3 HYPERPARAMETERS OPTIMIZED FOR LG*

| *Hyperparameters* | 'clf__C' | 'clf__loss' |
|---|---|---|
| Range | [0.001,0.01,0.1,1,10,100] | ['l1','l2'] |

KNN

*Table 4 HYPERPARAMETERS OPTIMIZED FOR KNN*

| *Hyperparameters* | clf__algorithm | clf__n_neighbors | clf__leaf_size |
|---|---|---|---|
| Range | ['auto', 'kd_tree'] | [2, 4,6,8,9,10,11,12] | [2, 9,16,20,26,31,50,70] |

## B.2 Results of the evaluation of the best functions.

| | Number of functions | functions | f1-score | evaluation | accuracy | precision | recall |
|---|---|---|---|---|---|---|---|
| **1** | 0 | [] | 0.86 | normal | 0.916 | 0.874576 | 0.845902 |
| **2** | 1 | [<Remove duplicate>] | 0.864322 | increase | 0.919 | 0.883562 | 0.845902 |
| **3** | 1 | [<Light stemming>] | 0.865874 | increase | 0.921 | 0.897887 | 0.836066 |
| **4** | 1 | [<Remove Kashida>] | 0.863333 | increase | 0.918 | 0.877966 | 0.84918 |
| **5** | 1 | [<Replace ى with ي>] | 0.866667 | increase | 0.92 | 0.881356 | 0.852459 |
| **6** | 1 | [<Replace ي and ئ with ى>] | 0.868114 | increase | 0.921 | 0.884354 | 0.852459 |
| **7** | 1 | [<Lemmatization>] | 0.863558 | increase | 0.921 | 0.912409 | 0.819672 |
| **8** | 1 | [<Replace ة with ه>] | 0.866779 | increase | 0.921 | 0.892361 | 0.842623 |
| **9** | 2 | [<Remove duplicate>, <Light stemming>] | 0.861486 | increase | 0.918 | 0.888502 | 0.836066 |
| **10** | 2 | [<Remove duplicate>, <Remove Kashida>] | 0.868114 | increase | 0.921 | 0.884354 | 0.852459 |
| **11** | 2 | [<Remove duplicate>, <Replace ى with ي>] | 0.869128 | increase | 0.922 | 0.890034 | 0.84918 |
| **12** | 2 | [<Remove duplicate>, <Replace ي and ئ with ى>] | 0.869128 | increase | 0.922 | 0.890034 | 0.84918 |
| **13** | 2 | [<Remove duplicate>, <Lemmatization>] | 0.86532 | increase | 0.92 | 0.889273 | 0.842623 |
| **14** | 2 | [<Remove duplicate>, <Replace ة with ه>] | 0.866779 | increase | 0.921 | 0.892361 | 0.842623 |
| **15** | 2 | [<Light stemming>, <Remove Kashida>] | 0.865517 | increase | 0.922 | 0.912727 | 0.822951 |
| **16** | 2 | [<Light stemming>, <Replace ى with ي>] | 0.863866 | increase | 0.919 | 0.886207 | 0.842623 |
| **17** | 2 | [<Light stemming>, <Replace ي and ئ with ى>] | 0.871622 | increase | 0.924 | 0.898955 | 0.845902 |
| **18** | 2 | [<Light stemming>, <Lemmatization>] | 0.863406 | increase | 0.919 | 0.888889 | 0.839344 |
| **19** | 2 | [<Light stemming>, <Replace ة with ه>] | 0.865874 | increase | 0.921 | 0.897887 | 0.836066 |
| **20** | 2 | [<Remove Kashida>, <Replace ى with ي>] | 0.870432 | increase | 0.922 | 0.882155 | 0.859016 |
| **21** | 2 | [<Remove Kashida>, <Replace ي and ئ with ى>] | 0.87188 | increase | 0.923 | 0.885135 | 0.859016 |
| **22** | 2 | [<Remove Kashida>, <Lemmatization>] | 0.863558 | increase | 0.921 | 0.912409 | 0.819672 |
| **23** | 2 | [<Remove Kashida>, <Replace ة with ه>] | 0.870588 | increase | 0.923 | 0.893103 | 0.84918 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **24** | 2 | [<Replace ى with ي>, <Replace ي and ئ with ى>] | 0.868114 | increase | 0.921 | 0.884354 | 0.852459 |
| **25** | 2 | [<Replace ى with ي>, <Lemmatization>] | 0.862944 | increase | 0.919 | 0.891608 | 0.836066 |
| **26** | 2 | [<Replace ى with ي>, <Replace ة with ه>] | 0.868114 | increase | 0.921 | 0.884354 | 0.852459 |
| **27** | 2 | [<Replace ي and ئ with ى>, <Lemmatization>] | 0.866667 | increase | 0.92 | 0.881356 | 0.852459 |
| **28** | 2 | [<Replace ي and ئ with ى>, <Replace ة with ه>] | 0.871022 | increase | 0.923 | 0.890411 | 0.852459 |
| **29** | 2 | [<Lemmatization>, <Replace ة with ه>] | 0.869863 | increase | 0.924 | 0.910394 | 0.832787 |
| **30** | 3 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>] | 0.869863 | increase | 0.924 | 0.910394 | 0.832787 |
| **31** | 3 | [<Remove duplicate>, <Light stemming>, <Replace ى with ي>] | 0.865417 | increase | 0.921 | 0.900709 | 0.832787 |
| **32** | 3 | [<Remove duplicate>, <Light stemming>, <Replace ي and ئwith ى>] | 0.866438 | increase | 0.922 | 0.90681 | 0.829508 |
| **33** | 3 | [<Remove duplicate>, <Light stemming>, <Lemmatization>] | 0.863481 | increase | 0.92 | 0.900356 | 0.829508 |
| **34** | 3 | [<Remove duplicate>, <Light stemming>, <Replace ة with ه>] | 0.869863 | increase | 0.924 | 0.910394 | 0.832787 |
| **35** | 3 | [<Remove duplicate>, <Remove Kashida>, <Replace ى with ي>] | 0.871452 | increase | 0.923 | 0.887755 | 0.855738 |
| **36** | 3 | [<Remove duplicate>, <Remove Kashida>, <Replace ي and ئwith ى>] | 0.870152 | increase | 0.923 | 0.895833 | 0.845902 |
| **37** | 3 | [<Remove duplicate>, <Remove Kashida>, <Lemmatization>] | 0.86532 | increase | 0.92 | 0.889273 | 0.842623 |
| **38** | 3 | [<Remove duplicate>, <Remove Kashida>, <Replace ة with ه>] | 0.876667 | increase | 0.926 | 0.891525 | 0.862295 |
| **39** | 3 | [<Remove duplicate>, <Replace ى with ي>, <Replace ي and ئwith ى>] | 0.869128 | increase | 0.922 | 0.890034 | 0.84918 |
| **40** | 3 | [<Remove duplicate>, <Replace ى with ي>, <Lemmatization>] | 0.868687 | increase | 0.922 | 0.892734 | 0.845902 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **41** | 3 | [<Remove duplicate>, <Replace ى with ي>, <Replace ة with ه>] | 0.876667 | increase | 0.926 | 0.891525 | 0.862295 |
| **42** | 3 | [<Remove duplicate>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.863333 | increase | 0.918 | 0.877966 | 0.84918 |
| **43** | 3 | [<Remove duplicate>, <Replace ي and ئ with ى>, <Replace ةwith ه>] | 0.876254 | increase | 0.926 | 0.894198 | 0.859016 |
| **44** | 3 | [<Remove duplicate>, <Lemmatization>, <Replace ة with ه>] | 0.870152 | increase | 0.923 | 0.895833 | 0.845902 |
| **45** | 3 | [<Light stemming>, <Remove Kashida>, <Replace ى with ي>] | 0.867672 | increase | 0.921 | 0.886986 | 0.84918 |
| **46** | 3 | [<Light stemming>, <Remove Kashida>, <Replace ي and ئ with ى>] | 0.870152 | increase | 0.923 | 0.895833 | 0.845902 |
| **47** | 3 | [<Light stemming>, <Remove Kashida>, <Lemmatization>] | 0.863406 | increase | 0.919 | 0.888889 | 0.839344 |
| **48** | 3 | [<Light stemming>, <Remove Kashida>, <Replace ة with ه>] | 0.868376 | increase | 0.923 | 0.907143 | 0.832787 |
| **49** | 3 | [<Light stemming>, <Replace ى with ي>, <Replace ي and ئ with ى>] | 0.871622 | increase | 0.924 | 0.898955 | 0.845902 |
| **50** | 3 | [<Light stemming>, <Replace ى with ي>, <Lemmatization>] | 0.861017 | increase | 0.918 | 0.891228 | 0.832787 |
| **51** | 3 | [<Light stemming>, <Replace ى with ي>, <Replace ة with ه>] | 0.873524 | increase | 0.925 | 0.899306 | 0.84918 |
| **52** | 3 | [<Light stemming>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.862479 | increase | 0.919 | 0.894366 | 0.832787 |
| **53** | 3 | [<Light stemming>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.871186 | increase | 0.924 | 0.901754 | 0.842623 |
| **54** | 3 | [<Light stemming>, <Lemmatization>, <Replace ة with ه>] | 0.867347 | increase | 0.922 | 0.90106 | 0.836066 |
| **55** | 3 | [<Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>] | 0.87188 | increase | 0.923 | 0.885135 | 0.859016 |
| **56** | 3 | [<Remove Kashida>, <Replace ى with ي>, <Lemmatization>] | 0.862944 | increase | 0.919 | 0.891608 | 0.836066 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **57** | 3 | [<Remove Kashida>, <Replace ى with ي>, <Replace ة with ه>] | 0.87395 | increase | 0.925 | 0.896552 | 0.852459 |
| **58** | 3 | [<Remove Kashida>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.866667 | increase | 0.92 | 0.881356 | 0.852459 |
| **59** | 3 | [<Remove Kashida>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.876667 | increase | 0.926 | 0.891525 | 0.862295 |
| **60** | 3 | [<Remove Kashida>, <Lemmatization>, <Replace ة with ه>] | 0.869863 | increase | 0.924 | 0.910394 | 0.832787 |
| **61** | 3 | [<Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.866667 | increase | 0.92 | 0.881356 | 0.852459 |
| **62** | 3 | [<Replace ى with ي>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.871022 | increase | 0.923 | 0.890411 | 0.852459 |
| **63** | 3 | [<Replace ى with ي>, <Lemmatization>, <Replace ة with ه>] | 0.866328 | increase | 0.921 | 0.895105 | 0.839344 |
| **64** | 3 | [<Replace ي and ئ with ى>, <Lemmatization>, <Replace ةwith ه>] | 0.87188 | increase | 0.923 | 0.885135 | 0.859016 |
| **65** | 4 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ى with ي>] | 0.86927 | increase | 0.923 | 0.901408 | 0.839344 |
| **66** | 4 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ي and ئ with ى>] | 0.870307 | increase | 0.924 | 0.907473 | 0.836066 |
| **67** | 4 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Lemmatization>] | 0.863481 | increase | 0.92 | 0.900356 | 0.829508 |
| **68** | 4 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ة with ه>] | 0.870307 | increase | 0.924 | 0.907473 | 0.836066 |
| **69** | 4 | [<Remove duplicate>, <Light stemming>, <Replace ى with ي>, <Replace ي and ئ with ى>] | 0.866438 | increase | 0.922 | 0.90681 | 0.829508 |
| **70** | 4 | [<Remove duplicate>, <Light stemming>, <Replace ى with ي>, <Lemmatization>] | 0.85906 | decrease | 0.916 | 0.879725 | 0.839344 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **71** | 4 | [<Remove duplicate>, <Light stemming>, <Replace ى with ي>, <Replace ة with ه>] | 0.862479 | increase | 0.919 | 0.894366 | 0.832787 |
| **72** | 4 | [<Remove duplicate>, <Light stemming>, <Replace ي and ئwith ى>, <Lemmatization>] | 0.864957 | increase | 0.921 | 0.903571 | 0.829508 |
| **73** | 4 | [<Remove duplicate>, <Light stemming>, <Replace ي and ئwith ى>, <Replace ة with ه>] | 0.86927 | increase | 0.923 | 0.901408 | 0.839344 |
| **74** | 4 | [<Remove duplicate>, <Light stemming>, <Lemmatization>, <Replace ة with ه>] | 0.865517 | increase | 0.922 | 0.912727 | 0.822951 |
| **75** | 4 | [<Remove duplicate>, <Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>] | 0.870152 | increase | 0.923 | 0.895833 | 0.845902 |
| **76** | 4 | [<Remove duplicate>, <Remove Kashida>, <Replace ى with ي>, <Lemmatization>] | 0.868687 | increase | 0.922 | 0.892734 | 0.845902 |
| **77** | 4 | [<Remove duplicate>, <Remove Kashida>, <Replace ى with ي>, <Replace ة with ه>] | 0.878536 | increase | 0.927 | 0.891892 | 0.865574 |
| **78** | 4 | [<Remove duplicate>, <Remove Kashida>, <Replace ي and ئwith ى>, <Lemmatization>] | 0.863333 | increase | 0.918 | 0.877966 | 0.84918 |
| **79** | 4 | [<Remove duplicate>, <Remove Kashida>, <Replace ي and ئwith ى>, <Replace ة with ه>] | 0.872483 | increase | 0.924 | 0.893471 | 0.852459 |
| **80** | 4 | [<Remove duplicate>, <Remove Kashida>, <Lemmatization>, <Replace ة with ه>] | 0.870152 | increase | 0.923 | 0.895833 | 0.845902 |
| **81** | 4 | [<Remove duplicate>, <Replace ى with ي>, <Replace ي and ئwith ى>, <Lemmatization>] | 0.863333 | increase | 0.918 | 0.877966 | 0.84918 |
| **82** | 4 | [<Remove duplicate>, <Replace ى with ي>, <Replace ي and ئwith ى>, <Replace ة with ه>] | 0.876254 | increase | 0.926 | 0.894198 | 0.859016 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **83** | 4 | [<Remove duplicate>, <Replace ى with ي>, <Lemmatization>, <Replace ة with ه>] | 0.866667 | increase | 0.92 | 0.881356 | 0.852459 |
| **84** | 4 | [<Remove duplicate>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.875421 | increase | 0.926 | 0.899654 | 0.852459 |
| **85** | 4 | [<Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>] | 0.870152 | increase | 0.923 | 0.895833 | 0.845902 |
| **86** | 4 | [<Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Lemmatization>] | 0.861017 | increase | 0.918 | 0.891228 | 0.832787 |
| **87** | 4 | [<Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Replace ة with ه>] | 0.867347 | increase | 0.922 | 0.90106 | 0.836066 |
| **88** | 4 | [<Light stemming>, <Remove Kashida>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.862479 | increase | 0.919 | 0.894366 | 0.832787 |
| **89** | 4 | [<Light stemming>, <Remove Kashida>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.873096 | increase | 0.925 | 0.902098 | 0.845902 |
| **90** | 4 | [<Light stemming>, <Remove Kashida>, <Lemmatization>, <Replace ة with ه>] | 0.867347 | increase | 0.922 | 0.90106 | 0.836066 |
| **91** | 4 | [<Light stemming>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.862479 | increase | 0.919 | 0.894366 | 0.832787 |
| **92** | 4 | [<Light stemming>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.871186 | increase | 0.924 | 0.901754 | 0.842623 |
| **93** | 4 | [<Light stemming>, <Replace ى with ي>, <Lemmatization>, <Replace ة with ه>] | 0.860544 | increase | 0.918 | 0.893993 | 0.829508 |
| **94** | 4 | [<Light stemming>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.865874 | increase | 0.921 | 0.897887 | 0.836066 |
| **95** | 4 | [<Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ | 0.866667 | increase | 0.92 | 0.881356 | 0.852459 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | with ى>, <Lemmatization>] | | | | | |
| **96** | 4 | [<Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.876667 | increase | 0.926 | 0.891525 | 0.862295 |
| **97** | 4 | [<Remove Kashida>, <Replace ى with ي>, <Lemmatization>, <Replace ة with ه>] | 0.866328 | increase | 0.921 | 0.895105 | 0.839344 |
| **98** | 4 | [<Remove Kashida>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.87188 | increase | 0.923 | 0.885135 | 0.859016 |
| **99** | 4 | [<Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.87188 | increase | 0.923 | 0.885135 | 0.859016 |
| **100** | 5 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>] | 0.870307 | increase | 0.924 | 0.907473 | 0.836066 |
| **101** | 5 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Lemmatization>] | 0.85906 | decrease | 0.916 | 0.879725 | 0.839344 |
| **102** | 5 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Replace ة with ه>] | 0.86927 | increase | 0.923 | 0.901408 | 0.839344 |
| **103** | 5 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.864957 | increase | 0.921 | 0.903571 | 0.829508 |
| **104** | 5 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.87395 | increase | 0.925 | 0.896552 | 0.852459 |
| **105** | 5 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Lemmatization>, <Replace ة with ه>] | 0.865517 | increase | 0.922 | 0.912727 | 0.822951 |
| **106** | 5 | [<Remove duplicate>, <Light stemming>, <Replace | 0.864957 | increase | 0.921 | 0.903571 | 0.829508 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | ی with ي>, <Replace ي and ئ with ى>, <Lemmatization>] | | | | | |
| **107** | 5 | [<Remove duplicate>, <Light stemming>, <Replace ی with ي>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.86927 | increase | 0.923 | 0.901408 | 0.839344 |
| **108** | 5 | [<Remove duplicate>, <Light stemming>, <Replace ی with ي>, <Lemmatization>, <Replace ة with ه>] | 0.85473 | decrease | 0.914 | 0.881533 | 0.829508 |
| **109** | 5 | [<Remove duplicate>, <Light stemming>, <Replace ي and ئwith ى>, <Lemmatization>, <Replace ة with ه>] | 0.865417 | increase | 0.921 | 0.900709 | 0.832787 |
| **110** | 5 | [<Remove duplicate>, <Remove Kashida>, <Replace ی with ي>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.863333 | increase | 0.918 | 0.877966 | 0.84918 |
| **111** | 5 | [<Remove duplicate>, <Remove Kashida>, <Replace ی with ي>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.872483 | increase | 0.924 | 0.893471 | 0.852459 |
| **112** | 5 | [<Remove duplicate>, <Remove Kashida>, <Replace ی with ي>, <Lemmatization>, <Replace ة with ه>] | 0.866667 | increase | 0.92 | 0.881356 | 0.852459 |
| **113** | 5 | [<Remove duplicate>, <Remove Kashida>, <Replace ي and ئwith ى>, <Lemmatization>, <Replace ة with ه>] | 0.875421 | increase | 0.926 | 0.899654 | 0.852459 |
| **114** | 5 | [<Remove duplicate>, <Replace ی with ي>, <Replace ي and ئwith ى>, <Lemmatization>, <Replace ة with ه>] | 0.875421 | increase | 0.926 | 0.899654 | 0.852459 |
| **115** | 5 | [<Light stemming>, <Remove Kashida>, <Replace ی with ي>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.862479 | increase | 0.919 | 0.894366 | 0.832787 |
| **116** | 5 | [<Light stemming>, <Remove Kashida>, <Replace ی with ي>, | 0.873096 | increase | 0.925 | 0.902098 | 0.845902 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | <Replace ي and ئ with ى>, <Replace ة with ه>] | | | | | |
| **117** | 5 | [<Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Lemmatization>, <Replace ة with ه>] | 0.860544 | increase | 0.918 | 0.893993 | 0.829508 |
| **118** | 5 | [<Light stemming>, <Remove Kashida>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.865874 | increase | 0.921 | 0.897887 | 0.836066 |
| **119** | 5 | [<Light stemming>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.865874 | increase | 0.921 | 0.897887 | 0.836066 |
| **120** | 5 | [<Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.87188 | increase | 0.923 | 0.885135 | 0.859016 |
| **121** | 6 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>] | 0.864957 | increase | 0.921 | 0.903571 | 0.829508 |
| **122** | 6 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Replace ة with ه>] | 0.87395 | increase | 0.925 | 0.896552 | 0.852459 |
| **123** | 6 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Lemmatization>, <Replace ةwith ه>] | 0.85473 | decrease | 0.914 | 0.881533 | 0.829508 |
| **124** | 6 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.865417 | increase | 0.921 | 0.900709 | 0.832787 |
| **125** | 6 | [<Remove duplicate>, <Light stemming>, <Replace ى with ي>, <Replace ي and ئ with ى>, | 0.865417 | increase | 0.921 | 0.900709 | 0.832787 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | <Lemmatization>, <Replace ة with ه>] | | | | | |
| **126** | 6 | [<Remove duplicate>, <Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.875421 | increase | 0.926 | 0.899654 | 0.852459 |
| **127** | 6 | [<Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.865874 | increase | 0.921 | 0.897887 | 0.836066 |
| **128** | 7 | [<Remove duplicate>, <Light stemming>, <Remove Kashida>, <Replace ى with ي>, <Replace ي and ئ with ى>, <Lemmatization>, <Replace ة with ه>] | 0.865417 | increase | 0.921 | 0.900709 | 0.832787 |

# C. Appendix for Chapter 4

## C.1 Tweets ids with their labels used in the study

File tweets ids with labels

# References


1. Statista. Number of social network users worldwide from 2017 to 2025. 2021; Available from: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.
2. Yeung AWK, Kletecka-Pulker M, Eibensteiner F, Plunger P, Völkl-Kernstock S, Willschke H, et al. Implications of Twitter in Health-Related Research: A Landscape Analysis of the Scientific Literature. Frontiers in Public Health. 2021 2021-July-09;9. doi: 10.3389/fpubh.2021.654481.
3. Park H, Reber BH, Chon M-G. Tweeting as Health Communication: Health Organizations' Use of Twitter for Health Promotion and Public Engagement. Journal of Health Communication. 2016 2016/02/01;21(2):188-98. PMID: 26716546. doi: 10.1080/10810730.2015.1058435.
4. Tang L, Bie BJ, Park SE, Zhi DG. Social media and outbreaks of emerging infectious diseases: A systematic review of literature. Am J Infect Control. 2018 Sep;46(9):962-72. PMID: 29628293. doi: 10.1016/j.ajic.2018.02.010.
5. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EH, Olsen JM, et al. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PLoS One. 2015;10(10):e0139701. PMID: 26437454. doi: 10.1371/journal.pone.0139701.
6. Abbass MAA, Keshava HB, Delaney CP. The Bigger Picture: Picking the Right Soap Box-Is it Possible to Connect with Different Audience Targets (Practitioners and Patients) from the Same Platform? Clin Colon Rectal Surg. 2017 Sep;30(4):281-90. PMID: 28924403. doi: 10.1055/s-0037-1604258.
7. Alsobayel H. Use of Social Media for Professional Development by Health Care Professionals: A Cross-Sectional Web-Based Survey. JMIR Med Educ. 2016 Sep 12;2(2):e15. PMID: 27731855. doi: 10.2196/mededu.6232.
8. Alanzi T, Al-Habib DK. The Use of Social Media by Healthcare Quality Personnel in Saudi Arabia. Journal of Environmental and Public Health. 2020 2020/05/22;2020:1417478. doi: 10.1155/2020/1417478.
9. Al-Khalifa KS, Al-Swuailem AS, AlSheikh R, Muazen YY, Al-Khunein YA, Halawany H, et al. The use of social media for professional purposes among dentists in Saudi Arabia. BMC Oral Health. 2021 2021/01/12;21(1):26. doi: 10.1186/s12903-021-01390-w.
10. Omair MA, AlOhaly RY, Alashgar LM. Awareness and Misconceptions of Female Students in King Saud University on Systemic Lupus Erythematosus. Rheumatology: Current Research. 2015;05(03). doi: 10.4172/2161-1149.1000165.
11. Antheunis ML, Tates K, Nieboer TE. Patients' and health professionals' use of social media in health care: Motives, barriers and expectations. Patient Education and Counseling. 2013 2013/09/01/;92(3):426-31. doi: https://doi.org/10.1016/j.pec.2013.06.020.
12. Alnemer KA, Alhuzaim WM, Alnemer AA, Alharbi BB, Bawazir AS, Barayyan OR, et al. Are Health-Related Tweets Evidence Based? Review and Analysis of Health-Related Tweets on Twitter. J Med Internet Res. 2015 Oct 29;17(10):e246. PMID: 26515535. doi: 10.2196/jmir.4898.
13. Clement J. Countries with the most Twitter users 2019. Statista; 2019; Available from: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.
14. Boghardt L. Saudi Arabia's War on Twitter. The Washington Institute for Near East Policy, 2013.
15. Radcliffe D, Bruni P. State of social media. Middle East. 2018;1.
16. Ghenai A, Mejova Y. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. Proc ACM Hum-Comput Interact. 2018;2(CSCW):1-20. doi: 10.1145/3274327.

17. Wang Y, McKee M, Torbica A, Stuckler D. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. Social Science & Medicine. 2019 2019/11/01/;240:112552. doi: https://doi.org/10.1016/j.socscimed.2019.112552.

18. Krittanawong C, Narasimhan B, Virk HUH, Narasimhan H, Hahn J, Wang Z, et al. Misinformation Dissemination in Twitter in the COVID-19 Era. Am J Med. 2020;133(12):1367-9. PMID: 32805227. doi: 10.1016/j.amjmed.2020.07.012.

19. Valecha R, Volety T, Rao HR, Kwon KH. Misinformation Sharing on Twitter During Zika: An Investigation of the Effect of Threat and Distance. IEEE Internet Computing. 2021;25(1):31-9. doi: 10.1109/MIC.2020.3044543.

20. Chen L, Wang X, Peng T-Q. Nature and Diffusion of Gynecologic Cancer–Related Misinformation on Social Media: Analysis of Tweets. J Med Internet Res. 2018;20(10):e11515. PMID: 30327289. doi: 10.2196/11515.

21. Jamison A, Broniatowski DA, Smith MC, Parikh KS, Malik A, Dredze M, et al. Adapting and Extending a Typology to Identify Vaccine Misinformation on Twitter. Am J Public Health. 2020;110(S3):S331-S9. PMID: 33001737. doi: 10.2105/AJPH.2020.305940.

22. Bautista JR, Zhang Y, Gwizdka J. Healthcare professionals' acts of correcting health misinformation on social media. International Journal of Medical Informatics. 2021 2021/04/01/;148:104375. doi: https://doi.org/10.1016/j.ijmedinf.2021.104375.

23. Fung ICH, Duke CH, Finch KC, Snook KR, Tseng PL, Hernandez AC, et al. Ebola virus disease and social media: A systematic review. Am J Infect Control. 2016 Dec 1;44(12):1660-71. PMID: 27425009. doi: 10.1016/j.ajic.2016.05.011.

24. Organization WH. MERS situation update. 2021 [cited 2022]; Available from: Available at

http://www.emro.who.int/health-topics/mers-cov/mers-outbreaks.html.

25. Omrani AS, Shalhoub S. Middle East respiratory syndrome coronavirus (MERS-CoV): what lessons can we learn? J Hosp Infect. 2015 Nov;91(3):188-96. PMID: 26452615. doi: 10.1016/j.jhin.2015.08.002.

26. Hamasha AA, Alghofaili N, Obaid A, Alhamdan M, Alotaibi A, Aleissa M, et al. Social Media Utilization among Dental Practitioner in Riyadh, Saudi Arabia. The Open Dentistry Journal. 2019;13(1). doi: 10.2174/1874210601913010101.

27. Jamal A, Khan SA, AlHumud A, Al-Duhyyim A, Alrashed M, Bin Shabr F, et al. Association of Online Health Information-Seeking Behavior and Self-Care Activities Among Type 2 Diabetic Patients in Saudi Arabia. Journal of Medical Internet Research. 2015 Aug;17(8). PMID: 26268425. doi: ARTN e196

10.2196/jmir.4312.

28. El Tantawi M, Bakhurji E, Al-Ansari A, AlSubaie A, Al Subaie HA, AlAli A. Indicators of adolescents' preference to receive oral health information using social media. Acta Odontol Scand. 2019 Apr;77(3):213-8. PMID: 30632864. doi: 10.1080/00016357.2018.1536803.

29. Al-Ayyoub M, Khamaiseh AA, Jararweh Y, Al-Kabi MN. A comprehensive survey of arabic sentiment analysis. Information Processing & Management. 2019 2019/03/01/;56(2):320-42. doi: https://doi.org/10.1016/j.ipm.2018.07.006.

30. alOwisheq A, alHumoud S, alTwairesh N, alBuhairi T, editors. Arabic Sentiment Analysis Resources: A Survey. Social Computing and Social Media; 2016 2016//; Cham: Springer International Publishing.

31. Abu Farha I, Magdy W. A comparative study of effective approaches for Arabic sentiment analysis. Information Processing & Management. 2021 2021/03/01/;58(2):102438. doi: https://doi.org/10.1016/j.ipm.2020.102438.

32. Badaro G, Baly R, Hajj H, El-Hajj W, Shaban KB, Habash N, et al. A Survey of Opinion Mining in Arabic: A Comprehensive System Perspective Covering Challenges and Advances in Tools, Resources, Models, Applications, and Visualizations. ACM Trans Asian Low-Resour Lang Inf Process. 2019;18(3):Article 27. doi: 10.1145/3295662.

33. Al-Ayyoub M, Nuseir A, Alsmearat K, Jararweh Y, Gupta B. Deep learning for arabic nlp: a survey. Journal of computational science. 2018;26:522-31. doi: 10.1016/j.jocs.2017.11.011.
34. SalloumSaid A, Mostafa A-E, Azza AM, Khaled S. A survey of text mining in social media facebook and twitter perspectives. Advances in Science, Technology and Engineering Systems Journal. 2017;2:127-33.
35. Hagg E, Dahinten VS, Currie LM. The emerging use of social media for health-related purposes in low and middle-income countries: A scoping review. International Journal of Medical Informatics. 2018 Jul;115:92-105. PMID: 29779724. doi: 10.1016/j.ijmedinf.2018.04.010.
36. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: A systematic review. Journal of Biomedical Informatics. 2020 2020/08/01/;108:103500. doi: https://doi.org/10.1016/j.jbi.2020.103500.
37. Karami A, Lundy M, Webb F, Dwivedi YK. Twitter and Research: A Systematic Literature Review Through Text Mining. IEEE Access. 2020;8:67698-717. doi: 10.1109/ACCESS.2020.2983656.
38. Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognition Letters. 2021 2021/01/01/;141:61-7. doi: https://doi.org/10.1016/j.patrec.2020.07.042.
39. Hussein DME-DM. A survey on sentiment analysis challenges. Journal of King Saud University - Engineering Sciences. 2018 2018/10/01/;30(4):330-8. doi: https://doi.org/10.1016/j.jksues.2016.04.002.
40. Popel M, Tomkova M, Tomek J, Kaiser Ł, Uszkoreit J, Bojar O, et al. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. Nature Communications. 2020 2020/09/01;11(1):4381. doi: 10.1038/s41467-020-18073-9.
41. Alanazi HO, Abdullah AH, Qureshi KN. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. Journal of medical systems. 2017;41(4):69.
42. Simsek Z, Vaara E, Paruchuri S, Nadkarni S, Shaw JD. New Ways of Seeing Big Data. Academy of Management Journal. 2019;62(4):971-8. doi: 10.5465/amj.2019.4004.
43. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health information & libraries journal. 2009;26(2):91-108.
44. Cronin P, Ryan F, Coughlan M. Undertaking a literature review: a step-by-step approach. British journal of nursing. 2008;17(1):38-43.
45. Paré G, Kitsiou S. Methods for literature reviews. Handbook of eHealth Evaluation: An Evidence-based Approach [Internet]: University of Victoria; 2017.
46. Albalawi Y, Carroll, N., & Nikolov, N. S. Opportunities for social media to support health communication in health crises. The 3rd European Technology Assessment Conference; Cork, Ireland2017.
47. Saunders M, Lewis P, Thornhill A, Jenkins M, Bolton D. Critically reviewing the literature. Research methods for business students. Harlow: Pearson Education Limited; 2019. p. 72-119.
48. Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering; London, England, United Kingdom: Association for Computing Machinery; 2014. p. Article 38.
49. Collis J, Hussey R. Business research : a practical guide for undergraduate & postgraduate students2014. ISBN: 9780230301832 0230301835.
50. Reis HT, Reis HT, Judd CM. Handbook of research methods in social and personality psychology: Cambridge University Press; 2000. ISBN: 0521559030.
51. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012;8(1):23-34. PMID: 22833776. doi: 10.20982/tqmp.08.1.p023.

52. Cohen J. Weighted Kappa - Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. Psychol Bull. 1968;70(4):213-&. PMID: 19673146.
53. Fleiss L BL, Myunghee C. The Measurement of Interrater Agreement. Statistical Methods for Rates and Proportions2003. p. 598-626.
54. Bhavsar P, Safro I, Bouaynaya N, Polikar R, Dera D. Machine learning in transportation data analytics. Data analytics for intelligent transportation systems: Elsevier; 2017. p. 283-307.
55. Kotthoff L. Algorithm selection for combinatorial search problems: A survey. Data mining and constraint programming: Springer; 2016. p. 149-90.
56. Beulah M, Pitchai Manickam B. Detection of DDoS Attack Using Ensemble Machine Learning Techniques. Soft Computing for Security Applications: Springer; 2022. p. 889-903.
57. Duwairi R, El-Orfali M. A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. Journal of Information Science. 2014;40(4):501-13. doi: 10.1177/0165551514534143.
58. Abo MEM, Raj RG, Qazi A. A review on arabic sentiment analysis: state-of-the-art, taxonomy and open research challenges. IEEE Access. 2019;7:162008-24. doi: 10.1109/ACCESS.2019.2951530.
59. Forman G. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research. 2003;3(Mar):1289-305.
60. Brownlee J. Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning: Machine Learning Mastery; 2020.
61. Kalyanam J, Velupillai S, Doan S, Conway M, Lanckriet GRG, editors. Facts and Fabrications about Ebola: A Twitter Based Study. ACM KDD Workshop on Connected Health in the Big Data Era; 2015; Sydney, Australia: ACM.
62. Elhadad MK, Li KF, Gebali F, editors. An Ensemble Deep Learning Technique to Detect COVID-19 Misleading Information. Advances in Networked-Based Information Systems; 2021 2021//; Cham: Springer International Publishing.
63. Sicilia R, Giudice SL, Pei Y, Pechenizkiy M, Soda P, editors. Health-related rumour detection on Twitter. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017 13-16 Nov. 2017.
64. Twitter. Data dictionary: Premium v1.1. 2019.
65. Lowe B, Laffey D. Is Twitter for the Birds?: Using Twitter to Enhance Student Learning in a Marketing Course. Journal of Marketing Education. 2011 2011/08/01;33(2):183-92. doi: 10.1177/0273475311410851.
66. Jiang L, Erdem M. Twitter-marketing in multi-unit restaurants: Is it a viable marketing tool? Journal of Foodservice Business Research. 2017 2017/10/20;20(5):568-78. doi: 10.1080/15378020.2016.1222746.
67. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res. 2013 Apr 23;15(4):e85. PMID: 23615206. doi: 10.2196/jmir.1933.
68. Crook B, Glowacki EM, Suran M, Harris JK, Bernhardt JM. Content Analysis of a Live CDC Twitter Chat During the 2014 Ebola Outbreak. Commun Res Rep. 2016;33(4):349-55. doi: 10.1080/08824096.2016.1224171.
69. Wong R, Harris JK, Staub M, Bernhardt JM. Local Health Departments Tweeting About Ebola: Characteristics and Messaging. J Public Health Man. 2017 Mar-Apr;23(2):E16-E24. PMID: 26334537. doi: 10.1097/Phh.0000000000000342.
70. McInnes CJ, Hornmoen H. 'Add Twitter and Stir': The Use of Twitter by Public Authorities in Norway and UK during the 2014-15 Ebola outbreak. Observatorio (OBS*). 2018;12(2). doi: 10.15847/obsOBS12220181173.
71. Levac Jl, O'Sullivan T. Social Media And Its Use In Health Promotion. interdisciplinary journal of health sciences. 2016;1:47–53. doi: 10.18192/riss-ijhs.v1i1.1534.

72. Heldman AB, Schindelar J, Weaver JB. Social Media Engagement and Public Health Communication: Implications for Public Health Organizations Being Truly “Social”. Public Health Reviews. 2013 2013/06/03;35(1):13. doi: 10.1007/BF03391698.
73. Oyeyemi SO, Gabarron E, Wynn R. Ebola, Twitter, and misinformation: a dangerous combination? BMJ. 2014 Oct 14;349:g6178. PMID: 25315514. doi: 10.1136/bmj.g6178.
74. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One. 2010 Nov 29;5(11):e14118. PMID: 21124761. doi: 10.1371/journal.pone.0014118.
75. Gold J, Pedrana AE, Sacks-Davis R, Hellard ME, Chang S, Howard S, et al. A systematic examination of the use of online social networking sites for sexual health promotion. BMC Public Health. 2011 Jul 21;11:583. PMID: 21777470. doi: 10.1186/1471-2458-11-583.
76. Capurro D, Cole K, Echavarria MI, Joe J, Neogi T, Turner AM. The use of social networking sites for public health practice and research: a systematic review. J Med Internet Res. 2014 Mar 14;16(3):e79. PMID: 24642014. doi: 10.2196/jmir.2679.
77. Worldometers. Population by Country. 2019; Available from: https://www.worldometers.info/world-population/population-by-country/.
78. Noha M, Khalil R. Political Islam and Global Media: The Boundaries of Religious Identity: Routledge.
79. Donohue JM, Huskamp HA, Wilson IB, Weissman J. Whom Do Older Adults Trust Most to Provide Information About Prescription Drugs? Am J Geriatr Pharmac. 2009 Apr;7(2):105-16. PMID: 19447363. doi: 10.1016/j.amjopharm.2009.04.005.
80. Gray NJ, Klein JD, Noyce PR, Sesselberg TS, Cantrill JA. Health information-seeking behaviour in adolescence: the place of the internet. Social Science & Medicine. 2005 Apr;60(7):1467-78. PMID: 15652680. doi: 10.1016/j.socscimed.2004.08.010.
81. Hesse BW, Nelson DE, Kreps GL, Croyle RT, Arora NK, Rimer BK, et al. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. Arch Intern Med. 2005 Dec 12-26;165(22):2618-24. PMID: 16344419. doi: 10.1001/archinte.165.22.2618.
82. Medlock S, Eslami S, Askari M, Arts DL, Sent D, de Rooij SE, et al. Health information-seeking behavior of seniors who use the Internet: a survey. Journal of medical Internet research. 2015;17(1):e10-e. PMID: 25574815. doi: 10.2196/jmir.3749.
83. Worsley A. Perceived reliability of sources of health information. Health Education Research. 1989;4(3):367-76. doi: 10.1093/her/4.3.367.
84. Ghenai A, Mejova Y. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. 2017 Ieee International Conference on Healthcare Informatics (Ichi). 2017:518-23. doi: 10.1109/Ichi.2017.58.
85. Yin XX, Han JW, Yu PS. Truth discovery with multiple conflicting information providers on the Web. Ieee T Knowl Data En. 2008 Jun;20(6):796-808. doi: Doi 10.1109/Tkde.2007.190745.
86. Zhao L, Hua T, Lu C-T, Chen I-R. A topic-focused trust model for Twitter. Computer Communications. 2016 2016/02/15/;76:1-11. doi: https://doi.org/10.1016/j.comcom.2015.08.001.
87. Nastasi A, Bryant T, Canner JK, Dredze M, Camp MS, Nagarajan N. Breast Cancer Screening and Social Media: a Content Analysis of Evidence Use and Guideline Opinions on Twitter. J Cancer Educ. 2018 Jun;33(3):695-702. PMID: 28097527. doi: 10.1007/s13187-017-1168-9.
88. Altshuler AL, Storey HLG, Prager SW. Exploring abortion attitudes of US adolescents and young adults using social media. Contraception. 2015 Mar;91(3):226-33. PMID: 25537853. doi: 10.1016/j.contraception.2014.11.009.
89. Bender JL, Jimenez-Marroquin MC, Jadad AR. Seeking Support on Facebook: A Content Analysis of Breast Cancer Groups. Journal of Medical Internet Research. 2011 Jan-Mar;13(1):165-75. PMID: 21371990. doi: ARTN e16

10.2196/jmir.1560.

90. Davis MA, Anthony DL, Pauls SD. Seeking and receiving social support on Facebook for surgery. Social Science & Medicine. 2015 Apr;131:40-7. doi: 10.1016/j.socscimed.2015.02.038.
91. Guidry JPD, Carlyle KE, LaRose JG, Perrin P, Messner M, Ryan M. Using the Health Belief Model to Analyze Instagram Posts about Zika for Public Health Communications. Emerg Infect Dis. 2019 Jan;25(1):179-80. PMID: 30561302. doi: 10.3201/eid2501.180824.
92. Schmidt AL, Zollo F, Scala A, Betsch C, Quattrociocchi W. Polarization of the vaccination debate on Facebook. Vaccine. 2018 Jun 14;36(25):3606-12. PMID: 29773322. doi: 10.1016/j.vaccine.2018.05.040.
93. Salzmann-Erikson M, Hicdurmaz D. Use of Social Media Among Individuals Who Suffer From Post-Traumatic Stress: A Qualitative Analysis of Narratives. Qual Health Res. 2017 Jan;27(2):285-94. PMID: 26825480. doi: 10.1177/1049732315627364.
94. Bhattacharya S, Srinivasan P, Polgreen P. Social media engagement analysis of US Federal health agencies on Facebook. Bmc Med Inform Decis. 2017 Apr 21;17. PMID: 28431582. doi: ARTN 49

10.1186/s12911-017-0447-z.
95. Straton N, Hansen K, Mukkamala RR, Hussain A, Gronli TM, Langberg H, et al. Big Social Data Analytics for Public Health: Facebook Engagement and Performance. 2016 Ieee 18th International Conference on E-Health Networking, Applications and Services (Healthcom). 2016:442-7. doi: 10.1109/HealthCom.2016.7749497.
96. Kim Y, Huang JD, Emery S. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. Journal of Medical Internet Research. 2016 Feb;18(2). PMID: 26920122. doi: ARTN e41

10.2196/jmir.4738.
97. Kalyanam J, Mackey TK. A Review of Digital Surveillance Methods and Approaches to Combat Prescription Drug Abuse. Current Addiction Reports. 2017 December 01;4(4):397-409. doi: 10.1007/s40429-017-0169-4.
98. Chan WSY, Leung AYM. Use of Social Network Sites for Communication Among Health Professionals: Systematic Review. Journal of Medical Internet Research. 2018 Mar;20(3). PMID: 29592845. doi: ARTN e117

10.2196/jmir.8382.
99. Fu KW, Liang H, Saroha N, Tse ZTH, Ip P, Fung ICH. How people react to Zika virus outbreaks on Twitter? A computational content analysis. Am J Infect Control. 2016 Dec 1;44(12):1700-2. PMID: 27566874. doi: 10.1016/j.ajic.2016.04.253.
100. Glowacki EM, Lazard AJ, Wilcox GB, Mackert M, Bernhardt JM. Identifying the public's concerns and the Centers for Disease Control and Prevention's reactions during a health crisis: An analysis of a Zika live Twitter chat. Am J Infect Control. 2016 Dec 1;44(12):1709-11. PMID: 27544795. doi: 10.1016/j.ajic.2016.05.025.
101. Young R, Tully M, Dalrymple KE. Engagement: use of Twitter chats to construct nominal participatory spaces during health crises. Inform Commun Soc. 2018;21(4):499-515. doi: 10.1080/1369118x.2017.1301518.
102. Miller M, Banerjee T, Muppalla R, Romine W, Sheth A. What Are People Tweeting About Zika? An Exploratory Study Concerning Its Symptoms, Treatment, Transmission, and Prevention. JMIR Public Health Surveill. 2017;3(2):e38. PMID: 28630032. doi: 10.2196/publichealth.7157.
103. Juric R, Kim I, Panneerselvam H, Tesanovic I, editors. Analysis of ZIKA Virus Tweets: Could Hadoop Platform Help in Global Health Management? The 50th Hawaii International Conference on System Sciences; 2017.
104. Vos SC, Buckner MM. Social Media Messages in an Emerging Health Crisis: Tweeting Bird Flu. Journal of Health Communication. 2016 Mar 3;21(3):301-8. PMID: 26192209. doi: 10.1080/10810730.2015.1064495.

105. Bhattacharya S, Srinivasan P, Polgreen P. Engagement with Health Agencies on Twitter. Plos One. 2014 Nov 7;9(11). PMID: 25379727. doi: 10.1371/journal.pone.0112235.
106. Neiger BL, Thackeray R, Van Wagenen SA, Hanson CL, West JH, Barnes MD, et al. Use of Social Media in Health Promotion: Purposes, Key Performance Indicators, and Evaluation Metrics. Health Promotion Practice. 2012 2012/03/01;13(2):159-64. PMID: 22382491. doi: 10.1177/1524839911433467.
107. Dumbrell D, Steele R, editors. Twitter and Health in the Australian Context: What Types of Information Are Health-Related Organizations Tweeting? 2013 46th Hawaii International Conference on System Sciences; 2013 7-10 Jan. 2013.
108. Passerello GL, Hazelwood JE, Lawrie S. Using Twitter to assess attitudes to schizophrenia and psychosis. BJPsych Bulletin. 2019:1-9. PMID: 30784393. doi: 10.1192/bjb.2018.115.
109. Meadows CZ, Tang L, Liu W. Twitter message types, health beliefs, and vaccine attitudes during the 2015 measles outbreak in California. Am J Infect Control. 2019 Jun 29. PMID: 31266661. doi: 10.1016/j.ajic.2019.05.007.
110. Chung JE. A Smoking Cessation Campaign on Twitter: Understanding the Use of Twitter and Identifying Major Players in a Health Campaign. J Health Commun. 2016 May;21(5):517-26. PMID: 27115179. doi: 10.1080/10810730.2015.1103332.
111. Stautz K, Bignardi G, Hollands GJ, Marteau TM. Reactions on Twitter to updated alcohol guidelines in the UK: a content analysis. BMJ Open. 2017 Feb 28;7(2):e015493. PMID: 28246145. doi: 10.1136/bmjopen-2016-015493.
112. Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, et al. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. Am J Public Health. 2018 Oct;108(10):1378-84. PMID: 30138075. doi: 10.2105/AJPH.2018.304567.
113. Allem JP, Ferrara E, Uppu SP, Cruz TB, Unger JB. E-Cigarette Surveillance With Social Media Data: Social Bots, Emerging Topics, and Trends. JMIR Public Health Surveill. 2017 Dec 20;3(4):e98. PMID: 29263018. doi: 10.2196/publichealth.8641.
114. Jamison AM, Broniatowski DA, Quinn SC. Malicious Actors on Twitter: A Guide for Public Health Researchers. Am J Public Health. 2019 May;109(5):688-92. PMID: 30896994. doi: 10.2105/AJPH.2019.304969.
115. Yang K-C, Varol O, Davis CA, Ferrara E, Flammini A, Menczer F. Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies. 2019;1(1):48-61. doi: 10.1002/hbe2.115.
116. Howard PN, Woolley S, Calo R. Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. J Inf Technol Politi. 2018;15(2):81-93. doi: 10.1080/19331681.2018.1448735.
117. Edwards C, Beattie AJ, Edwards A, Spence PR. Differences in perceptions of communication quality between a Twitterbot and human agent for information seeking and learning. Comput Hum Behav. 2016 Dec;65:666-71. doi: 10.1016/j.chb.2016.07.003.
118. Albalawi Y, Sixsmith J. Identifying Twitter influencer profiles for health promotion in Saudi Arabia. Health Promot Int. 2017 Jun 1;32(3):456-63. PMID: 26516181. doi: 10.1093/heapro/dav103.
119. *Social Authority*. [2019-09-30]; Available from: https://followerwonk.com/social-authority.
120. *PeerIndex*. [updated 2019-09-30]; Available from: http://peerindex.com.
121. *Kred*. [2019-09-30]; Available from: https://home.kred/.
122. *Klou*. [2019-09-30]; Available from: https://khoros.com/platform/klout.
123. Suh B, Hong L, Pirolli P, Chi EH, editors. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. 2010 IEEE Second International Conference on Social Computing; 2010 20-22 Aug. 2010.

124. Kim AE, Hansen HM, Murphy J, Richards AK, Duke J, Allen JA. Methodological considerations in analyzing Twitter data. J Natl Cancer Inst Monogr. 2013 Dec;2013(47):140-6. PMID: 24395983. doi: 10.1093/jncimonographs/lgt026.
125. Graham M, Hale SA, Gaffney D. Where in the World Are You? Geolocation and Language Identification in Twitter. The Professional Geographer. 2014 2014/10/02;66(4):568-78. doi: 10.1080/00330124.2014.907699.
126. Sloan L, Morgan J. Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. PLoS One. 2015;10(11):e0142209. PMID: 26544601. doi: 10.1371/journal.pone.0142209.
127. Twitter. Search Tweets. 2015; Available from: [https://developer.twitter.com/en/docs/tweets/search/api-reference](https://developer.twitter.com/en/docs/tweets/search/api-reference) Archieved at: [http://www.webcitation.org/74byAfJy0](http://www.webcitation.org/74byAfJy0)
128. Bobicev V, Sokolova M. Thumbs Up and Down: Sentiment Analysis of Medical Online Forums. Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task: Association for Computational Linguistics; 2018.
129. Team RC. R: A language and environment for statistical computing. Vienna, Austria, 2019.
130. Castillo C, Mendoza M, Poblete B. Information credibility on twitter. Proceedings of the 20th international conference on World wide web; Hyderabad, India. 1963500: ACM; 2011. p. 675-84.
131. Kim A, Miano T, Chew R, Eggers M, Nonnemaker J. Classification of Twitter Users Who Tweet About E-Cigarettes. JMIR Public Health Surveill. 2017;3(3):e63. PMID: 28951381. doi: 10.2196/publichealth.8060.
132. Morris MR, Counts S, Roseway A, Hoff A, Schwarz J. Tweeting is believing?: understanding microblog credibility perceptions. Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work; Seattle, Washington, USA. 2145274: ACM; 2012. p. 441-50.
133. Zhuang KC, Shen HB, Zhang H. User spread influence measurement in microblog. Multimed Tools Appl. 2017 Feb;76(3):3169-85. doi: 10.1007/s11042-016-3818-z.
134. Riquelme F, Gonzalez-Cantergiani P. Measuring user influence on Twitter: A survey. Information Processing & Management. 2016 Sep;52(5):949-75. doi: 10.1016/j.ipm.2016.04.003.
135. Neamat El Gayar CYS. Computational Linguistics, Speech and Image Processing for Arabic Language: WORLD SCIENTIFIC; 2017. 288 p. ISBN: 978-981-322-938-9.
136. Al-Nazer A, Gutub A. Exploit Kashida Adding to Arabic e-Text for High Capacity Steganography. Nss: 2009 3rd International Conference on Network and System Security. 2009:447-+. doi: 10.1109/NSS.2009.21.
137. Volkova S, Van Durme B. Online Bayesian Models for Personal Analytics in Social Media. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; Austin, Texas2015. p. 2325-31.
138. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One. 2013;8(9):e73791. PMID: 24086296. doi: 10.1371/journal.pone.0073791.
139. An L, Yi X, Han Y, Li G. An Influence Prediction Model for Microblog Entries on Public Health Emergencies. Data and Information Management. 2019 2019/01/17;0(0). doi: 10.2478/dim-2018-0013.
140. Prabhu V, Lee T, Loeb S, Holmes JH, Gold HT, Lepor H, et al. Twitter response to the United States Preventive Services Task Force recommendations against screening with prostate-specific antigen. BJU International. 2014 2015/07/01;116(1):65-71. PMID: 24661474. doi: 10.1111/bju.12748.
141. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract. 2004 May;10(2):307-12. PMID: 15189396. doi: 10.1111/j.2002.384.doc.x.

142. de Winter JCF, Dodou D. Five-point likert items: T test versus Mann-Whitney-Wilcoxon. Practical Assessment, Research and Evaluation. 2010;15(11).
143. Mehta CR, Patel NR. ALGORITHM 643: FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered rxc contingency tables. ACM Trans Math Softw. 1986;12(2):154-61. doi: 10.1145/6497.214326.
144. Clarkson DB, Fan Y-a, Joe H. A remark on algorithm 643: FEXACT: an algorithm for performing Fisher's exact test in r x c contingency tables. ACM Trans Math Softw. 1993;19(4):484-8. doi: 10.1145/168173.168412.
145. Yoo E, Rand W, Eftekhar M, Rabinovich E. Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. Journal of Operations Management. 2016;45:123-33. doi: 10.1016/j.jom.2016.05.007.
146. Boudad N, Faizi R, Thami ROH, Chiheb R. Sentiment classification of Arabic tweets: A supervised approach. Journal of Mobile Multimedia. 2017;13(3-4):233-43.
147. Oueslati O, Cambria E, HajHmida MB, Ounelli H. A review of sentiment analysis research in Arabic language. Future Generation Computer Systems. 2020 2020/11/01/;112:408-30. doi: https://doi.org/10.1016/j.future.2020.05.034.
148. Alabbas W, al-Khateeb HM, Mansour A, Epiphaniou G, Frommholz I. Classification of colloquial arabic tweets in real-time to detect high-risk floods. International Conference on Social Media, Wearable and Web Analytics2017. p. 1-8.
149. Alayba AM, Palade V, England M, Iqbal R. Arabic Language Sentiment Analysis on Health Services. 1st International Workshop on Arabic Script Analysis and Recognition2017. p. 114-8.
150. Mohaouchane H, Mourhir A, Nikolov NS, editors. Detecting Offensive Language on Arabic Social Media Using Deep Learning. 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS); 2019 22-25 Oct. 2019.
151. Abdullah M, AlMasawa M, Makki I, Alsolmi M, Mahrous S. Emotions extraction from Arabic tweets. International Journal of Computers and Applications. 2018:661-75. doi: 10.1080/1206212X.2018.1482395.
152. Soliman AB, Eissa K, El-Beltagy SR. AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. Procedia Computer Science. 2017 2017/01/01/;117:256-65. doi: https://doi.org/10.1016/j.procs.2017.10.117.
153. Fouad MM, Mahany A, Aljohani N, Abbasi RA, Hassan S-U. ArWordVec: efficient word embedding models for Arabic tweets. Soft Computing. 2020 2020/06/01;24(11):8061-8. doi: 10.1007/s00500-019-04153-6.
154. Farha IA, Magdy W, editors. Mazajak: An online Arabic sentiment analyser. Proceedings of the Fourth Arabic Natural Language Processing Workshop; 2019.
155. Kanan T, Sadaqa O, Aldajeh A, Alshwabka H, dolime WA-, AlZu'bi S, et al., editors. A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT); 2019 9-11 April 2019.
156. Alrifai K, Rebdawi G, Ghneim N. Arabic tweeps gender and dialect prediction: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings2017. p. 1-9.
157. HaCohen-Kerner Y, Yigal Y, Shayovitz E, Miller D, Breckon T, editors. Author profiling: Gender prediction from Tweets and images: Notebook for PAN at CLEF 2018. CEUR Workshop Proceedings; 2018.
158. Alrifai K, Rebdawi G, Ghneim N. Comparison of machine learning approaches in arabic tweeps gender prediction. International Journal of Scientific and Technology Research. 2019;8(11):2892-5.
159. Alkhair M, Meftouh K, Smaïli K, Othman N. An Arabic Corpus of Fake News: Collection, Analysis and Classification. Communications in Computer and Information Science2019. p. 292-302.

160. Kaliyar RK, Goswami A, Narang P, Sinha S. FNDNet – A deep convolutional neural network for fake news detection. Cognitive Systems Research. 2020 2020/06/01/;61:32-44. doi: https://doi.org/10.1016/j.cogsys.2019.12.005.
161. Albalawi Y, Nikolov NS, Buckley J. Trustworthy health-related tweets on social media in Saudi Arabia: tweet metadata analysis. Journal of medical internet research. 2019;21(10):e14731. doi: 10.2196/14731.
162. Park HW, Park S, Chong M. Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea. Journal of Medical Internet Research. 2020;22(5):e18897. doi: 10.2196/18897.
163. Lazard AJ, Scheinfeld E, Bernhardt JM, Wilcox GB, Suran M. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. Am J Infect Control. 2015;43(10):1109-11. doi: 10.1016/j.ajic.2015.05.025.
164. Paul MJ, Dredze M. Discovering health topics in social media using topic models. PloS one. 2014;9(8):e103408. doi: 10.1371/journal.pone.0103408.
165. Tuarob S, Tucker CS, Salathe M, Ram N. Discovering health-related knowledge in social media using ensembles of heterogeneous features. Proceedings of the 22nd ACM international conference on Information & Knowledge Management; San Francisco, California, USA2013. p. 1685-90.
166. Maw M, Balakrishnan V, Rana O, Ravana SD. Trends and Patterns of Text Classification Techniques: A Systematic Mapping Study. Malaysian Journal of Computer Science. 2020;33(2):102-17. doi: 10.22452/mjcs.vol33no2.2.
167. Al-Rubaiee H, Qiu RX, Li DY. Identifying mubasher software products through sentiment analysis of arabic tweets. 2016 International Conference on Industrial Informatics and Computer Systems2016.
168. Castillo P, Habib M, Aljarah I, Faris H. Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context. Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods2020. p. 453-60.
169. Albadi N, Kurdi M, Mishra S, editors. Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2018 28-31 Aug. 2018.
170. Alakrot A, Murray L, Nikolov NS. Towards Accurate Detection of Offensive Language in Online Communication in Arabic. Procedia Computer Science2018. p. 315-20.
171. Shannag FB, Hammo BH. Lessons Learned from Event Detection from Arabic Tweets: The Case of Jordan Flash Floods near Dead Sea. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT); 9-11 April 20192019. p. 806-11.
172. Oussous A, Lahcen AA, Belfkih S. Impact of Text Pre-processing and Ensemble Learning on Arabic Sentiment Analysis. Proceedings of the 2nd International Conference on Networking, Information Systems & Security; Rabat, Morocco: Association for Computing Machinery; 2019. p. 1-9.
173. Abdulla NA, Ahmed NA, Shehab MA, Al-Ayyoub M, Al-Kabi MN, Al-rifai S. Towards improving the lexicon-based approach for arabic sentiment analysis. International Journal of Information Technology and Web Engineering. 2014;9(3):55-71. doi: 10.4018/ijitwe.2014070104.
174. Oussous A, Benjelloun F-Z, Lahcen AA, Belfkih S. ASA: A framework for Arabic sentiment analysis. Journal of Information Science. 2019 2020/08/01;46(4):544-59. doi: 10.1177/0165551519849516.
175. Mikolov T, Chen K, Corrado G, Dean J, editors. Efficient estimation of word representations in vector space. In: ICLR: Proceeding of the International Conference on Learning Representations Workshop; 2013; Arizona, USA.
176. Sultana J, Usha Rani M, Farquad MAH. An Extensive Survey on Some Deep-Learning Applications. In: Venkata Krishna P, Obaidat MS, editors. Emerging Research in Data Engineering Systems and Computer Communications; Singapore: Springer Singapore; 2020. p. 511-9.

177. Heikal M, Torki M, El-Makky N. Sentiment Analysis of Arabic Tweets using Deep Learning. Procedia Computer Science2018. p. 114-22.
178. Li H, Caragea D, Li X, Caragea C. Comparison of word embeddings and sentence encodings as generalized representations for crisis Tweet classification tasks. Innovating for Resilience - 1st International Conference on Information Systems for Crisis Response and Management Asia Pacific; Auckland, New Zealand: Massey Univeristy2018. p. 480-93.
179. Abuzayed A, Elsayed T. Quick and simple approach for detecting hate speech in arabic tweets. Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection; may; Marseille, France: European Language Resource Association; 2020. p. 109-14.
180. Guellil I, Saâdane H, Azouaou F, Gueni B, Nouvel D. Arabic natural language processing: An overview. Journal of King Saud University-Computer and Information Sciences. 2019:497-507. doi: 10.1016/j.jksuci.2019.02.006.
181. Alwehaibi A, Roy K, editors. Comparison of Pre-Trained Word Vectors for Arabic Text Classification Using Deep Learning Approach. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); 2018 17-20 Dec. 2018.
182. Altowayan AA, Tao L. Word embeddings for Arabic sentiment analysis. Proceedings - 2016 IEEE International Conference on Big Data, Big Data 20162016. p. 3820-5.
183. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T, editors. Learning Word Vectors for 157 Languages. 2018 may; Miyazaki, Japan: European Language Resources Association (ELRA).
184. Abdullah M, Hadzikadicy M, Shaikhz S. SEDAT: Sentiment and emotion detection in arabic text Using cnn-lstm deep learning. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); 17-20 Dec. 20182018. p. 835-40.
185. Alali M, Sharef NM, Murad MAA, Hamdan H, Husin NA. Narrow Convolutional Neural Network for Arabic Dialects Polarity Classification. IEEE Access. 2019;7:96272-83. doi: 10.1109/ACCESS.2019.2929208.
186. Ghallab A, Mohsen A, Ali Y. Arabic Sentiment Analysis: A Systematic Literature Review. Applied Computational Intelligence and Soft Computing. 2020;2020. doi: 10.1155/2020/7403128.
187. Lulu L, Elnagar A. Automatic Arabic Dialect Classification Using Deep Learning Models. In: Shaalan K, ElBeltagy SR, editors. Arabic Computational Linguistics2018. p. 262-9.
188. Srinivasulu K. Health-Related Tweets Classification: A Survey. Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: Springer; 2021. p. 259-68.
189. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006. xx, 738 p. p. ISBN: 9781493938438

0387310738.
190. Chung Y, Haas PJ, Upfal E, Kraska T. Unknown Examples & Machine Learning Model Generalization. ArXiv. 2018;abs/1808.08294.
191. Collis J, Hussey R. Business Research: A Practical Guide for Undergraduate and Postgraduate Students. London: Macmillan Education UK; 2014. ISBN: 0230301835.
192. Collier N, Goodwin RM, McCrae J, Doan S, Kawazoe A, Conway M, et al. An ontology-driven system for detecting global health events. Proceedings of the 23rd International Conference on Computational Linguistics; Beijing, China: Association for Computational Linguistics; 2010. p. 215–22.
193. Hicks A, Hogan WR, Rutherford M, Malin B, Xie M, Fellbaum C, et al. Mining Twitter as a First Step toward Assessing the Adequacy of Gender Identification Terms on Intake Forms. Annual Symposium proceedings; 20152015. p. 611-20.
194. Pruss D, Fujinuma Y, Daughton AR, Paul MJ, Arnot B, Albers Szafir D, et al. Zika discourse in the Americas: A multilingual topic analysis of Twitter. PloS one. 2019;14(5):e0216922. doi: 10.1371/journal.pone.0216922.

195. Zhang Z, Ahmed W. A comparison of information sharing behaviours across 379 health conditions on Twitter. International Journal of Public Health. 2019 2019/04/01;64(3):431-40. doi: 10.1007/s00038-018-1192-5.
196. Shoukry A, Rafea A. Sentence-level Arabic sentiment analysis. 2012 International Conference on Collaboration Technologies and Systems (CTS); 21-25 May 20122012. p. 546-50.
197. Roccetti M, Delnevo G, Casini L, Mirri S. An alternative approach to dimension reduction for pareto distributed data: a case study. Journal of big Data. 2021;8(1):1-23. doi: 10.1186/s40537-021-00428-8.
198. Sun Y, Wong AKC, Kamel MS. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence. 2009;23(04):687-719. doi: 10.1142/s0218001409007326.
199. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl. 2004;6(1):20–9. doi: 10.1145/1007730.1007735.
200. Sun A, Lim E-P, Liu Y. On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems. 2009 2009/12/01/;48(1):191-201. doi: https://doi.org/10.1016/j.dss.2009.07.011.
201. Somasundaram A, Reddy US. Data imbalance: Effects and solutions for classification of large and highly imbalanced data. Proceedings of the of 1st international conference on research in engineering, computers and technology2016. p. 1-16.
202. Twitter. Developer agreement and policy. 2020 [updated 10th March 2020; cited 2021 06th March]; Available from: https://developer.twitter.com/en/developer-terms/agreement-and-policy.
203. Barnaghi P, Ghaffari P, Breslin JG. Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment. 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService); 29 March-1 April 20162016. p. 52-7.
204. Abdellaoui H, Zrigui M. Using Tweets and Emojis to Build TEAD: an Arabic Dataset for Sentiment Analysis. Computacion Y Sistemas. 2018;22(3):777-86. PMID: WOS:000471005100007. doi: 10.13053/CyS-22-3-3031.
205. Abo MEM, Shah NAK, Balakrishnan V, Kamal M, Abdelaziz A, Haruna K, et al. SSA-SDA: Subjectivity and sentiment analysis of sudanese dialect arabic. International Conference on Computer and Information Sciences2019. p. 206-10.
206. Abozinadah EA, Jones JH, Jr. A statistical learning approach to detect abusive twitter accounts. ACM International Conference Proceeding Series2017. p. 6-13.
207. Hussien WA, Tashtoush YM, Al-Ayyoub M, Al-Kabi MN. Are Emoticons Good Enough to Train Emotion Classifiers of Arabic Tweets? 2016 7th International Conference on Computer Science and Information Technology2016. p. 1-6.
208. Mahmoud A, Elghazaly T. Using twitter to monitor political sentiment for Arabic slang. Studies in Computational Intelligence2018. p. 53-66.
209. Almanie T, Aldayel A, Alkanhal G, Alesmail L, Almutlaq M, Althunayan R. Saudi Mood: A Real-Time Informative Tool for Visualizing Emotions in Saudi Arabia Using Twitter2018. ISBN: 978-1-5386-4110-1.
210. Abuelenin S, Elmougy S, Naguib E. Twitter sentiment analysis for arabic tweets. Advances in Intelligent Systems and Computing2018. p. 467-76.
211. Alzu'bi S, Badarneh O, Hawashin B, Al-Ayyoub M, Alhindawi N, Jararweh Y. Multi-Label Emotion Classification for Arabic Tweets. Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)2019. p. 499-504.
212. Samy AE, El-Beltagy SR, Hassanien E. A Context Integrated Model for Multi-label Emotion Detection. Procedia Computer Science2018. p. 61-71.

213. Symeonidis S, Effrosynidis D, Arampatzis A. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications. 2018 2018/11/15/;110:298-310. doi: https://doi.org/10.1016/j.eswa.2018.06.022.
214. Baali M, Ghneim N. Emotion analysis of Arabic tweets using deep learning approach. Journal of Big Data. 2019;6(1). doi: 10.1186/s40537-019-0252-x.
215. Ismail R, Omer M, Tabir M, Mahadi N, Amin I. Sentiment Analysis for Arabic Dialect Using Supervised Learning. International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE); 12-14 Aug. 20182018. p. 1-6.
216. Nayel HA, Medhat W, Rashad M. BENHA@IDAT: Improving irony detection in Arabic tweets using ensemble approach. CEUR Workshop Proceedings2019. p. 401-8.
217. Alqarafi A, Adeel A, Hawalah A, Swingler K, Hussain A. A Semi-supervised Corpus Annotation for Saudi Sentiment Analysis Using Twitter. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)2018. p. 589-96.
218. Al-Saif H, Al-Dossari H. Detecting and classifying crimes from arabic twitter posts using text mining techniques. International Journal of Advanced Computer Science and Applications. 2018 Oct;9(10):377-87. PMID: WOS:000454189600046. doi: 10.14569/ijacsa.2018.091046.
219. Maamouri M, Bies A, Kulick S. Diacritization: A challenge to Arabic treebank annotation and parsing. Proceedings of the Conference of the Machine Translation SIG of the British Computer Society2006. p. 35-47.
220. El Gayar N, Suen CY. Computational Linguistics, Speech and Image Processing for Arabic Language: World Scientific; 2019. ISBN: 9813229381.
221. Oussous A, Lahcen AA, Belfkih S. Improving Sentiment Analysis of Moroccan Tweets Using Ensemble Learning. In: Tabii Y, Lazaar M, AlAchhab M, Enneya N, editors. Big Data, Cloud and Applications, Bdca 20182018. p. 91-104.
222. Oumi NA, Safadi LA, Chorfi H, editors. Mining Tweets to Indicate Hidden/Potential Networks. 2018 21st Saudi Computer Society National Computer Conference (NCC); 2018 25-26 April 2018.
223. Abdelhade N, Soliman THA, Ibrahim HM. Detecting twitter users' opinions of arabic comments during various time episodes via deep neural network. Advances in Intelligent Systems and Computing2018. p. 232-46.
224. Al-Wehaibi RN, Khan MB. Predicting arabic tweet popularity by use of data and text mining techniques. MEDES 2014 - 6th International Conference on Management of Emergent Digital EcoSystems, Proceedings2014. p. 183-9.
225. Habash N, Soudi A, Buckwalter T. On Arabic Transliteration. In: Soudi A, Bosch Avd, Neumann G, editors. Arabic Computational Morphology: Knowledge-based and Empirical Methods. Dordrecht: Springer Netherlands; 2007. p. 15-22.
226. Elayeb B, Bounhas I. Arabic Cross-Language Information Retrieval: A Review. ACM Trans Asian Low-Resour Lang Inf Process. 2016;15(3):1-44. doi: 10.1145/2789210.
227. Xu J, Fraser A, Weischedel R. Empirical studies in strategies for Arabic retrieval. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval; Tampere, Finland: Association for Computing Machinery; 2002. p. 269–74.
228. Wang H, Hanafy A, Bahgat M, Noeman S, Emam OS, Bommireddipalli VR, editors. A System for Extracting Sentiment from Large-Scale Arabic Social Data. 2015 First International Conference on Arabic Computational Linguistics (ACLing); 2015 17-20 April 2015.
229. Alahmary RM, Al-Dossari HZ, Emam AZ, Ieee. Sentiment Analysis of Saudi Dialect Using Deep Learning Techniques. 2019 International Conference on Electronics, Information, and Communication2019. p. 562-7.
230. Alsaif H, Alotaibi T. Arabic Text Classification using Feature-Reduction Techniques for Detecting Violence on Social Media. International Journal of Advanced Computer Science and

Applications. 2019 Apr;10(4):77-87. PMID: WOS:000467916400010. doi: 10.14569/ijacsa.2019.0100409.
231. Al-Twairesh N, Al-Negheimish H. Surface and deep features ensemble for sentiment analysis of arabic tweets. IEEE Access. 2019;7:84122-31. doi: 10.1109/ACCESS.2019.2924314.
232. El-Fishawy N, Hamouda A, Attiya GM, Atef M. Arabic summarization in Tw itter social network. Ain Shams Engineering Journal. 2014;5(2):411-20. doi: 10.1016/j.asej.2013.11.002.
233. El-Naggar N, El-Sonbaty Y, El-Nasr MA, editors. Sentiment analysis of modern standard Arabic and Egyptian dialectal Arabic tweets. 2017 Computing Conference; 2017 18-20 July 2017.
234. Al-Osaimi S, Badruddin KM. Role of emotion icons in sentiment classification of arabic tweets. MEDES 2014 - 6th International Conference on Management of Emergent Digital EcoSystems, Proceedings2014. p. 167-71.
235. Aldayel HK, Azmi AM. Arabic tweets sentiment analysis - a hybrid scheme. Journal of Information Science. 2016 Dec;42(6):782-97. PMID: WOS:000389709500004. doi: 10.1177/0165551515610513.
236. Atoum JO, Nouman M. Sentiment Analysis of Arabic Jordanian Dialect Tweets. International Journal of Advanced Computer Science and Applications. 2019 Feb;10(2):256-62. PMID: WOS:000463078000035. doi: 10.14569/IJACSA.2019.0100234.
237. Elshakankery K, Ahmed MF. HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis. Egyptian Informatics Journal. 2019;20(3):163-71. doi: 10.1016/j.eij.2019.03.002.
238. Al-Smadi M, Jaradat Z, Al-Ayyoub M, Jararweh Y. Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. Information Processing & Management. 2017 May;53(3):640-52. PMID: WOS:000396972300006. doi: 10.1016/j.ipm.2017.01.002.
239. Zerrouki T. Tashaphyne, Arabic light stemmer. 2020; Available from: https://pypi.org/project/Tashaphyne/.
240. Elghazaly T, Mahmoud A, Hefny HA. Political sentiment analysis using twitter data. ACM International Conference Proceeding Series2016. p. 1-5.
241. Nizar YH. Introduction to Arabic Natural Language Processing: Morgan & Claypool; 2010. 1 p. ISBN: 9781598297966.
242. Zerrouki T. Towards An Open Platform For Arabic Language Processing: École nationale supérieure d'informatique; 2020.
243. Alhaj YA, Al-qaness MA, Dahou A, Elaziz MA, Zhao D, Xiang J. Effects of Light Stemming on Feature Extraction and Selection for Arabic Documents Classification. Recent Advances in NLP: The Case of Arabic Language: Springer; 2020. p. 59-79.
244. Srinivasa-Desikan B. Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras: Packt Publishing Ltd; 2018. ISBN: 1788837037.
245. Junejo KN, Karim A, Hassan MT, Jeon M. Terms-based discriminative information space for robust text classification. Information Sciences. 2016;372:518-38. doi: 10.1016/j.ins.2016.08.073.
246. Al-Horaibi L, Khan MB. Sentiment analysis of arabic tweets using text mining techniques. Proceedings of SPIE - The International Society for Optical Engineering2016. p. 288-92.
247. Raj RJR, Srinivasulu S, Ashutosh A, editors. A Multi-classifier Framework for Detecting Spam and Fake Spam Messages in Twitter. 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT); 2020 10-12 April 2020.
248. Gallagher C, Furey E, Curran K. The application of sentiment analysis and text analytics to customer experience reviews to understand what customers are really saying. International Journal of Data Warehousing and Mining (IJDWM). 2019;15(4):21-47. doi: 10.4018/ijdwm.2019100102.

249. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011;2(3):1-27. doi: 10.1145/1961189.1961199.
250. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A library for large linear classification. Journal of machine learning research. 2008;9(Aug):1871-4. doi: 10.1145/1390681.1442794.
251. Al-Azani S, El-Alfy ESM. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. Procedia Computer Science2017. p. 359-66.
252. Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Science Advances. 2018;4(1). doi: 10.1126/sciadv.aao5580.
253. Raschka S, Mirjalili V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition: Packt Publishing; 2017. ISBN: 1787125939.
254. Kuhn M, Johnson K. Applied predictive modeling: Springer; 2013.
255. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
256. Yenigalla P, Kar S, Singh C, Nagar A, Mathur G. Addressing unseen word problem in text classification. 2018: Springer. p. 339-51.
257. Elekes Á, Di Stefano AS, Schäler M, Böhm K, Keller M. Learning from few samples: Lexical substitution with word embeddings for short text classification. 2019: IEEE. p. 111-9.
258. Gulli A, Kapoor A, Pal S. Deep Learning with TensorFlow 2 and Keras: Regression, ConvNets, GANs, RNNs, NLP, and More with TensorFlow 2 and the Keras API: Packt Publishing, Limited; 2019. ISBN: 1838827722.
259. Soufan A. Deep Learning for Sentiment Analysis of Arabic Text. Proceedings of the ArabWIC 6th Annual International Conference Research Track; Rabat, Morocco: Association for Computing Machinery; 2019. p. 1-8.
260. Elnagar A, Al-Debsi R, Einea O. Arabic text classification using deep learning models. Information Processing & Management. 2020 2020/01/01/;57(1). doi: https://doi.org/10.1016/j.ipm.2019.102121.
261. Hutter F, Lücke J, Schmidt-Thieme L. Beyond Manual Tuning of Hyperparameters. KI - Künstliche Intelligenz. 2015 2015/11/01;29(4):329-37. doi: 10.1007/s13218-015-0381-0.
262. Feurer M, Hutter F. Hyperparameter Optimization. In: Hutter F, Kotthoff L, Vanschoren J, editors. Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer International Publishing; 2019. p. 3-33.
263. O'Malley T, Bursztein E, Long J, Chollet F, Jin H, Invernizzi L. Hyperparameter tuning with Keras Tuner. 2019; Available from: https://github.com/keras-team/keras-tuner.
264. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference2019. p. 4171-86.
265. Antoun W, Baly F, Hajj H. AraBERT: Transformer-based Model for Arabic Language Understanding. may; Marseille, France: European Language Resource Association; 2020. p. 9-15.
266. Vraga EK, Bode L. Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation. Political Communication. 2020 2020/01/02;37(1):136-44. doi: 10.1080/10584609.2020.1716500.
267. Al-Twairesh N. The Evolution of Language Models Applied to Emotion Analysis of Arabic Tweets. Information. 2021;12(2). doi: 10.3390/info12020084.
268. El-Alami F-z, Ouatik El Alaoui S, En Nahnahi N. Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. Journal of King Saud University - Computer and Information Sciences. 2021 2021/02/18/. doi: https://doi.org/10.1016/j.jksuci.2021.02.005.

269. El-Razzaz M, Fakhr MW, Maghraby FA. Arabic Gloss WSD Using BERT. Applied Sciences. 2021;11(6):2567. PMID: doi:10.3390/app11062567.
270. Gómez-Pérez J, Denaux R, Garcia-Silva A. A Practical Guide to Hybrid Natural Language Processing: Combining Neural Models and Knowledge Graphs for NLP. A Practical Guide to Hybrid Natural Language Processing. 2020.
271. Saeed F, M.S W, Al-Sarem M. Detecting Health-Related Rumors on Twitter using Machine Learning Methods. International Journal of Advanced Computer Science and Applications. 2020 01/01;11. doi: 10.14569/IJACSA.2020.0110842.
272. Ott BL. The age of Twitter: Donald J. Trump and the politics of debasement. Critical Studies in Media Communication. 2017 2017/01/01;34(1):59-68. doi: 10.1080/15295036.2016.1266686.
273. Tang Y, Hew KF. Using Twitter for education: Beneficial or simply a waste of time? Computers & Education. 2017 2017/03/01/;106:97-118. doi: 10.1016/j.compedu.2016.12.004.
274. Justinia T, Alyami A, Al-Qahtani S, Bashanfar M, El-Khatib M, Yahya A, et al. Social Media and the Orthopaedic Surgeon: a Mixed Methods Study. Acta Inform Med. 2019 Mar;27(1):23-8. PMID: 31213739. doi: 10.5455/aim.2019.27.23-28.
275. Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. Information Processing & Management. 2021 2021/01/01/;58(1):102390. doi: 10.1016/j.ipm.2020.102390.
276. Al-Rakhami MS, Al-Amri AM. Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter. IEEE Access. 2020;8:155961-70. doi: 10.1109/ACCESS.2020.3019600.
277. Alhaddad MS. The use of social media among Saudi residents for medicines related information. Saudi Pharmaceutical Journal. 2018 2018/12/01/;26(8):1106-11. doi: 10.1016/j.jsps.2018.05.021.
278. Chou W-YS, Oh A, Klein WM. Addressing health-related misinformation on social media. Jama. 2018;320(23):2417-8. doi: 10.1001/jama.2018.16865.
279. Tan ASL, Lee C-j, Chae J. Exposure to Health (Mis)Information: Lagged Effects on Young Adults' Health Behaviors and Potential Pathways. Journal of Communication. 2015;65(4):674-98. doi: 10.1111/jcom.12163.
280. Nyhan B, Reifler J. When Corrections Fail: The Persistence of Political Misperceptions. Political Behavior. 2010 2010/06/01;32(2):303-30. doi: 10.1007/s11109-010-9112-2.
281. Pool J, Fatehi F, Akhlaghpour S. Infodemic, Misinformation and Disinformation in Pandemics: Scientific Landscape and the Road Ahead for Public Health Informatics Research. Stud Health Technol Inform. 2021 May 27;281:764-8. PMID: 34042681. doi: 10.3233/shti210278.
282. Swetland SB, Rothrock AN, Andris H, Davis B, Nguyen L, Davis P, et al. Accuracy of health-related information regarding COVID-19 on Twitter during a global pandemic. World Medical & Health Policy. 2021 2021/07/29;n/a(n/a). doi: 10.1002/wmh3.468.
283. Sharma M, Yadav K, Yadav N, Ferdinand KC. Zika virus pandemic-analysis of Facebook as a social media health information platform. Am J Infect Control. 2017 Mar 1;45(3):301-2. PMID: 27776823. doi: 10.1016/j.ajic.2016.08.022.
284. Seltzer EK, Horst-Martz E, Lu M, Merchant RM. Public sentiment and discourse about Zika virus on Instagram. Public Health. 2017 2017/09/01/;150:170-5. doi: 10.1016/j.puhe.2017.07.015.
285. Sell TK, Hosangadi D, Trotochaud M. Misinformation and the US Ebola communication crisis: analyzing the veracity and content of social media messages related to a fear-inducing infectious disease outbreak. BMC Public Health. 2020 2020/05/07;20(1):550. doi: 10.1186/s12889-020-08697-3.
286. Albalawi Y, Buckley J, Nikolov NS. Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting Arabic health information on social media. Journal of Big Data. 2021 2021/07/02;8(1):95. doi: 10.1186/s40537-021-00488-w.
287. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: A survey. Science China Technological Sciences. 2020 2020/10/01;63(10):1872-97. doi: 10.1007/s11431-020-1647-3.

288. Azzouza N, Akli-Astouati K, Ibrahim R, editors. TwitterBERT: Framework for Twitter Sentiment Analysis Based on Pre-trained Language Model Representations. Emerging Trends in Intelligent Computing and Informatics; 2020 2020; Cham: Springer International Publishing.
289. Gomez-Perez JM, Denaux R, Garcia-Silva A. Understanding Word Embeddings and Language Models. In: Gomez-Perez JM, Denaux R, Garcia-Silva A, editors. A Practical Guide to Hybrid Natural Language Processing: Combining Neural Models and Knowledge Graphs for NLP. Cham: Springer International Publishing; 2020. p. 17-31.
290. Broniatowski DA, Kerchner D, Farooq F, Huang X, Jamison AM, Dredze M, et al. Twitter and Facebook posts about COVID-19 are less likely to spread misinformation compared to other health topics. PLOS ONE. 2022;17(1):e0261768. doi: 10.1371/journal.pone.0261768.
291. Twitter. 2020 [06 Mar 2021]; Available from: https://developer.twitter.com/en/developer-terms/agreement-and-policy.
292. Xu K, Xie L, Yao K, editors. Investigating LSTM for punctuation prediction. 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP); 2016 17-20 Oct. 2016.
293. Shi S, Tang Z, Chu X, Liu C, Wang W, Li B. A Quantitative Survey of Communication Optimizations in Distributed Deep Learning. IEEE Network. 2021;35(3):230-7. doi: 10.1109/MNET.011.2000530.
294. Safaya A, Abdullatif M, Yuret D, editors. KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media. 2020 dec; Barcelona (online): International Committee for Computational Linguistics.
295. Antoun W, Baly F, Hajj H. AraBERT v1 & v2 : Pre-training BERT for Arabic Language Understanding. 2021 [15 / 06 / 2021]; Available from: https://huggingface.co/aubmindlab/bert-large-arabertv2.
296. Faraj D, Faraj D, Abdullah M, editors. SarcasmDet at Sarcasm Detection Task 2021 in Arabic using AraBERT Pretrained Model. 2021 apr; Kyiv, Ukraine (Virtual): Association for Computational Linguistics.
297. Darwish K, Mubarak H, editors. Farasa: A New Fast and Accurate Arabic Word Segmenter. LREC; 2016.
298. Abdelali A, Hassan S, Mubarak H, Darwish K, Samih Y. Pre-Training BERT on Arabic Tweets: Practical Considerations. arXiv preprint arXiv:210210684. 2021.
299. Abdul-Mageed M, Elmadany A, Nagoudi EMB. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Association for Computational Linguistics; 2021. p. 7088--105.
300. Abu Farha I, Magdy W, editors. Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection. 2021 apr; Kyiv, Ukraine (Virtual): Association for Computational Linguistics.
301. Zhenyan L, Dan M, Weiping W, Chunxia Z, editors. A Supervised Parameter Estimation Method of LDA. Web Technologies and Applications; 2015; Cham: Springer International Publishing.
302. Yamaguchi T, Shimizu J, Oya Y, Horio Y, Hida T. Drug-Induced Liver Injury in a Patient with Nonsmall Cell Lung Cancer after the Self-Administration of Fenbendazole Based on Social Media Information. Case Reports in Oncology. 2021;14(2):886-91. doi: 10.1159/000516276.
303. Powers D. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies [Internet]. 2011; 2(1):[37-63 pp.]. Available from: https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf.
304. Hand D, Christen P. A note on using the F-measure for evaluating record linkage algorithms. Statistics and Computing. 2018 2018/05/01;28(3):539-47. doi: 10.1007/s11222-017-9746-6.

305. Tran T, Valecha R, Rad P, Rao HR. An Investigation of Misinformation Harms Related to Social Media during Two Humanitarian Crises. Information Systems Frontiers. 2021 2021/08/01;23(4):931-9. doi: 10.1007/s10796-020-10088-3.
306. Sunarya POA, Refianti R, Benny A, Octaviani W. Comparison of Accuracy between Convolutional Neural Networks and Naïve Bayes Classifiers in Sentiment Analysis on Twitter. International Journal of Advanced Computer Science and Applications. 2019.
307. Neppalli VK, Caragea C, Caragea D, editors. Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters. ISCRAM; 2018.
308. Tammina S, Annareddy S, editors. Sentiment Analysis on Customer Reviews using Convolutional Neural Network. 2020 International Conference on Computer Communication and Informatics (ICCCI); 2020 22-24 Jan. 2020.
309. Hu Y, Li Y, Yang T, Pan Q, editors. Short Text Classification with A Convolutional Neural Networks Based Method. 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV); 2018 18-21 Nov. 2018.
310. Straton N, Jang H, Ng R, editors. Stigma Annotation Scheme and Stigmatized Language Detection in Health-Care Discussions on Social Media. 2020 may; Marseille, France: European Language Resources Association.
311. Galit S. To Explain or to Predict? Statistical Science. 2010 8/1;25(3):289-310. doi: 10.1214/10-STS330.