applied sciences

MDPI

Article

Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning

Shoayee Alotaibi [1], Rashid Mehmood [2,*], Iyad Katib [3], Omer Rana [4] and Aiiad Albeshri [5]

[1] Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia; SAlotaibi0372@stu.kau.edu.sa
[2] High-Performance Computing Center, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[3] Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia; IAKatib@kau.edu.sa
[4] School of Computer Science, Cardiff University, Cardiff, CF10 3AT, UK; RanaOF@cardiff.ac.uk
[5] Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia; AAAlbeshri@kau.edu.sa
* Correspondence: RMehmood@kau.edu.sa

Received: 21 January 2020; Accepted: 11 February 2020; Published: 19 February 2020

Abstract: Smartness, which underpins smart cities and societies, is defined by our ability to engage with our environments, analyze them, and make decisions, all in a timely manner. Healthcare is the prime candidate needing the transformative capability of this smartness. Social media could enable a ubiquitous and continuous engagement between healthcare stakeholders, leading to better public health. Current works are limited in their scope, functionality, and scalability. This paper proposes Sehaa, a big data analytics tool for healthcare in the Kingdom of Saudi Arabia (KSA) using Twitter data in Arabic. Sehaa uses Naive Bayes, Logistic Regression, and multiple feature extraction methods to detect various diseases in the KSA. Sehaa found that the top five diseases in Saudi Arabia in terms of the actual afflicted cases are dermal diseases, heart diseases, hypertension, cancer, and diabetes. Riyadh and Jeddah need to do more in creating awareness about the top diseases. Taif is the healthiest city in the KSA in terms of the detected diseases and awareness activities. Sehaa is developed over Apache Spark allowing true scalability. The dataset used comprises 18.9 million tweets collected from November 2018 to September 2019. The results are evaluated using well-known numerical criteria (Accuracy and F1-Score) and are validated against externally available statistics.

Keywords: smart cities; healthcare; Apache Spark; disease detection; symptoms detection; Arabic language; Saudi dialect; Twitter; machine learning; big data; high performance computing (HPC)

1. Introduction

Smart cities and societies are driving unparalleled technological growth manifested in our daily lives [1]. We are witnessing a rapid evolution, rather a transformation, of our societies. Novel solutions are being developed and adopted in work and life, benefitting from the growing ability to monitor and analyze our environments in near real-time. A range of devices and technologies are being used for monitoring purposes including the Internet of Things (IoT), GPS, cameras, (radio-frequency identification) RFIDs, smartphones, smartwatches, other smart wearables, and social media. These devices produce diverse data that are analyzed using artificial intelligence (AI) and other computational intelligence methods, and used for decision-making purposes. The key to this

transformative "smartness" is our ability to "engage with the environment", analyze it, and make decisions, all in a timely manner.

The healthcare industry is the prime candidate in need of the transformative capability of this "smartness" [2]. This is for a number of reasons. Healthcare is necessary for our survival. Countries throughout the world are spending a significant portion of their GDPs on healthcare, and this expenditure is rising [3]. For example, the total spending on healthcare in the US equals nearly one-fifth of the US economy [4]. The healthcare services industry is known to be inefficient compared to other sectors, causing waste of resources and higher costs [4]. The cost of high-quality healthcare is increasing in general [2]. The average age of the populations in several countries around the world is also increasing. A surge in the consumption of processed foods and decline in physical activities have given rise to obesity and prevalence of chronic and other diseases including heart diseases, diabetes, cancer, dementia, depression, and hypertension [2]. It is well known that the traditional model of appointments between doctors and patients is not satisfactory, and there is a need to continuously (but non-intrusively) engage with the patients to manage their health [5]. There is an increasing emphasis on personalized disease prevention, leaving disease treatment as a last resort [5]. In short, the key to reversing this trend of the near-ubiquity of chronic, lifelong diseases is the ability of governments, healthcare providers, stakeholders, and the public to "engage" with each other in a dynamic, adaptive, and timely manner. With half of the world population connected to social networks, social media provides a vital solution for a ubiquitous and timely engagement among healthcare stakeholders.

It is reported that roughly 58% of the global "eligible population" (70% of the eligible population in 100 countries around the world) uses social media [6]. Therefore, social media is the ultimate stakeholders' "engagement" platform for healthcare. Social media could provide a two-way communication channel for governments, healthcare providers, stakeholders, and citizens to engage with each other, understand the requirements of all parties, crowdsource and develop innovative solutions, identify specific targets, redefine and improve the public health management experience, improve healthcare services and educational and awareness programs, reduce healthcare costs, provide guidance on alternative personalized routes for treatment of specific diseases, improve transparency and participatory decision-making, and more.

Twitter is one of the most popular social media platforms today [7]. Tweeters, people or organizations, post short messages on Twitter, called tweets, sharing various types of information, including personal and organizational news, status, events, and more [8]. These tweets provide important information on public, technical, or business matters such as on product ratings, new technologies, healthcare, transportation, and politics. People discuss everyday matters related to all aspects of their lives, seek knowledge, and form opinions about various matters. Twitter is being used to interact with the public and with customers, promote services, products, and policies, and gain feedback on service requirements and government matters. Twitter has 330 million monthly active users worldwide, 40% of whom are active on Twitter daily [8]. 500 million tweets are sent every day, equating 5.79 tweets per second [8]. Twitter's influence is well known. For example, according to a 2019 report by the Digital Marketing Institute, 40% of Twitter users purchased something after seeing it on Twitter [8]. As of October 2019, Statista reports that Saudi Arabia has the fifth largest number of Twitter users in the world (10.09 million, roughly one third of its population), only after the US, Japan, Russia, and the UK [7]. Moreover, Saudi Arabia has the highest proportion of Internet users who are active on Twitter. 80% of the Twitter users access it through mobile phones, that allowing connecting to people anywhere anytime, providing access to rich spatio-temporal information [9].

Our research focuses on the use of Twitter media for healthcare in Saudi Arabia, with the aim to develop technologies that provide enhanced healthcare in the country. We have performed a review of the relevant literature (see Section 3), and it reveals that there are two major challenges in this quest. First, the current state of research on the topic is in its infancy, both in terms of the scope of the works [10] as well as the investigation into the methods for healthcare analytics, including that of machine learning methods [11]. Several works exist on the analysis of tweets in the English language; however, much more is needed for developing robust analytics methods, tool functionalities, and

usability. The state of research in languages other than English is even more rudimentary. For the Arabic language, while some works (not in healthcare) are available in Modern Standard Arabic (MSA) [12,13], the works on the Arabic dialects are very limited in numbers and scope [10,14]. Moreover, we have found only three works in Arabic specific to healthcare [15–17], but these are limited in scope, depth, and/or functionalities. Specifically, there is no known work in data analytics of the Saudi Arabic dialect in healthcare.

The second challenge relates to the fact that the scalability and interoperability of the Twitter data analytics tools for healthcare have not been considered. The challenges in this respect include management, integration, and distributed computation of data, including the difficulties related to managing the 4V characteristics of big data, i.e., volume, velocity, variety, and veracity of data. There have been some works on the use of big data platforms in Twitter data analysis in various application domains [18–23]. However, to the best of our knowledge, no work exists that uses big data technologies for data analytics in healthcare using tweets in the Arabic language.

This paper proposes Sehaa (an Arabic word meaning health), a big data analytics system for healthcare. Sehaa is composed of four main modules. The Data Collection module captures and stores public tweet messages from Saudi Arabia using a Twitter streaming API (application programming interfaces) according to a set of predefined parameters. Social media contents are unstructured and contain many errors (e.g., the veracity of big data). The Pre-Processing module cleans and manually labels the data to prepare it for the actual learning and classification stages. The Classification Module comprises six classification models (classifiers) that are used in two classification stages. The classification methods used include Naive Bayes (NB) and Logistic Regression (LR), in combination with four feature extraction methods: BiGram, TriGram, HashingTF, and CountVectorizer. The Validation Module evaluates the classifiers' performance against two widely used numerical evaluation criteria, F1-Score and Accuracy. The results are visualized and validated against external sources such as national statistics, research reports, and news media.

The Sehaa tool is used in this paper to detect symptoms, diseases, and medications in Saudi Arabia. The results are collected and analyzed in terms of the detected diseases and the level of awareness generated in five major cities in the country—Riyadh, Jeddah, Dammam, Makkah, and Taif. Sehaa found that the top five diseases in Saudi Arabia in terms of the actual afflicted cases are dermal diseases, heart diseases, hypertension, cancer, and diabetes. Riyadh, the capital and the biggest city by population, has the highest ratio of awareness to afflicted cases for six of the fourteen diseases that Sehaa has detected. However, the top two diseases, dermal diseases and heart diseases (HRD) are not included in these six diseases, implying that Riyadh should do more in creating awareness of these diseases. Jeddah, the second major city in Saudi Arabia, does not have a good awareness-to-afflicted number of tweets for any of the top five diseases in Saudi Arabia. Jeddah needs to do much more in creating awareness of the top five national diseases. Taif is the fifth major city with a population that is one-eighth of Riyadh's. However, it has a ratio of awareness to afflicted cases for several diseases comparable to Riyadh. We have found that Taif is the healthiest city in Saudi Arabia. It has the lowest number of disease cases in Saudi Arabia (seven out of 14 diseases) while maintaining a high number of awareness activities.

Sehaa is developed over Apache Spark, which is an open-source big data distributed computing platform allowing scalability and feasibility for integration with other datasets and tools. The dataset was formed by collecting tweets from Saudi Arabia during the period between November 2018 and September 2019. A total of 18.9 million tweets were used in the analytics reported in this paper. This study is the first of its kind in Saudi Arabia using Apache Spark and tweets in the Arabic language. Sehaa is an excellent example of integrating artificial intelligence (AI), distributed big data computing, and human cognition, brought together as a convenient tool for the betterment of public health and the economy. The system methodology and design are generic and it can be adopted globally. Our focus in this work is on Saudi Arabia and therefore the tool currently works with tweets only in the Arabic language (it can be used in other Arabic speaking countries, such as UAE, Kuwait, and Egypt). Potential users of this tool are hospitals and other healthcare organizations, ministries of health, pharmaceutical companies, and other healthcare stakeholders.

The rest of this paper consists of six sections. Section 2 presents a brief overview on the background material and Section 3 presents a review of the relevant literature. The methodology, design and architecture of the Sehaa tool are described in Section 4. Section 5 presents a detailed analysis of the results obtained through Sehaa. Section 6 discusses the numerical evaluation and external validation of the Sehaa system. Section 7 concludes and proposes future directions.

2. Background

This section briefly describes important background concepts related to this study, including big data, Apache Spark, the machine learning algorithms, and the feature extraction techniques.

2.1. Big Data

Recently, the term big data has been widely used to describe specific type of datasets. Big data is defined as "the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time" [24]. Big data also refers to the "emerging technologies that are designed to extract value from data having the four Vs characteristics; volume, variety, velocity and veracity" [25]. Hadoop, Apache Spark, and Tableau are examples of technologies that provide solutions for big data.

2.2. Apache Spark

Apache Spark is "a unified analytics engine for large-scale data processing" [26]. Spark has many features, which make it an optimal choice for big data analytics. For example, in parallel processing, applications can be easily constructed using more than 80 high-level operators that are provided by Spark. Spark also offers an interactive shell written in different programming languages such as Python, Scala, or R. Moreover, Spark contains most of the required libraries for big data analytics steps. Figure 1 shows the programming libraries provided by Spark: Spark SQL, Spark Streaming, MLib, and GraphX.

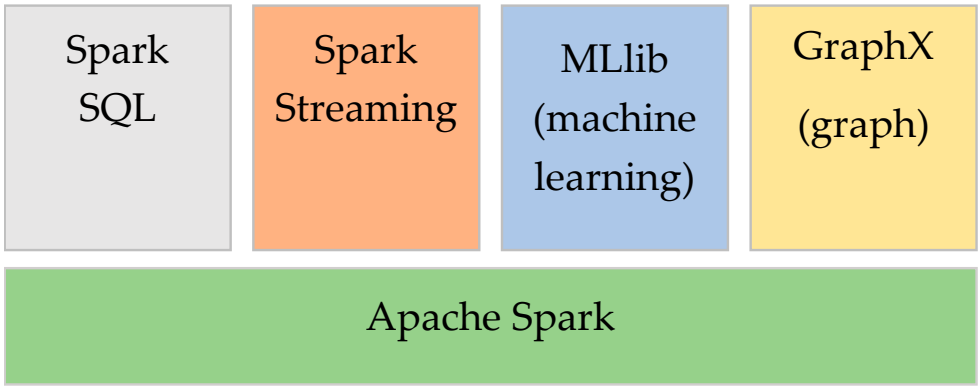

Figure 1. Components of Apache Spark [26].

2.3. Machine Learning

The term machine learning can be defined as the process of learning from input data in order to build adequate experience and then generate the required output [27]. The learning process can be supervised or unsupervised. In supervised learning, the scenario includes predicting missing information (usually a class or a label) for some data. The prediction is made after learning from provided information in the training data. By contrast, in unsupervised learning, the data is not divided into training and testing sets, and the learning process is achieved by grouping data into subsets of similar objects according to various features. The following algorithms are examples of supervised learning algorithms.

2.3.1. Logistic Regression (LR)

In logistic regression, the learner uses the logistic sigmoid function (Equation (1)) to calculate the probability value for the testing data. The appropriate label or class is then assigned according to the resulting probability value [28]. In Equation (1), $s(z)$ is the output between 0 and 1 (probability value), z is the input to the function, and e is the base of the natural logarithm.

$$s\ (z) = 1/(1 + e^{-z}), \tag{1}$$

2.3.2. Naïve Bayes (NB)

For the Naïve Bayes classifier, the learner applies the Bayes theorem assuming independency between the extracted features of the training data. Hence, the NB algorithm is highly scalable according to the number of features included [29]. This algorithm is commonly used in a wide range of big data analytics research.

2.4. Feature Extraction

Feature extraction is a crucial step in constructing a machine learning classifier. It aims to reduce the raw data into a manageable number of variables (set of features) while maintaining the data accuracy. This allows us to select the significant features when building the classifications models. There are several techniques for feature extraction; N-Gram and TF-IDF, which have been used in this study, are defined below.

2.4.1. N-Gram

In machine learning, one of the feature extraction methods is N-Gram. It involves converting the input data into a sequence of separate n tokens, which are usually words. N is an integer number, which is usually one in the one-gram method, two in the bi-gram method, or three in the tri-gram method. In practice, pyspark libraries implement this method using N-Gram class instances [30].

2.4.2. TF-IDF

The term TF-IDF refers to Term Frequency-Inverse Document Frequency. It measures the term importance in a document in a corpus by considering its frequency [30]. If the term appears frequently, that means it does not have special information about a particular document—for example, "and", "a", "the", and "of". For a given corpus D, which contains the number of documents d, the TF-IDF numeric value for a term t is calculated as shown in Equation (2). $TF(t, d)$ is the term t frequency in a document d. $DF(t, D)$ is the number of documents in the corpus D that contain the term t. The term $IDF\ (t, D)$ in Equation (2) is calculated using Equation (3).

$$TFTDF(t, d, D) = TF(t, d) * IDF\ (t, D) \tag{2}$$

$$IDF\ (t, D) = \log \frac{|D| + 1}{DF(t, D) + 1} \tag{3}$$

2.4.3. CountVectorizer

The idea of CountVectorizer is to transform a collection of text documents to vectors of token counts. In cases where an a priori dictionary does not exist, CountVectorizer can be used to extract the vocabulary and construct the required dictionary [30].

2.4.4. HashingTF

For each sentence (bag of words), HashingTF can be used to hash the sentence into a feature vector. We use IDF to rescale the feature vectors; this is done in order to improve performance when using text as features [30].

3. Literature Review

Smart cities provide "state-of-the-art approaches for urbanization … The notion of smart cities can be extended to smart societies … digitally enabled, knowledge-based societies, aware of and working towards social, environmental, and economic sustainability" [1,31]. The foundations of these smart cities and societies are laid out with smart systems and applications. A range of technologies is contributing to the development of these smart systems. These include the Internet of Things (IoT) [32–36], social media [21–23,37,38], big data [39–44], high performance computing (HPC) [45–48], cloud, fog, and edge computing [34,49–52], and machine learning [36,53]. The applications include healthcare [34,39,54–56], transportation [57,58], and others [59,60]. Social media and IoT provide the pulse for sensing and engaging with the environments. Sentiment analysis, or opinion mining, is a vital tool in natural language processing (NLP), defined as "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions toward entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" [61]. Many of the notable works on sentiment analysis rely on machine learning and social media, including Twitter.

In this section, we review the literature related to the use of Twitter sentiment analysis in healthcare. Section 3.1 focuses on healthcare-related analytics in English, while Section 3.2 focuses on the same but for the Arabic language. Section 3.3 describes the research gap.

3.1. Twitter Data Analytics in Healthcare

The data generated by social media, such as Twitter, provides unexpected opportunities to enrich the health care sector (see, e.g., [62]). Fields of applications in health care that have benefitted from social media data include, for example, building surveillance tools to track a certain disease, studying side effects of medications, and exploring healthcare-related habits. The most common methodologies found in the literature are statistical analysis and text mining (including sentiment analysis) of Twitter data using machine learning algorithms. The notable works are reviewed below.

Parker et al. [63] analyzed tweets without aiming to detect a particular illness. The target was to produce "interest curves" that document the generation of hypotheses regarding which health-related conditions/topics have occurred frequently. Unlike other studies, this approach is not dedicated to discovering one certain disease. The main contribution was to convert the stream of tweets to a list of health-related topics. Paul and Dredze developed a model, called the Ailment Topic Aspect Model (ATAM), to link each disease with its possible symptoms, medications, and related words [64]. The ATAM model has been constructed using Twitter data and trained by a support vector machine algorithm (SVM). The experiments demonstrated the efficiency of ATAM in aligning different groups of diseases with their symptoms, medications, and related words. The authors mentioned that ATAM could be used as a surveillance tool for general health topics. The authors extended in [65] the original ATAM to discover mentions of additional illnesses, including allergies, obesity, and insomnia.

The idea of detecting influenza cases from Twitter data was explored by Aramaki et al. [66]. Using a support vector machine (SVM)-based classifier, they detected the influenza patients. The experiments' results demonstrated the practicability of the proposed approach, which showed acceptable correlation comparing with medical reports statistics, especially at the outbreak and early spread (early epidemic) stage. The authors extended their work in [67] and implemented a robust influenza prediction model that enabled the use of direct and indirect information using tweets from urban and rural areas in Japan. This work was further extended in [68]. The authors constructed a more generalized diseases surveillance tool that performs multi-label and cross-language tasks. The multi-label approach was used to classify tweets into eight different diseases symptoms and label them appropriately with patients' symptoms. The tool is cross-language and works on three different languages: English, Japanese, and Chinese.

Lamb et al. [69] found that deeper content analysis of tweets leads to a remarkable improvement in influenza's surveillance performance. More precisely, flu-related tweets do not reflect infections alone; rather the tweets might be related to suspicions related to flu infections, worries about the

infection, or discussions about the disease itself. Hence, they relied on trained classifiers that can distinguish between the awareness and infection tweets using a pre-annotated set of trained data. Their results demonstrated that by distinguishing between types of flu tweets to identify reports of infection, reasonable surveillance could be recovered. The obtained results brought to the attention of the NLP community that deeper content analysis of tweets is worth investigation. Smith et al. [70] developed a real-time surveillance tool for disease awareness rather than monitoring infection cases. This system offered an opportunity not only for public health officials to identify awareness trends, for which they often have no other data sources, but also to study what drives awareness of influenza in a population.

The studies are not limited to tracking the spread of a specific epidemic; they also examined following the side effects of certain medications. For example, Bian et al. [71] reported that sentiment analysis of Twitter data revealed some of the unreported side effects for five drugs related to cancer medications. The researchers firstly specified the drug users' Twitter accounts using a Support Vector machine (SVM) classifier. Then, they evolved an analytic framework that integrates natural language processing and machine learning methods to capture drug-related adverse events from the Twitter messages. Their findings showed the possibility of supporting pharmacovigilance by extracting knowledge from tweets.

Mayslin and Zhu explored in [72] the role of Twitter data analysis in tobacco consumption surveillance. The extracted sentiment was linked in complex ways with social image, personal experience, and recently popular products such as the hookah and electronic cigarettes. It also showed the need for public awareness of their health effects. Taken together, these findings suggest a role for machine classification of tobacco-related posts over strictly keyword-based approaches in enhancing tobacco surveillance applications. Jashinsky et al. [73] investigated whether Twitter data can be a promising source for researchers to identify suicide-related risk factors. They indicated in their study that Twitter could be a useful tool for the early detection of individuals who are at risk of suicide. Their findings demonstrated a strong correlation between Twitter data and actual suicide data for a certain state in the United States.

Achrekar et al. [74] implemented a novel approach in estimating influenza statistics using Twitter data and real data provided by Centers for Disease Control and Prevention (CDC). Their model showed that to some extent, Twitter data could be conferred as a reliable source for the real-time assessment of epidemic conditions and could offset the lack of real statistics. They showed that text mining significantly improves the correlation between Twitter and the Influenza like Illness (ILI) rates provided by the Centers for Disease Control and Prevention (CDC). They built a model using a support vector machine algorithm and simple bag-of-words text processing to detect flu cases. Due to the high correlation of their model results with real statistics, they then built an estimation tool using the same techniques that could compensate for the absence of real statistical cases.

Furthermore, for a certain disease such as influenza, surveillance approaches have proved their efficiency when tested on larger scale regions [75]. However, not only influenza was a public health concern; asthma is another concern that implies readiness from emergency rooms in hospitals. In [76], a robust model was developed using different sources of real data to predict the number of asthma-related emergency department (ED) visits in a specific area. Unlike the traditional surveillance tools, which rely on EMRs (electronic medical records), the study relied on employing machine learning in social media and environmental sensors data. In practice, for a specific geographic area within a time period, they collected Twitter data and examined the relationship between Twitter data, internet users' search interests from Google, ED asthma-related visits data, and pollution sensor data. The association between these different data was examined to build a prediction tool for asthma-related visits. After that, a prediction tool was built to estimate the number of ED asthma related visits. The tool successfully estimated the rate of asthma ED visits using a combination of independent variables from the above-mentioned data sources.

Culotta in [77] developed a flu tracking system using a supervised learning approach for the analysis of flu-related tweets. The results of the tracking system showed a high correlation with the real statistics generated by the national health authorities. That provided further confidence in the

usefulness of Twitter as a data resource for health-related research and for the robustness of natural language processing NLP algorithms.

3.2. Twitter Data Analytics in Healthcare (Arabic)

We have found only three works on Twitter data analytics for healthcare in the Arabic language. Alayba et al. [15] introduce a new Arabic annotated dataset about health services which they state is a necessary component in sentiment analysis studies. The dataset consists of about 2026 tweets in the Arabic language. The tweets were collected over a six-month period using the four most popular hashtags about health services. Pre-processing including the normalizing steps was applied. The clean tweets were annotated by three annotators to be either positive or negative. The classification of tweets was limited to two classes only, due to the difficulty of rating the opinion in the Arabic language compared to English. Their experiments were performed by various machine learning algorithms and deep neural networks using different settings. They reported to have obtained a best classifier results using SVM with Linear Support Vector Classification and Stochastic Gradient Descent. Alkouz and Aghbari [16] detect influenza in the UAE from Arabic tweets. They classified tweets and used them to predict the number of future hospital visits using a linear regression model. Their work focused on analyzing tweets in Arabic MSA and in the UAE dialect. The authors reported correlations between their reported results and those obtained from the UAE Ministry of Health. This work did not use any big data technologies. Ilyas and Alowibdi [17] used tweets in Arabic to track diseases in the Gulf Cooperation Council (GCC) countries. This work used a small number of tweets and did not use any AI and big data technologies.

3.3. Research Gap

The discussions on the related works provided above clearly establish the immense potential of Twitter data analytics in healthcare. Two major challenges are evident. First, the current state of research on the topic is limited, both in terms of the scope of the works [10] as well as the investigation into and comparison of the methods for healthcare analytics, including machine learning methods [11]. A number of works exist in the analysis of tweets in the English language; however much more is needed to develop robust analytics methods, tool functionalities, and usability. Further works are needed in languages other than English, particularly cross-language works. There are several challenges that hinder the development of tools for Twitter data analytics in the Arabic language, the greatest being the complexity of the language itself. Research on Twitter data analytics in Arabic has begun to appear in recent years in various application domains (detecting authors' genders [12], detecting traffic related events [18,20,38], finding restaurants' reputations [13]) but the progress has been slow. Moreover, some works are available in Modern Standard Arabic (MSA), but in general (not specific to healthcare), the works on Arabic dialects are very limited in number and scope [10,14]. The three works in Arabic specific to healthcare [15–17] that we have discussed in the previous section are limited in scope, depth, and/or functionalities. There is no known work in data analytics specifically on Saudi Arabic dialect in healthcare.

The second challenge related to Twitter data analytics in healthcare concerns the scalability and interoperability of the Twitter data analytics systems. The challenges in this respect include management, integration, and distributed computation of data including the difficulties related to managing the 4V characteristics of big data, i.e., volume, velocity, variety, and veracity. There have been some works on the use of big data platforms in Twitter data analytics in various languages but in different application domains [18–23,37,78,79]. To the best of our knowledge, no work has been reported that uses big data technologies for data analytics in healthcare using tweets in the Arabic language.

Healthcare is among the most data-intensive generating and consuming sectors. The use of big data distributed computing technologies is important for scalability and integration with other sources of healthcare information. Motivated by these research gaps, we have attempted the design of a tool that provides healthcare analytics capabilities from tweets in the Arabic language using big data technologies.

4. Sehaa Tool: Methodology and Design

In this section, we describe the methodology and design of our proposed Sehaa system. We have built a software tool based on the design of the proposed Sehaa system and we will refer to it as a tool or system interchangeably. The tool comprises four modules and these are described in separate subsections (Section 4.2 and 4.5) subsequent to the first Section (4.1), which provides an overview. The dataset is described in Section 4.2.

4.1. Sehaa: An Overview

We built the Sehaa system in order to detect the most frequent health symptoms and diseases. The system methodology and design are generic and can be adopted globally. However, our focus in this work is on Saudi Arabia and therefore the tool currently works with the Arabic language. The architecture of the Sehaa tool is illustrated in Figure 2**Error! Reference source not found.**. Sehaa is composed of four main modules.

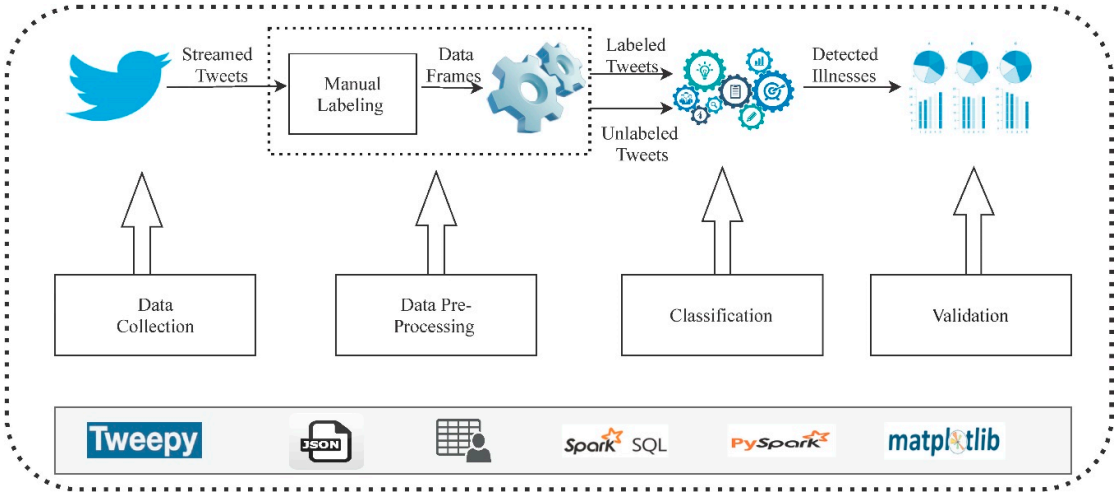

Figure 2. Proposed architecture for the Sehaa system.

The overall process of the Sehaa system can be summarized as follows. First, we capture and download public tweet messages from Saudi Arabia using a Twitter streaming API according to a set of predefined parameters. These parameters are the (Arabic) language, the location, and a set of search keywords which represent health symptoms and diseases (see Section 4.2, [80], and Table 1).

It is well known that social media contents are unstructured and contain many errors (i.e., veracity of big data). Therefore, in the Pre-Processing module, the acquired data are cleaned and pre-processed to ensure its readiness for the actual learning, classification, and prediction stages. This forms our dataset. We then divide the dataset into training and testing datasets. 60% of the tweets are included in the training set and the rest in the testing dataset. There is no lexicon or libraries available in the Arabic language for machine learning, particularly for healthcare, and, therefore, we have manually labeled the tweets; see details of the Pre-Processing Module in Section 4.3. Subsequently, we build six classification models (classifiers) to be used in two classification stages, first to classify related and unrelated tweets, and then to detect symptoms and diseases; see Section 4.4. Finally, the results are validated using multiple numerical criteria and external sources (news media) and visualized; see Section 4.5.

Table 1. Sehaa: list of symptoms, medications, and diseases.

Keywords (Symptoms, Diseases, and Medications)	Corresponding Disease	Abbreviation	Symptoms (Symptoms, Diseases, and Medications)	Corresponding Disease	Abbreviation

السكر سكري السكري انسلوين أنسولين الانسولين	Diabetes السكري	DBT	زكام زكمه أنفلونزا	Influenza الانفلونزا	INF
سرطان الخبيث	Cancer السرطان	CNR	كحة كحه ساسية ساسيه	Cough الكحة	CGH
شريان جلطة جلطه سكته سكتة السكتة السكته قلب	Heart Disease أمراض القلب	HRD	قولون قالون قيلون	Colon القولون	CLN
الضغط	Hypertension غط الدم	HYT	سل درن	Tuberculosis السل	TBC
غدة غده الغده درقيه الدرقية الدرقيه	Thyroid Diseases الغدة الدرقية	THD	ارتجاع رقان رقان معدة	Stomach Disorders ألآم المعدة	STD
أرق	Insomnia الارق	INS	قر ة معوي نزلة معوية نزله معويه	Intestinal Disorders ألآم الأمعاء	IND
كولسترو كلسترو	Cholesterol الكوليسترو	CHT	اكزيما أكزيما برص بهاق جر الجر	Dermal Diseases أمراض الجلدية	DRD
ربو كتمه	Asthma الربو	AST			

4.2. Data Collection Module

The main purpose of this module is to collect relevant tweets. Our aim is to capture tweets in the Arabic language that are related to healthcare in Saudi Arabia. We have used a stream listener for the purpose.

4.2.1. Keywords and Geolocation

Firstly, we collected the tweets using a filter based on a list of keywords. These keywords were partly taken from the latest statistics published by the Saudi Ministry of Health (MOH) on their official website [80]. Some of the other keywords were acquired from the domain experts through personal discussions, and from common knowledge and vocabularies in the Saudi Dialect. Table 1**Error! Reference source not found.** lists common vocabularies for most common symptoms and diseases in Saudi Arabia. Column 1 lists the set of keywords that could be used by Tweeters to mention symptoms, diseases, or medications and hence were included in the search keywords. These keywords are grouped based on particular representative diseases in English and Arabic, and are listed in Column 2. The abbreviations for the diseases are provided in Column 3. The next three columns provide similar details for other diseases.

Secondly, we filtered the tweets based on the geographical location of the tweets to ensure that the tweets are generated from Saudi Arabia. This was achieved by specifying the top-right and bottom-left geographical coordinates and defining a bounding square box around the country. The tweets that originated from within the defined bounding box were extracted by the stream listener. This is different from the paid Twitter Search API, where the user's location field is used to filter

tweets. A Python function using the Tweepy library was written that incorporated the two filters described above into the stream listener.

We understand that some tweets originating outside Saudi Arabia could also be relevant to this work and should have been collected. However, we used a free streaming Twitter API and were limited in resources. Moreover, the tweets originating from outside Saudi Arabia would form a small proportion of the overall tweets and we plan to consider this in future work.

4.2.2. The Data Set

The data were collected between 20 November 2018 and 9 September 2019. However, we were unable to run the collection script several times during the period due to technical difficulties and personal circumstances. The periods when the tweets were collected are: 20 November 2018 to 8 January 2019, 13 February to 29 May 2019, and 31 July to 9 September 2019. This makes a total of 195 days of data. A total of 18.9 million tweets were collected.

4.2.3. The JSON Parser

The tweets from the streaming API as described above are tweet objects in the JSON (JavaScript Object Notation) format. The structured JSON format is the default response of the Twitter API. A part of a tweet object in the JSON format is illustrated in Figure 3. The JOSN format consists of pairs of attributes and values for different objects. Tweets and Users are the two main objects, and each object has its own attributes. The values of the attributes can be accessed using appropriate indexing.


```
{
"created_at": "Tues Dec 20 20:19:24 +0000 2018",
"id": 1050118,
"id_str": "1050118621198921728",
"text": "يتم تشخيص سكري الحمل بعمل اختبار الدم قبل المحلول وبعده وعلي حسب القراءات يبلغك الطبيب",
"user": {"location": Riyadh
},
"entities": {}
}
```


Figure 3. Example of the JSON objects of a Tweet.

There are a number of difficulties related to handling tweets in the JSON format, including programming and computational complexities. Moreover, the existing parsers are dedicated to the English language and many encoding issues are encountered when they are used for Arabic. Therefore, we built a parser to extract the required attributes of the tweets in the JSON format and to store them into the CSV format. Each tweet in the JSON format is stored as a separate file. The parser takes these files containing JSON tweets, extracts all the required attributes of each tweet, and stores them in CSV files. This time, however, each tweet is not stored as a separate file; rather all the tweets related to a specific keyword (see Table 1) are stored in a single file. A new file is used when the maximum file size limit is reached.

The parser algorithm is given as Algorithm 1. It starts by iterating over all the tweet objects in the JSON format. The necessary attributes, such as tweet ids, text, time, and location are extracted and stored in RAM in multiple lists. Finally, these lists are exported to CSV files in the secondary storage (see Figure 4). The generated CSV files are relatively easy to modify for annotation and labeling purposes.

Algorithm 1: Sehaa JSON Parser

Input: tweets_json [ ]
// A list of files each containing tweet objects in json format
Output: tweets_csv [ ]
// A list of files each containing tweet objects in csv format

```
1 tweet_csv_temp=[ ]
  label1=null
  label2= null
  for tweet in tweets_json do
2   tweet_json_temp=[ ]
    id=tweet["id"]
    text=tweet["text"]
    time=tweet["created_at"]
    location = tweet["User"]["location"]
    tweet_json_temp.append(id, time, text, label1, label2)
    tweet_csv_temp.append(tweet_json_temp)
3 header=[id, text, time, location, label1, label2]
  csv_writer.writerow(tweets_csv, header)
  csv_writer.writerows(tweets_csv, tweet_csv_temp)
```

ID | Text | Time | Location | Label1 | Label2
1050118 | يتم تشخيص سكري الحمل بحمل اختبار الدم قبل المحلول وبعده وعلي حسب القراءات يبلغك الطبيب | Tues Dec 20 20:19:24 +0000 2018 | Riyadh | NULL | NULL
1050212 | سكري له فترة معلق على ٤٠٠ كثير يعني من يوم الجمعة الى اليوم تقريبًا ٥ مرات اروح للمستشفى؟ | Tues Dec 20 22:00:00 +0000 2018 | Jeddah | NULL | NULL
1062222 | ي اللهم اشفي مرضى السرطان | Wed Dec 21 07:13:24 +0000 2018 | Jeddah | NULL | NULL
1078882 | مارق الرياض تالي الليل | Fri Dec 23 01:15:55 +0000 2018 | Riyadh | NULL | NULL

Figure 4. Sehaa: the output of the JSON parser.

4.3. Data Pre-Processing Module

Data pre-processing or preparation is a crucial step within a data analytics pipeline, due to the unstructured and informal nature of the data generated by social media. It involves applying several techniques to the acquired data set in order to clean the data. Data pre-processing should be performed to ensure the data readiness for the subsequent steps wherein the actual analytics will be performed. It also enhances the quality and accuracy of data analytics. Some libraries are available for the pre-processing of text in various languages. The NLTK (natural language toolkit) library is an example. It is used to pre-process text in a wide range of languages including English. However, it does not provide satisfactory support for the Arabic language.

Based on our design preferences and the nature of our data set, we have created a specific pre-processing algorithm appropriate for the Arabic language (see Algorithm 2). It starts by removing all advertisement tweets; then incomplete and non-Arabic tweets are removed. Duplicated and unwanted characters such as emoticons characters are also removed.

4.3.1. Labeling the Tweets

Subsequent to the data cleaning phase, the dataset needs to be labeled for training and testing purposes. We randomly selected a part of the data set and divided it into training (60%) and testing (40%) datasets. We manually labeled all the selected tweets using two levels of labeling. At the first

level, we label tweets to distinguish between "related" and "unrelated" tweets, using the labels "R" and "U", respectively. Tweets that express sickness cases and include information about a specific disease or news about awareness events regarding certain health phenomena are considered as related. However, if the text contains one of the search keywords but the context does not reflect a health concern, such as supplications, jokes, and poems, it is labeled as unrelated. A few examples of tweets and their labels are shown in Table 2.

The second level of labeling is performed on the related tweets to distinguish between the tweets that communicate to create "awareness" about diseases and those which are reporting actual cases of being "inflicted by" a disease. The former category of the related tweets is labeled as "A" and the latter category of tweets as "I".

Algorithm 2: Sehaa Pre-Processing Algorithm

```
Input: tweets_csv [ ]; stop_words_list
Output: tokens [ ]
tokens =[ ]
 tweets_df = load(tweets_csv)   // load csv files to spark dataframe (df)
for tweet in tweets_df do
    temp=[ ]                    // a temporary variable carries a tweet tokens
    /* Check if the Tweets is advertisement                              */
    if tweet[text].contains(phone_number) then
        tweets_df.remove(tweet)
        continue
    else
        tweet.replace(URL, Emoticons, NonArabicChar, RT,
         Hashtags)with(" ")                  // Using Regular Expressions
        temp = pyarabic.arabic.tokenize(tweet)        // Pyarabic library
        for t in temp not in stop_words_list do
            if len(tweet) < 3 then
                tweets_df.remove(tweet)
                continue                              // delete short tweets
        tokens_list.append(temp)
tweets_df["filteredText"]= tokens
```

Table 2. The manual labeling of the tweets in Sehaa: a few examples.

No.	Tweet's Text (Arabic and Its Literal Translation)	Context Type	1st Level Label	2nd Level Label
1	يتم تشخيص سكري الحمل بعمل اختبار الدم قبل المحلول وبعده وعلي حسب القراءات يبلغك الطبيب Pregnancy diabetes is diagnosed by blood test before and after the syrup and based on the results, the doctor will tell you.	Medical information (awareness tweet)	R	A
2	سكري له فترة معلق على ٤٠٠ كثير يعني من يوم الجمعة الى اليوم تقريبًا ٥ مرات اروح للمستشفى؟ I am diabetic and my sugar level is more than 400 since Friday, five times a day, do I need to go to the hospital?	Complaint (afflicted by disease)	R	I
3	أرق على أرق ومن مثلي يأ رق Sleepless upon sleepless and who is like me?	Complaint	R	I
4	مارق الرياض تالي الليل Riyadh is being delicate at late night.	Poem	U	U

5	اللهم اشفي مرضى السرطان May Allah cure cancer's patients	Supplication	U	U
6	ليس مرض بل شخص كـ قطعة سكر It is not a disease, just a person such as sugar.	Joke	U	U
7	لايفوتكم تمر سكري مكنوز مجروش فاخر للتواصل وتس أب كيلو ب 60 ريال والتوصيل لجميع مناطق المملكة Do not miss the chance, luxury mjarwsh sukari dates, for delivery to all Saudi cities with 60 riyals.	Advertisement	U	U

4.4. Classification Module

In machine learning, classification aims to predict a category, a class, or a label of a given input data based on the rules generated during the learning or training phase. The label values are already known, and the class boundaries are well defined in the training data. In the learning phase, the classifier uses the labeled data (training data) to generate the rules of classification in order to learn how to predict the labels for the data provided in the future.

Several classification techniques have been suggested and implemented in various programming language libraries, such as Python and Pyspark. Support vector machine, decision tree and deep learning are examples of machine learning techniques [81]. In our study, we have used the Naïve Bayes (NB) and Logistic Regression (LR) algorithms in different combinations with four feature extraction methods: BiGram, TriGram, HashingTF, and CountVectorizer.

The Classification Module receives data from the Pre-Processing module containing labeled and unlabeled tweets and classifies the tweets using the NB and LR algorithms in different combinations with the four above-mentioned extraction techniques. At the first level, the tweets are classified into related and unrelated tweets. Subsequently, the related tweets undergo a second level of classification where we separate the tweets into the "awareness" and the "afflicted by" tweets (see Section 4.3 and Table 2).

4.5. Validation Module

The validation of the classification results is the most important task in data analytics. The task involves evaluating the classification models using a set of criteria. We have used two widely used numerical evaluation criteria, *Accuracy* and *F-1 Score*. These are calculated as follows:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (4)$$

$$F\text{-}1\ Score = 2/(1/Precision + 1/Recall), \quad (5)$$

$$Precision = TP/(TP + FP), \quad (6)$$

$$Recall = TP/(TP + FN), \quad (7)$$

where True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), False Negative (*FN*), Precision, and Recall are well known metrics. We use the *Accuracy* and *F-1 Score* criteria to select the best algorithm for the classification of the tweets.

4.5.1. Visualization

An important function in the Sehaa system is to prepare various statistics such as the numerical evaluation metrics mentioned above and to visualize them.

4.5.2. External Validation

We also validated the results obtained through the Sehaa system against various external sources. These sources include the Institute for Health Metrics and Evaluation (IHME), which is an independent global health research center at the University of Washington [82], the official World Health Organization (WHO) website [83], the Centers for Diseases Control and Prevention [84], and

the Saudi Ministry of Health (MOH). We looked at these sources and compared the relative occurrence of various diseases in Saudi Arabia. The results will be discussed in the next section.

5. Sehaa: Results and Discussions

This section summarizes the main findings of this research. In Section 5.1**Error! Reference source not found.** we analyze the pre-classification results. Section 5.2**Error! Reference source not found.** presents the post-classification results using the first level classification. The results differentiating the awareness and afflicted cases from the second level classifier are discussed in Section 5.3. See Section 4.4, for the classification methodology.

An important point in reporting the results in this section should be noted here. We understand that a tweet could mention multiple diseases (though our experiences suggest that tweets mentioning multiple diseases are rare). However, in this work, we made the decision to associate each tweet with only one disease, the one which is mentioned first in a tweet. Future work will explore this further and improve the quality of data analytics in Sehaa.

5.1. Pre-Classification Results

Figure 5 shows the frequency of symptoms, medications, and diseases in Saudi Arabia detected by the Sehaa system; see Table 1 for the list of keywords and their representative diseases. The x-axis represents symptoms, medications, and diseases in the Saudi dialect. The y-axis shows the number of tweets that mention each keyword. We have used the log-10 scale for the y-axis in order to be able to show both the low-frequency and high-frequency keywords. These are pre-classification results, and they also include irrelevant tweets such as advertisements. The first level classification has not been performed yet, and hence the tweets may include the use of various terms that are related to healthcare but are being used in the contexts other than healthcare. See Table 2, example 4, where the Arabic word (transliterated into English as) "araqq" could mean insomnia but is being used to mean delicate. Other examples include supplications listed as example 5 in Table 2.

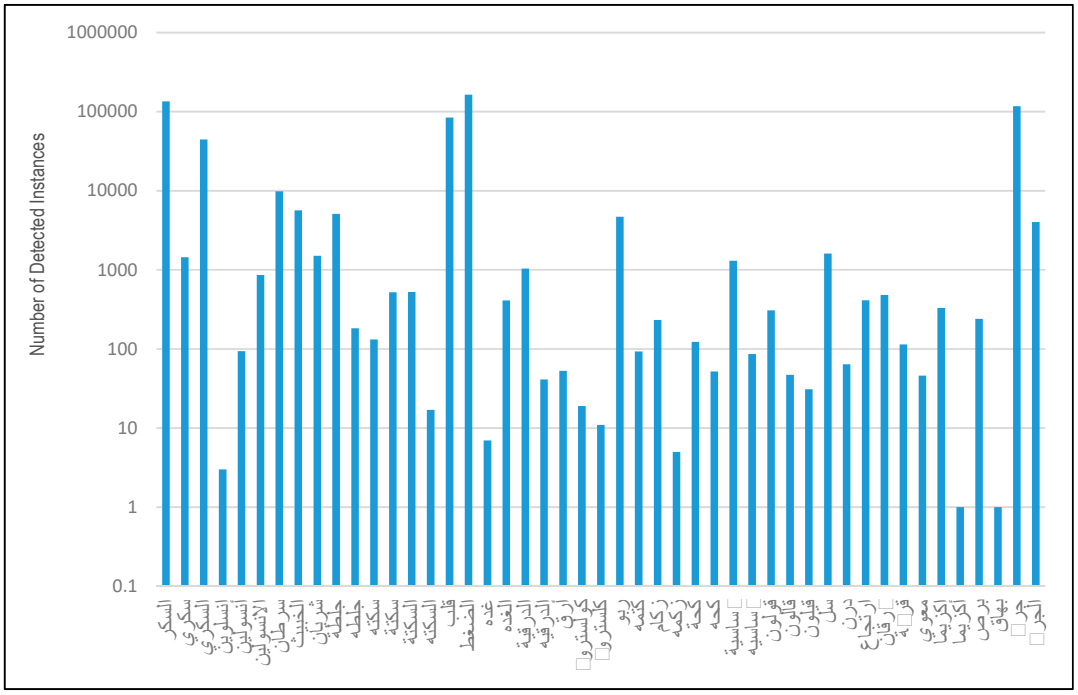

Figure 5. Sehaa: symptoms, medications, and diseases (Saudi Arabia) (pre-classification).

Note in the figure that the distribution of keywords has a high variance. The highly frequent keywords include (الضغط, قلب, السكري, السكر, جر□), (blood pressure, heart disease, diabetes, and scabies), while (بهاق, أكزيما), (eczema, and vitiligo), are the least frequent keywords.

Looking at the uses of various keywords for a specific disease, most of the tweets that mentioned heart disease used the keywords (قلب, جلطة). Similarly, for dermal diseases, the most frequently used keywords were (جر□) and (اكزيما). For cancer, the most frequently used keywords were (الخبيث, سرطان). For hypertension diseases there is the only word that is used in Arabic, (الضغط), and the figure shows that it is one of the most of common diseases in Saudi Arabia. The diabetes disease is represented by multiple keywords (see Table 1). Figure 5 shows that a number of these keywords have been used to talk about diabetes and with fairly high frequencies, implying that it is one of the most common diseases in Saudi Arabia. Finally, we note that scabies is one of the most common diseases in Saudi Arabia according to the figure. However, it is in fact not that common in Saudi Arabia based on our external validation (see Section 6.2). The reason for the high occurrence of the Scabies disease in Figure 5 is due to a number of epidemic cases that were reported in Saudi Arabia in late 2018 (the period when this data was collected).

Figure 6 plots the data as in Figure 5 but as a pie chart, and depicts the numerical proportions of various diseases. The symptoms, diseases, and medications are aggregated using the terminology given in Table 1. Note that the highest four common diseases in the figure are DBT (diabetes) with 31% of total mentions, HYT (hypertension) with 28%, DRD (dermal) with 21%, and HRD (heart disease) with 16% of total mentions. Diabetes is the most common disease in Saudi Arabia based on the unfiltered results.

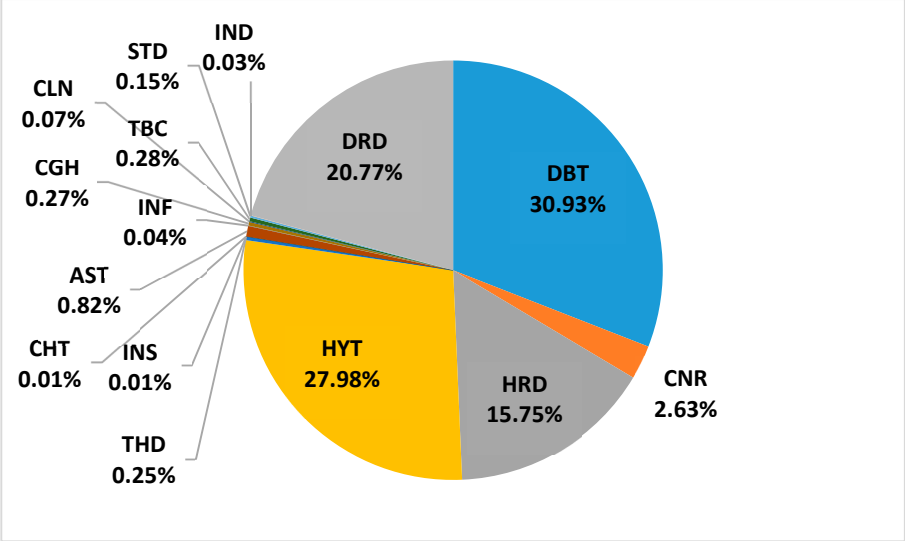

Figure 6. Sehaa: symptoms, medications, and diseases (Saudi Arabia) (pre-classification) (pie chart).

5.2. Post-Classification Results (First-Level)

This section presents and analyzes the results obtained from the first-level classifier, i.e., the classifier that removed unrelated (or irrelevant) tweets (for classifier details, see Section 4.4). The difference between these results compared to the ones presented in the previous section is as follows. Firstly, these results do not include irrelevant tweets as explained in the previous section. Secondly, the results are aggregated based on the common terminologies provided in Table 1. Thirdly, we provide results for major cities in addition to the whole of Saudi Arabia and also provide statistics based on the data, which were normalized with the population sizes of the examined cities.

Figure 7 illustrates the post-classification (first-level) distribution of the diseases detected by the Sehaa system. Each set of bars in the figure represents a unique disease and the detected numbers of occurrences are plotted using the log-10 scale on the y-axis. For detailed analysis, we have extracted the tweets for the five major cities in Saudi Arabia and plotted them for each unique disease. There

are a total of 14 diseases. The cities are Riyadh, Jeddah, Dammam, Makkah, and Taif. After the classification stages, we filtered the tweets according to their locations by retrieving the locations of the users. The location of a tweet is available in the location attribute of the user object, which is part of the tweet objects. Note that we were only able to find locations of tweets where these were enabled by the user. The sixth bar in the figure for each disease is the total number of occurrences in the whole of Saudi Arabia. The absence of bars implies no cases detected for a disease in a city. The three-lettered abbreviations of the diseases are provided in Table 1.

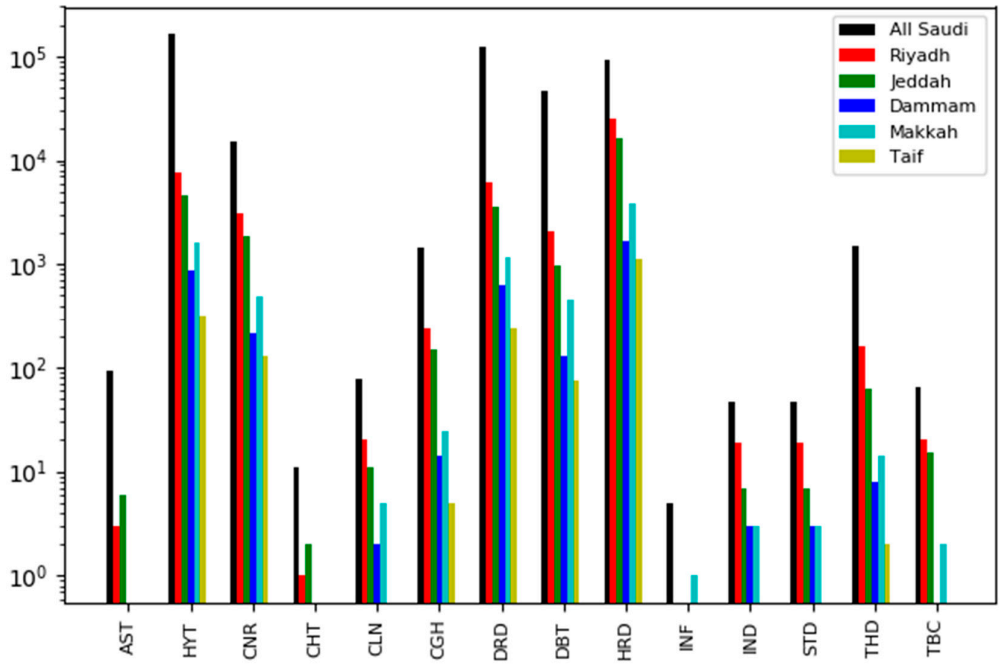

Figure 7. Sehaa: distribution of Diseases across major cities (post-classification) (first-level).

For further exploration, and to take into consideration the fact that bigger cities, such as Riyadh, have a higher population and hence potentially a higher number of tweeters, the frequency of each disease was normalized against the relevant population. That is, the frequency of a disease for a city was divided by the population of the city, and the frequency of a disease for the whole of Saudi Arabia was divided by the population of the country. These normalized values for the 14 diseases, five cities, and the whole of Saudi Arabia are plotted in Figure 8.

Figures 7 and 8 show that the top five prevalent diseases in Saudi Arabia are hypertension (HYT), dermal diseases (DRD), heart diseases (HRD), diabetes (DBT), and cancer (CNR). This pattern is somehow different from the pattern of all five major cities (Riyadh, Jeddah, Dammam, Makkah, and Taif). The top five prevalent diseases for the five cities are heart diseases, hypertension, dermal diseases, cancer, and diabetes. However, there are clear differences in some of these cases in terms of the precise frequencies of the top five diseases across the five cities.

Note in the normalized values (Figure 8) that Riyadh and Jeddah have the highest frequency among the five cities for most of the diseases. One exception is Makkah, which is the only city where influenza has been detected, and this could be due to a large number of people visiting from all over the world coming in close proximity, which could spread influenza in Makkah more than in other cities. This is generally known in Saudi Arabia and therefore the Sehaa findings have verified the common knowledge in Saudi Arabia. We would like to note here that we are not suggesting that influenza does not exist in other cities, but that perhaps people do not talk about it as much as they do in Makkah.

Note also in Figure 8 that the overall national normalized number of the occurrence of most diseases is lower than the frequencies for one or more of the five cities except asthma (AST),

hypertension (HYT), dermal diseases (DRD), diabetes (DBT), and thyroid diseases (THD). Riyadh, followed by Jeddah, have the highest normalized value among the five cities and the national value. Asthma (AST) and cholesterol (CHT) were only detected in Riyadh and Jeddah and not in the other three cities. For Makkah, there was an almost equal number of detected cases for diabetes and cancer. The most intriguing finding is that Taif city appears to be the healthiest city. It had the lowest number of detected diseases; only seven diseases were detected out of a total of fourteen diseases. It had the lowest normalized value for all the detected diseases.

Finally, note that the top five diseases detected by the pre-classification (Figure 6) and post-classification (Figures 7 and 8) results are different. A careful look at the numbers reveals that the pattern is more or less the same for all diseases except for diabetes. The reason for these differences is firstly the removal of irrelevant tweets, which makes the post-classification results a better indication of the reality. Secondly, in Saudi Arabia, the most common word used by the public for diabetes is "سكر" that literally means "sugar". People also use the word "سكر" to show love for each other, to (positively) make jokes to each other, and greet each other to have a sweet morning, evening, etc. (see Table 2**Error! Reference source not found.**, Row 6). There were a large number of such tweets that were filtered out by the classifier and therefore diabetes was not detected as the top disease. Moreover, the words "سكري" and "السكري" also refer to a type of date fruit (see Table 2, Row 7). These words in fruit connotation are also found very frequently in tweets and these were filtered out as advertisements and other types of unrelated tweets. Nevertheless, a deeper look into these results and the classifiers is needed and forms our future work.

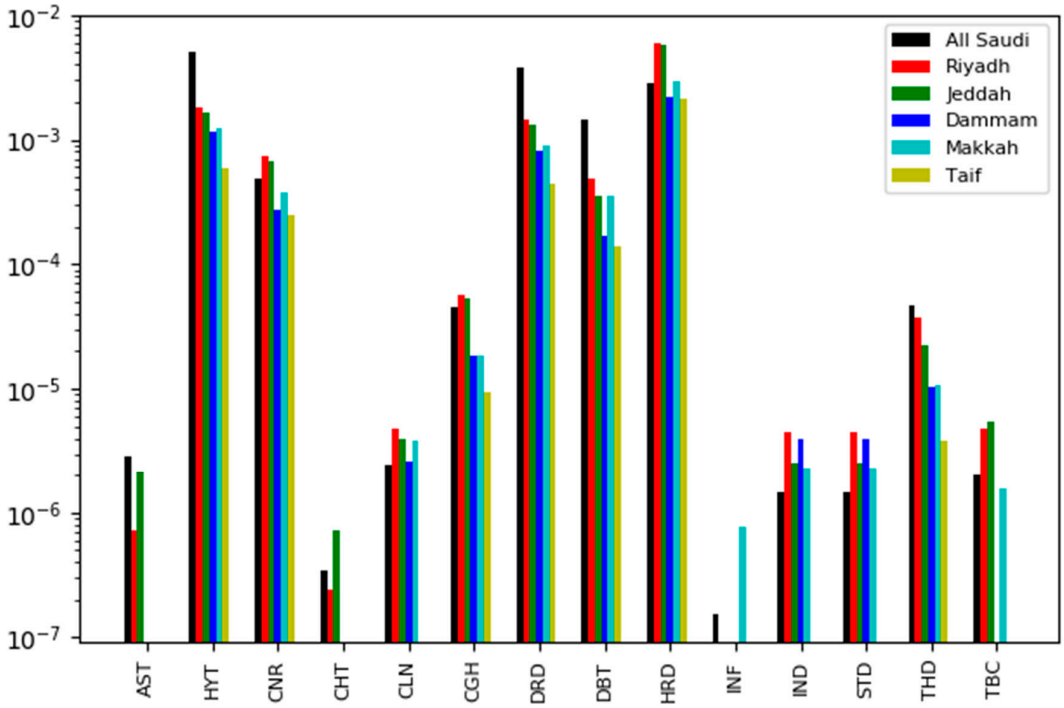

Figure 8. Sehaa: distribution of diseases across major cities (*Normalized*) (post-classification) (first-level).

5.3. Awareness vs. Afflicted Tweets

The tweets detected by the first-level classifiers comprise both the "afflicted by" and "awareness" tweets (see Section 4.3). The "afflicted by" tweets include cases of sickness, suffering, and medication. The awareness tweets communicate to create awareness about diseases. Usually, the awareness tweets are posted by medical practitioners who are tweeting for raising public awareness purposes. The purpose of the second-level classifier is to separate the actual cases from the awareness

tweets in order to compute the actual number of diseases cases. Moreover, this classification also helps with finding the level of awareness-related activities for particular diseases across various cities in the Kingdom of Saudi Arabia (KSA).

Figure 9 depicts the total number of awareness and afflicted cases for the set of fourteen diseases in the KSA. The top five diseases in terms of awareness tweets were hypertension (HYT), dermal diseases (DRD), heart diseases (HRD), diabetes (DBT), and cancer (CNR). The top five diseases in terms of the actual "afflicted by" cases were dermal diseases (DRD), heart diseases (HRD), hypertension (HYT), cancer (CNR), and diabetes (DBT). The rankings of the diseases using the awareness tweets are the same as for the first-level classification. However, the ranking of the occurrence of the diseases using the "afflicted by" cases is different from the first-level classification (see Figure 9). Although this appears to be a surprise, it is expected because the number of the awareness tweets is much higher than the number of the tweets reporting the actual cases, and these bigger numbers dominated the first-level classification trends in Figure 7. Compare these results with the pre-classification results in Figure 6 and note that the trend for the diseases is also different from the second-level classification results in Figure 9. We conclude that the most accurate top five diseases detected by the Sehaa system are the ones depicted in Figure 9 for "afflicted by" cases, that is, dermal diseases (DRD), heart diseases (HRD), hypertension (HYT), cancer (CNR), and diabetes (DBT).

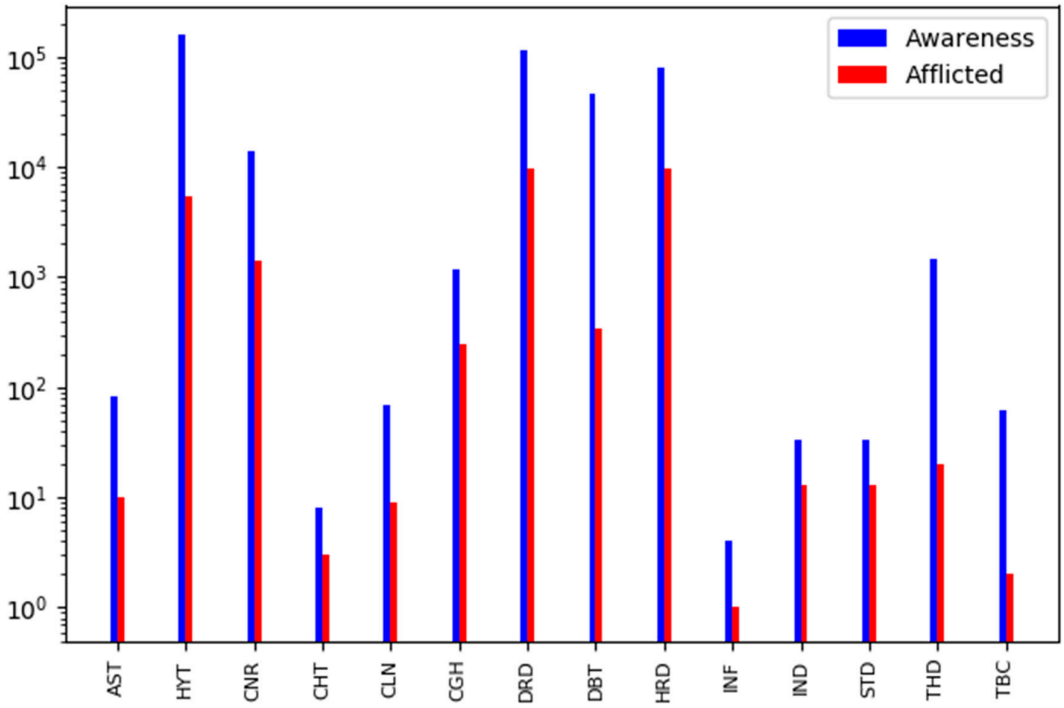

Figure 9. Sehaa: awareness vs. afflicted cases.

Figure 10 depicts the ratio of awareness and afflicted cases for Saudi Arabia and its five major cities. We infer from these values the level of awareness activities being carried out by various cities in comparison to the occurrence of each of the fourteen diseases. We expect bigger cities to carry out a higher number of awareness activities, as is usually the case due to the larger population, higher education levels, and use of technologies (Twitter). However, we reiterate that Figure 10 depicts the ratio and not the number of awareness activities. Note in Figure 10 that Riyadh has the highest ratio of awareness to afflicted cases for six of the fourteen diseases: AST, HYT, CNR, CGH, DBT, and THD. Interestingly, the top two diseases—dermal diseases (DRD) and heart diseases (HRD)—are not included in these six diseases, implying that Riyadh should do more in creating awareness for these diseases.

Jeddah is considered the second major city in Saudi Arabia. It has the highest ratio of awareness to afflicted cases for three of the fourteen diseases: CHT (cholesterol), CLN (colon), and TBC (tuberculosis). None of these are among the top five diseases in Saudi Arabia. Jeddah needs to do more in creating awareness for the top five national diseases. Taif is the fifth major city with a population that is one-eighth of that of the largest city Riyadh. However, it is commendable it has a comparable ratio of awareness to afflicted cases for several diseases. Note in Figure 7 that seven out of 14 diseases were not detected in Taif city and hence these diseases have zero values in Figure 10. Considering the two facts together, Taif has the lowest number of disease cases in Saudi Arabia while maintaining a high number of awareness activities.

Compare further Figures 7 and 10 and note that the national ratios for the fourteen diseases are significantly lower compared to the other cities (which is not the case in Figure 7). This shows that most of the awareness activities are happening in the five major cities and more effort is needed in other cities in Saudi Arabia.

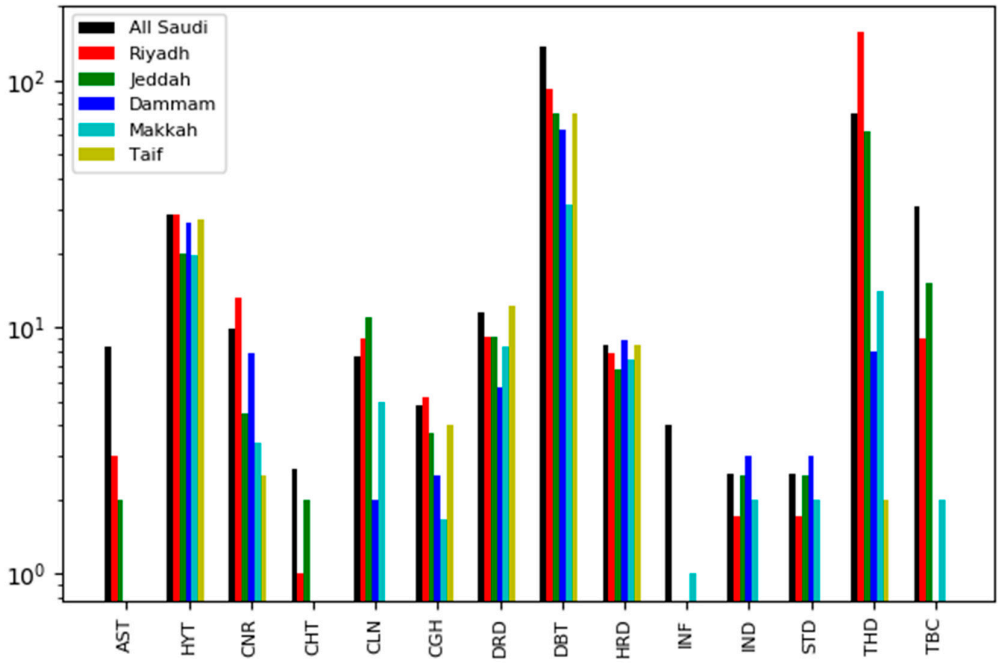

Figure 10. Sehaa: the ratio of awareness tweets to the afflicted cases.

6. Results Validation

We now discuss the numerical evaluation results of the classifiers and the external validation for the Sehaa system.

6.1. Numerical Evaluation

The Classification module provides critical functionality for the Sehaa system (see Section 4.4). The classification is accomplished in two levels. The first level is to distinguish the health-related tweets from the unrelated ones, while the second level is to classify the related tweets into awareness or afflicted tweets (see Table 2). At each level, the classification relies on machine learning algorithms.

In order to select the most efficient algorithm, we used different machine learning algorithms in the Spark ML packages, using Python with different feature extraction techniques to train the classification models. We split the manually labeled data into training sets (60%) and testing sets (40%). We built different model pipelines and trained these models using the Naïve Bayes (NB) and logistic regression (LR) algorithms. To find the best algorithm, we numerically evaluated them using the testing data set using the well-known evaluation criteria Accuracy and F1-score (see Section 4.5).

These scores for the first-level and second-level classifiers are plotted in Figures 11 and 12, respectively.

Figure 11 shows that Naïve Bayes with Trigram feature extraction provided the highest Accuracy (78.2%) compared to any other combinations of feature extraction techniques with Naïve Bayes or Logistic Regression. However, Logistic Regression with Trigram feature extraction provided the highest F1-score among all the classification and feature extraction methods. We had selected Naïve Bayes with Trigram (due to the higher Accuracy score) for the results presented in the previous sections.

Figure 12 shows the results for the second-level classifier. Note that it provides higher accuracies than the first-level classifier. For both Accuracy (86.7%) and F1-score (85.6%), Logistic Regression with HashingTF provided the best results, and therefore it was selected for the second-level classification.

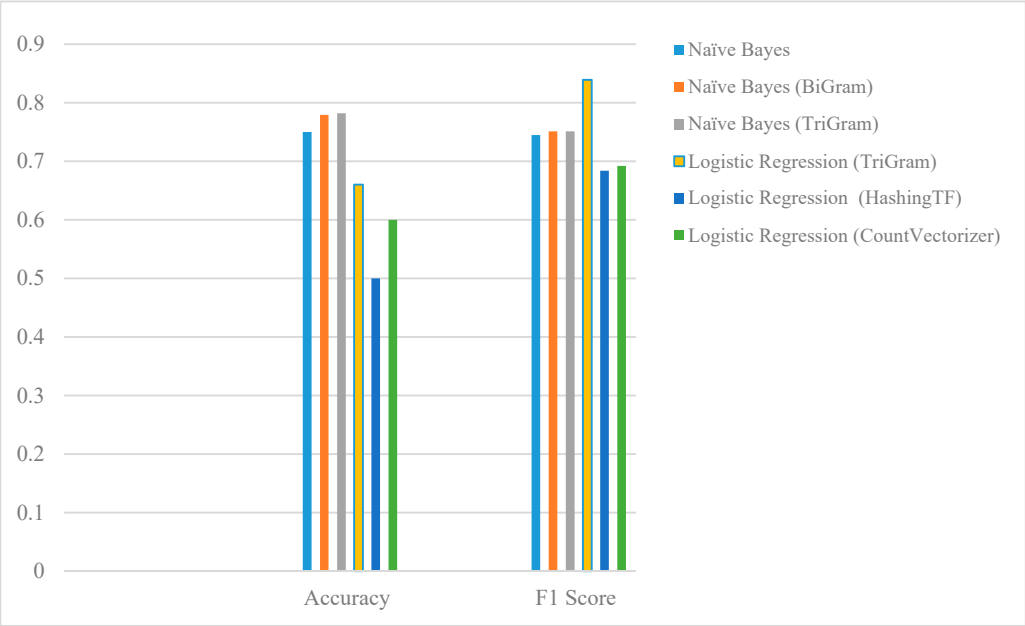

Figure 11. Sehaa: numerical evaluation of first-level classifiers and feature extraction methods.

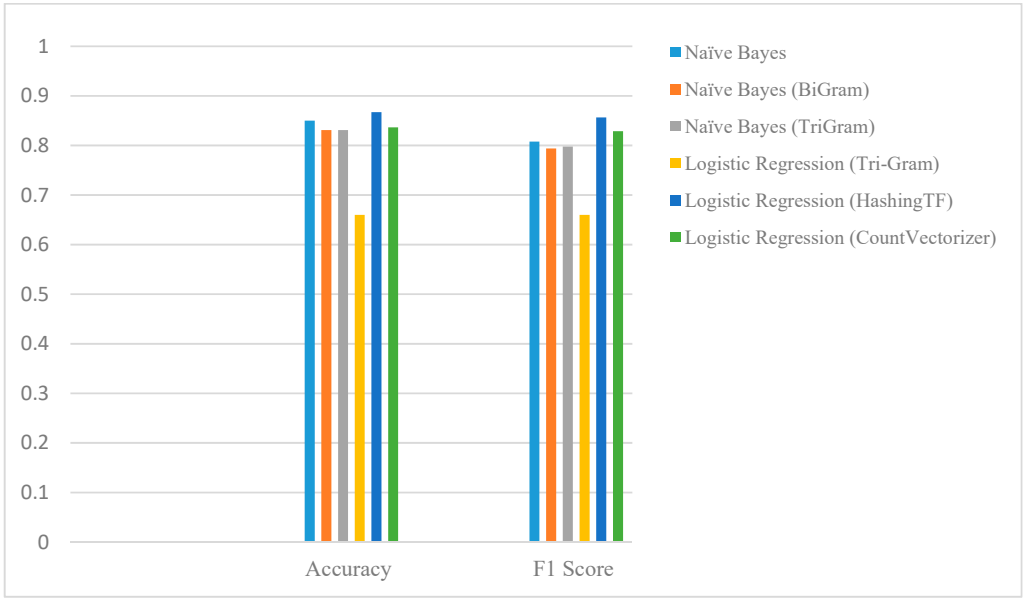

Figure 12. Sehaa: numerical evaluation of second-level classifiers and feature extraction methods.

6.2. External Validation

We have attempted to validate the results obtained from the Sehaa system against external sources, including reports from various relevant organizations, news media, and results reported in research articles. This involved comparing the statistics reported in various sources and the statistics reported from Sehaa. Unfortunately, we found limited information related to health and diseases in Saudi Arabia in the various media. The information is either incomplete or is old. Moreover, the existing information is limited to communicable diseases only. The sources we have looked at include the official website of the Saudi Ministry of Health [80], published articles from various organizations including the Institute for Health Metrics and Evaluation (IHME) (an independent global health research center at the University of Washington) [82], the official website of the World Health Organization (WHO) [83], and the Centers for Diseases Control and Prevention (CDCP).

According to the latest report published by CDCP [84], heart disease was considered among the top causes of death in Saudi Arabia in 2018. The Institute for Health Metrics and Evaluation reported that the risk of death from heart diseases increased by 36% from 2007 to 2017 [82]. These data confirm the findings of the Sehaa tool, where heart disease was detected as the second top disease in Saudi Arabia (see Section 5.3).

Al-Nozha et al. [85] reported in 1997 that more than a quarter of Saudi adults were suffering from hypertension. Aljohani reported, based on a study considering all hospitalized patients in one of Jeddah's hospitals, that hypertension was the sixth most widely prevalent disease in 2010 [86]. The Saudi Ministry of Health reported in 2017 that there was a remarkable increase in hypertension cases among Saudi adults [87]. These findings from as early as 1997 until recently show the high prevalence of hypertension in Saudi Arabia and hence a good correlation with the Sehaa findings.

MOH announced 1452 reported cases of Tuberculosis (TBC) in 2018 [88]; 451 of these cases were from Jeddah, 338 from Riyadh, and 21 from Taif. While this news item does not relate directly to the findings of the Sehaa tool, the low TBC numbers from Taif are in agreement with our findings for the low levels of diseases in Taif.

7. Conclusions

Smartness, which underpins smart cities and societies, is defined by our ability to engage with our environments, analyze them, and make decisions, all in a timely manner. Healthcare is the prime candidate needing the transformative capability of this smartness due to reasons including healthcare spending reaching a significant proportion of national GDPs (around one-fifth of the GDP in the US), an aging population, gross inefficiencies, and bad eating habits around the world, giving rise to the prevalence of lifelong diseases. With half of the world population connected to social networks, social media provides a vital solution for a ubiquitous and timely engagement among healthcare stakeholders. Twitter is one of the most popular social media platforms today. 500 million tweets are sent every day. Saudi Arabia has the fifth largest number of Twitter users in the world.

Our focus in this research has been on the use of Twitter media for healthcare in Saudi Arabia with the aim to develop technologies that provide enhanced healthcare in the country. We provided an extensive review of the relevant literature and identified two major challenges: first, the rudimentary level of the existing research (in terms of scope, functionalities, and usability) on Twitter data analytics in healthcare in English, and particularly in other languages, including Arabic; second, the scalability and interoperability of the analytics tools for healthcare, such as the management, integration, and distributed computations of big data.

We proposed Sehaa, a big data analytics tool for healthcare in Saudi Arabia using Twitter data in Arabic. Sehaa used Naive Bayes and Logistic Regression and multiple feature extraction methods to detect various diseases in Saudi Arabia. Sehaa was able to successfully detect various diseases. The top five diseases in Saudi Arabia in terms of the actual afflicted cases are dermal diseases, heart diseases, hypertension, cancer, and diabetes. Riyadh and Jeddah need to do more in creating awareness about the top diseases. Taif is the healthiest city in the KSA in terms of the detected diseases and awareness activities. Sehaa is developed over Apache Spark, allowing true scalability. The results were evaluated using the well-known numerical criteria Accuracy and F1-Score,

obtaining 83.9% and 86.7% scores for the two classification stages. The Sehaa results were validated against externally available statistics and shown to have a good correlation with them. For example, heart disease was found to be one of the top causes of death in Saudi Arabia, as reported in external sources, and this was in agreement with Sehaa, which detected heart diseases among the top five diseases in the country. Taif was shown to have low disease occurrence in external media and this was also in agreement with the results obtained through Sehaa.

Sehaa is an excellent example of integrating artificial intelligence (AI), distributed big data computing, and human cognition, brought together as a convenient tool for the betterment of public health and the economy. The system methodology and design are generic and can be extended to other countries in the Arab world as well as globally. Our focus in this work is on Saudi Arabia and therefore the tool currently works with tweets only in the Arabic language (it can be used in other Arabic speaking countries, such as UAE, Kuwait, and Egypt). Potential users of this tool are hospitals and other healthcare organizations, ministries of health, pharmaceutical companies, and other healthcare stakeholders.

This study is the first of its kind in Saudi Arabia using Apache Spark and tweets in the Arabic language. Sehaa is an important step in developing data analytics tools for Twitter (and other social media) in Arabic. The use of a scalable distributed computing platform for big data (Apache Spark) in this paper is also an important step in the right direction. The future will see the integration of more and more disparate systems to allow global system optimization [89–92]; the use of open-source scalable distributed computing platforms is very important for this purpose. The integration of Social media data with other smart city systems for real-time healthcare analytics, planning, and operations is a grand challenge with unimaginable benefits and applications. Another important challenge in Twitter data analytics is the labeling of data or Tweets. In this paper, we have used manual labeling, which is an extremely time- and resource-consuming task. An alternative is to use semi-supervised or unsupervised machine learning to automatically label large amounts of data (see, e.g., [93]). These methods are in their infancy and are limited due to low accuracies. Further investigation is planned for developing automatic labeling methods. Future work will work in these directions and improve the scope, functionality, scalability, analytics, data management, productivity, usability, and accuracy of the tool.

Modern living includes ubiquitous use of smartphones, wearables such as smartwatches, and other mobile devices. The concept of Smartness that we have discussed in this paper, i.e., our ability to engage with our environments, analyze them, and make decisions, requires embedding mobile devices and sensors in our environments. These sensor-rich environments undoubtedly have disadvantages in terms of the security and privacy risks they pose to us [94]. Twitter data, which is the focus of this research, is already public and our analysis only reports results on the population level; thus we believe that the privacy risks would either be non-existent or would be of a minor nature. However, this is an important concern and should be properly investigated. We have some background in developing privacy-preserving technologies [59,95–97], and we plan to look further into the privacy issues related to Twitter data and propose solutions to minimize these privacy risks.

Author Contributions: "conceptualization, S.A. and R.M.; methodology, S.A. and R.M.; software, S.A. and R.M.; validation, S.A. and R.M.; formal analysis, S.A. and R.M.; investigation, S.A. and R.M.; resources, S.A. and R.M.; data curation, S.A. and R.M.; writing—original draft preparation, S.A.; writing—review and editing, R.M. and O.R.; visualization, S.A. and R.M.; supervision, R.M. and I.K.; project administration, R.M. and A.A; funding acquisition, R.M., A.A, I.K., O.R.". All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant number RG-6-611-40. The authors, therefore, acknowledge with thanks the DSR for their technical and financial support.

Acknowledgments: The work carried out in this paper was supported by the HPC center at King Abdulaziz University.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mehmood, R.; Katib, S.S.I.; Chlamtac, I. (Eds.) *Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies*; EAI/Springer Innovations in Communication and Computing, Springer International Publishing, Springer Nature Switzerland AG: 2020.
2. Just How Big Is the Healthcare Industry? Here's What You Need to Know—Dreamit Ventures. Available online: https://www.dreamit.com/journal/2018/4/24/size-healthcare-industry (accessed on 8 February 2020).
3. Getting the Right Care to the Right People at the Right Cost: An Interview With Ron Walls | McKinsey. Available online: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/getting-the-right-care-to-the-right-people-at-the-right-cost-an-interview-with-ron-walls (accessed on 8 February 2020).
4. Sherman, E. U.S. Health Care Spending Hit $3.65 Trillion in 2018. *Fortune*, 21 February 2019. Available online: https://fortune.com/2019/02/21/us-health-care-costs-2/ (accessed on 12 January 2020).
5. Finding the Future of Care Provision: The Role of Smart Hospitals | McKinsey. Available online: https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/finding-the-future-of-care-provision-the-role-of-smart-hospitals (accessed on 8 February 2020).
6. Kemp, S. Digital Trends 2019: Every Single Stat You Need to Know about the Internet. *thenextweb.com*, 30 January 2019. Available online: https://thenextweb.com/contributors/2019/01/30/digital-trends-2019-every-single-stat-you-need-to-know-about-the-internet/ (accessed on 10 January 2020).
7. Statista. Countries with Most Twitter Users 2019 | Statistic. *Statista*, 20 November 2019. Available online: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed on 19 April 2019).
8. Lin, Y. 10 Twitter Statistics Every Marketer Should Know in 2020. *Oberlo*, 30 November 2019. Available online: https://www.oberlo.com/blog/twitter-statistics (accessed on 11 January 2020).
9. Twitter by the Numbers (2019): Stats, Demographics & Fun Facts. *Omnicore*, 10 February 2020. Available online: https://www.omnicoreagency.com/twitter-statistics/ (accessed on 11 January 2020).
10. Alotaibi, S.; Mehmood, R.; Katib, I. Sentiment Analysis of Arabic Tweets in Smart Cities: A Review of Saudi Dialect. In Proceedings of the 2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC), Rome, Italy, 10–13 June 2019; pp. 330–335.
11. Gohil, S.; Vuik, S.; Darzi, A. Sentiment analysis of health care tweets: Review of the methods used. *J. Med. Internet Res.* **2018**, *4*, 43.
12. AlSukhni, E.; Alequr, Q. Investigating the Use of Machine Learning Algorithms in Detecting Gender of the Arabic Tweet Author. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 319–328.
13. Al-Hussaini, H.; Al-Dossari, H. Lexicon-based Approach to Build Service Provider Reputation from Arabic Tweets in Twitter. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 445–454.
14. Al-Ayyoub, M.; Khamaiseh, A.A.; Jararweh, Y.; Al-Kabi, M.N. A comprehensive survey of arabic sentiment analysis. *Inf. Process. Manag.* **2019**, *56*, 320–342.
15. Alayba, A.M.; Palade, V.; England, M.; Iqbal, R. Arabic Language Sentiment Analysis on Health Services. In Proceedings of the International Workshop on Arabic and derived Script Analysis and Recognition, Nancy, France, 3–5 April 2017; pp. 114–118.
16. Alkouz, B.; Al Aghbari, Z. Analysis and prediction of influenza in the UAE based on Arabic tweets. In Proceedings of the 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA 2018), Shanghai, China, 9–12 March 2018; pp. 61–66.
17. Ilyas, M.U.; Alowibdi, J.S. Disease Tracking in GCC Region Using Arabic Language Tweets. In Proceedings of the Companion of the Web Conference 2018—WWW'18, Lyon, France, 13–17 April 2018; pp. 417–423.
18. Alomari, E.; Mehmood, R.; Katib, I. Sentiment Analysis of Arabic Tweets for Road Traffic Congestion and Event Detection. In *Smart Infrastructure and Applications*; Springer: Cham, Switzerland, 2020; pp. 37–54.
19. Suma, S.; Mehmood, R.; Albeshri, A. Automatic Detection and Validation of Smart City Events Using HPC and Apache Spark Platforms. In *Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies*, Springer: Cham, Switzerland, 2019; pp. 55–78.
20. Alomari, E.; Mehmood, R.; Katib, I. Road Traffic Event Detection Using Twitter Data, Machine Learning, and Apache Spark. In Proceedings of the 3rd IEEE International Conference on Smart City Innovations (SCI 2019), Leicester, UK, 19–23 August 2019.

21. Lau, R.Y. Toward a social sensor based framework for intelligent transportation. In Proceedings of the 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), Macau, China, 12–15 June 2017; pp. 1–6.
22. Pandhare, K.R.; Shah, M.A. Real time road traffic event detection using Twitter and spark. In Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 10–11 March 2017; pp. 445–449.
23. Salas, A.; Georgakis, P.; Nwagboso, C.; Ammari, A.; Petalas, I. Traffic Event Detection Framework Using Social Media. In Proceedings of the IEEE International Conference on Smart Grid and Smart Cities, Singapore, 23–26 July 2017; pp. 303–307.
24. Chen, M.; Mao, S.; Liu, Y. Big data: A survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209.
25. Mehmood, R.; Faisal, M.A.; Altowaijri, S. Future Networked Healthcare Systems: A Review and Case Study. In *Big Data: Concepts, Methodologies, Tools, and Applications*; Information Resources Management Association, Ed.; IGI Global, US, 2016; pp. 2429–2457.
26. "Apache Spark™ - Unified Analytics Engine for Big Data." [Online]. Available: https://spark.apache.org/. (Accessed on 28 December 2019)
27. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning : From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
28. "Logistic Regression — ML Glossary documentation." Available Online: https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html. (Accessed on 29 December 2019)
29. ""Graphical Models Lecture 2: Bayesian Network Representatioon."" Available Online: https://people.cs.umass.edu/~mccallum/courses/gm2011/02-bn-rep.pdf (Accessed: 02 January 2020).
30. Extracting, Transforming and Selecting Features—Spark 2.4.4 Documentation. Available Online: https://spark.apache.org/docs/latest/mL-features#tf-idf (accessed on 7 February 2020).
31. Mehmood, R.; Bhaduri, B.; Katib, I.; Chlamtac, I. (Eds.) Smart Societies, Infrastructure, Technologies and Applications. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST)*; Jeddah, Saudi Arabia, 27 - 29 Nov 2017, Springer, Cham: 2018; Volume 224.
32. Muhammed, T.; Mehmood, R.; Albeshri, A. Enabling reliable and resilient IoT based smart city applications. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (LNICST)*; Jeddah, Saudi Arabia, 27 - 29 Nov 2017, Springer, Cham: 2018; Volume 224, pp. 169–184.
33. Alam, F.; Mehmood, R.; Katib, I.; Albogami, N.N.; Albeshri, A. Data Fusion and IoT for Smart Ubiquitous Environments: A Survey. *IEEE Access* **2017**, *5*, 9533–9554.
34. Muhammed, T.; Mehmood, R.; Albeshri, A.; Katib, I. UbeHealth: A personalized ubiquitous cloud and edge-enabled networked healthcare system for smart cities. *IEEE Access* **2018**, *6*, 32258–32285.
35. Muhammed, T.; Mehmood, R.; Albeshri, A.; Alzahrani, A. HCDSR: A Hierarchical Clustered Fault Tolerant Routing Technique for IoT-Based Smart Societies. In *Smart Infrastructure and Applications*; Springer: Cham, Switzerland, 2020; pp. 609–628.
36. Mehmood, R.; Alam, F.; Albogami, N.N.; Katib, I.; Albeshri, A.; Altowaijri, S.M. UTiLearn: A Personalised Ubiquitous Teaching and Learning System for Smart Societies. *IEEE Access* **2017**, *5*, 2615–2635.
37. Alomari, K.M.; ElSherif, H.M.; Shaalan, K. Arabic Tweets Sentimental Analysis Using Machine Learning. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2017; pp. 602–610.
38. Alomari, E.; Mehmood, R. *Analysis of Tweets in Arabic Language for Detection of Road Traffic Conditions*; Springer: Cham, Switzerland, 2018; pp. 98–110.
39. Mehmood, R.; Graham, G. Big Data Logistics: A health-care Transport Capacity Sharing Model. *Procedia Comput. Sci.* **2015**, *64*, 1107–1114.
40. Mehmood, R.; Meriton, R.; Graham, G.; Hennelly, P.; Kumar, M. Exploring the influence of big data on city transport operations: A Markovian approach. *Int. J. Oper. Prod. Manag.* **2017**, *37*, 75–104.
41. Arfat, Y.; Usman, S.; Mehmood, R.; Katib, I. *Big Data Tools, Technologies, and Applications: A Survey*; Springer: Cham, Switzerland, 2020; pp. 453–490.
42. Arfat, Y.; Usman, S.; Mehmood, R.; Katib, I. *Big Data for Smart Infrastructure Design: Opportunities and Challenges*; Springer: Cham, Switzerland, 2020; pp. 491–518.
43. Arfat, Y.; Suma, S.; Mehmood, R.; Albeshri, A. *Parallel Shortest Path Big Data Graph. Computations of US Road Network Using Apache Spark: Survey, Architecture, and Evaluation*; Springer: Cham, Switzerland, 2020; pp. 185–214.

44. Usman, S.; Mehmood, R.; Katib, I. *Big Data and HPC Convergence for Smart Infrastructures: A Review and Proposed Architecture*; Springer: Cham, Switzerland, 2020; pp. 561–586.
45. Muhammed, T.; Mehmood, R.; Albeshri, A.; Katib, I. SURAA: A Novel Method and Tool for Loadbalanced and Coalesced SpMV Computations on GPUs. *Appl. Sci.* **2019**, *9*, 947.
46. Alyahya, H.; Mehmood, R.; Katib, I. Parallel Iterative Solution of Large Sparse Linear Equation Systems on the Intel MIC Architecture. In *Smart Infrastructure and Applications*; Springer: Cham, Switzerland, 2020; pp. 377–407.
47. Usman, S.; Mehmood, R.; Katib, I.; Albeshri, A.; Altowaijri, S.M. ZAKI: A Smart Method and Tool for Automatic Performance Optimization of Parallel SpMV Computations on Distributed Memory Machines. *Mob. Netw. Appl.* **2019**, 1–20, doi:10.1007/s11036-019-01318-3.
48. Usman, S.; Mehmood, R.; Katib, I.; Albeshri, A. ZAKI+: A Machine Learning Based Process Mapping Tool for SpMV Computations on Distributed Memory Architectures. *IEEE Access* **2019**, *7*, 81279–81296.
49. Arfat, Y.; Aqib, M.; Mehmood, R.; Albeshri, A.; Katib, I.; Albogami, N.; Alzahrani, A. Enabling Smarter Societies through Mobile Big Data Fogs and Clouds. *Procedia Comput. Sci.* **2017**, *109*, 1128–1133.
50. Mehmood, R.; Faisal, M.A.; Altowaijri, S. Future Networked Healthcare Systems: A Review and Case Study. In *Handbook of Research on Redesigning the Future of Internet Architectures*; Boucadair, M., Jacquenet, C., Eds.; IGI Global: Hershey, PA, USA, 2015; pp. 531–558.
51. Lo'ai, A.T.; Bakhader, W.; Mehmood, R.; Song, H. Cloudlet-Based Mobile Cloud Computing for Healthcare Applications. In Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 4–8 December 2016; pp. 1–6.
52. Schlingensiepen, J.; Mehmood, R.; Nemtanu, F.C.; Niculescu, M. Increasing Sustainability of Road Transport in European Cities and Metropolitan Areas by Facilitating Autonomic Road Transport Systems (ARTS). In *2013 Proceedings of the 5th International Conference on Sustainable Automotive Technologies (ICSAT 2013)*; Ingolstadt, Germany, 25 – 27 September 2013, Springer, Cham: 2014; pp. 201–210.
53. Alam, F.; Mehmood, R.; Katib, I.; Altowaijri, S.M.; Albeshri, A. TAAWUN: A Decision Fusion and Feature Specific Road Detection Approach for Connected Autonomous Vehicles. *Mob. Netw. Appl.* **2019**, 1–17, doi:10.1007/s11036-019-01319-2.
54. Alotaibi, S.; Mehmood, R.; Katib, I. The Role of Big Data and Twitter Data Analytics in Healthcare Supply Chain Management. In *Smart Infrastructure and Applications*; Springer: Cham, Switzerland, 2020; pp. 267–279.
55. Alamoudi, E.; Mehmood, R.; Albeshri, A.; Gojobori, T. A Survey of Methods and Tools for Large-Scale DNA Mixture Profiling. In *Smart Infrastructure and Applications*; Springer: Cham, Switzerland, 2020; pp. 217–248.
56. Alotaibi, S.; Mehmood, R. Big data enabled healthcare supply chain management: Opportunities and challenges. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (LNICST)*; Springer: Cham, Switzerland, 2018; Volume 224, pp. 207–215.
57. Aqib, M.; Mehmood, R.; Alzahrani, A.; Katib, I.; Albeshri, A.; Altowaijri, S.M. Altowaijri. Smarter Traffic Prediction Using Big Data, In-Memory Computing, Deep Learning and GPUs. *Sensors* **2019**, *19*, 2206.
58. Aqib, M.; Mehmood, R.; Alzahrani, A.; Katib, I.; Albeshri, A.; Altowaijri, S.M. Rapid Transit Systems: Smarter Urban Planning Using Big Data, In-Memory Computing, Deep Learning, and GPUs. *Sustainability* **2019**, *11*, 2736.
59. Al-Dhubhani, R.; Mehmood, R.; Katib, I.; Algarni, A. Location Privacy in Smart Cities Era. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*; Springer: Cham, Switzerland, 2018; Volume 224, pp. 123–138.
60. Khanum, A.; Alvi, A.; Mehmood, R. Towards a semantically enriched computational intelligence (SECI) framework for smart farming. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*; Springer: Cham, Switzerland, 2018; Volume 224, pp. 247–257.
61. Liu, B. Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167.
62. Andreu-Perez, J.; Poon, C.C.; Merrifield, R.D.; Wong, S.T.; Yang, G.Z. Big Data for Health. *IEEE J. Biomed. Heal. Inf.* **2015**, *19*, 1193–1208.
63. Parker, J.; Yates, A.; Goharian, N.; Frieder, O. Health-related hypothesis generation using social media data. *Soc. Netw. Anal. Min.* **2015**, *5*, 1–15.
64. Paul, M.J.; Dredze, M. A model for mining public health topics from Twitter. *Health* **2012**, *11*, 1.

65. Paul, M.J.; Dredze, M. You are what you Tweet: Analyzing Twitter for public health. In Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM-2011), Barcelona, Spain, 17–21 July 2011, pp. 265–272.
66. Aramaki, E. Twitter Catches the Flu : Detecting Influenza Epidemics Using Twitter. *Comput. Linguist.* **2011**, *2011*, 1568–1576.
67. Wakamiya, S.; Kawai, Y.; Aramaki, E. Twitter-based influenza detection after flu peak via tweets with indirect information: Text mining study. *J. Med. Internet Res.* **2018**, *4*, 65.
68. Wakamiya, S.; Morita, M.; Kano, Y.; Ohkuma, T.; Aramaki, E. Tweet classification toward twitter-based disease surveillance: New data, methods, and evaluations. *J. Med. Internet Res.* **2019**, *21*, e12783.
69. Lamb, A.; Paul, M.; Dredze, M. Separating fact from fear: Tracking flu infections on Twitter. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 789–795.
70. Smith, M.; Broniatowski, D.A.; Paul, M.J.; Dredze, M. Towards Real-Time Measurement of Public Epidemic Awareness: Monitoring Influenza Awareness through Twitter. In Proceedings of the AAAI Workshop on World Wide Web and Public Health Intelligence, Austin, TX, USA, 25–26 January 2015; Volume 20052.
71. Bian, J.; Topaloglu, U.; Yu, F. Towards large-scale twitter mining for drug-related adverse events. In Proceedings of the 2012 International Workshop on Smart Health and Wellbeing 2012, Maui, HI, USA, 29 October 2012; pp. 25–32, doi:10.1145/2389707.2389713.
72. Myslín, M.; Zhu, S.H.; Chapman, W.; Conway, M. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. *J. Med. Internet Res.* **2013**, *15*, e174.
73. Jashinsky, J.; Burton, S.H.; Hanson, C.L.; West, J.; Giraud-Carrier, C.; Barnes, M.D.; Argyle, T. Tracking Suicide Risk Factors through Twitter in the US. *Crisis* **2014**, *35*, 51–59.
74. Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.H.; Liu, B. Twitter Improves Seasonal Influenza Prediction. In Proceedings of the International Conference on Health Informatics (HEALTHINF 2012), Vilamoura, Algarve, 1–4 February 2012; pp. 61–70, doi:10.5220/0003780600610070
75. Broniatowski, D.A.; Paul, M.J.; Dredze, M. National and local influenza surveillance through twitter: An analysis of the 2012–2013 influenza epidemic. *PLoS ONE* **2013**, *8*, e83672.
76. Ram, S.; Zhang, W.; Williams, M.; Pengetnze, Y. Predicting Asthma-Related Emergency Department Visits Using Big Data. *IEEE J. Biomed. Heal. Inf.* **2015**, *19*, 1216–1223,;
77. Culotta, A. Detecting influenza outbreaks by analyzing Twitter messages. *arXiv* **2009**, arXiv:1007.4748.
78. Suma, S.; Mehmood, R.; Albugami, N.; Katib, I.; Albeshri, A. Enabling Next Generation Logistics and Planning for Smarter Societies. *Procedia Comput. Sci.* **2017**, *109*, 1122–1127.
79. Suma, S.; Mehmood, R.; Albeshri, A. Automatic event detection in smart cities using big data analytics. In *International Conference on Smart Cities, Infrastructure, Technologies and Applications (SCITA 2017): Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*; Springer: Cham, Switzerland, 2017; Volume 224, pp. 111–122.
80. Statistical Yearbook. Available Online:https://www.moh.gov.sa/en/Ministry/Statistics/book/Pages/default.aspx. (Accessed: 06 November 2019]
81. Suthaharan, S. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. *Integr. Ser. Inf. Syst*. **2015**, *36*, 1–12.
82. "Saudi Arabia | Institute for Health Metrics and Evaluation." Available Online: http://www.healthdata.org/saudi-arabia. (Accessed on 06 November 2019).
83. "WHO | Saudi Arabia." Available Online: https://www.who.int/countries/sau/en/. (Accessed on 06 November 2019)
84. "CDC Global Health-Saudi Arabia." Available Online: https://www.cdc.gov/globalhealth/countries/saudi_arabia/default.htm. (Accessed on 26 Novemebr 2019).
85. Al-Nozha, M.M.; Ali, M.S.; Osman, A.K. Arterial hypertension in Saudi Arabia. *Ann. Saudi Med.* **1997**, *17*, 170–174.
86. Aljohani, H.A. Association between Hemoglobin Level and Severity of Chronic Periodontitis. *JKAU Med. Sci*. **2010**, *17*, 53–64.
87. "Health Days 2017—World Hypertension Day "Available Online:

    https://www.moh.gov.sa/en/HealthAwareness/healthDay/2017/Pages/HealthDay-2017-05-17.aspx

    (Accessed on 09 January 2020)

88. حالات الدرن الرئوي حسب المنطقة وفئة العمر خلال عام 1439 هـ (2018م) - البيانات - البوابة السعودية للبيانات المفتوحة." Available Online: https://data.gov.sa/Data/ar/dataset/pulmonary_tuberculosis_by_region-_age_group_during_1439h_-2018g-. (Accessed: 17 Decemeber 2019)
89. Ahmad, N.; Mehmood, R. Enterprise systems and performance of future city logistics. *Prod. Plan. Control.* **2016**, *27*, 500–513.
90. Ahmad, N.; Mehmood, R. Enterprise Systems for Networked Smart Cities. In *Smart Infrastructure and Applications*; Springer: Cham, Switzerland, 2020; pp. 1–33.
91. Graham, G.; Ahmad, N.; Mehmood, R. Enterprise systems: Are we ready for future sustainable cities. *Supply Chain Manag.* **2015**, *20*, 264–283.
92. How Data Science Is Shaping the Modern NHS. Available Online: https://www.newstatesman.com/science-tech/technology/2018/11/how-data-science-shaping-modern-nhs (accessed on 8 February 2020).
93. Shafiabady, N.; Lee, L.H.; Rajkumar, R.; Kallimani, V.P.; Akram, N.A.; Isa, D. Using unsupervised clustering approach to train the Support Vector Machine for text classification. *Neurocomputing* **2016**, *211*, 4–10.
94. Giraldo, J.; Sarkar, E.; Cardenas, A.A.; Maniatakos, M.; Kantarcioglu, M. Security and Privacy in Cyber-Physical Systems: A Survey of Surveys. *IEEE Des. Test.* **2017**, *34*, 7–17.
95. Ayres, G.; Mehmood, R. LocPriS: A security and privacy preserving location based services development framework. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNAI*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6279, pp. 566–575.
96. Ayres, G.; Mehmood, R.; Mitchell, K.; Race, N.J. Localization to enhance security and services in Wi-Fi networks under privacy constraints. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 16, pp. 175–188.
97. Al-Dhubhani, R.S.; Cazalas, J.; Mehmood, R.; Katib, I.; Saeed, F. A framework for preserving location privacy for continuous queries. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2020; Volume 1073, pp. 819–832.