Sentiment Analysis Model for Arabic Tweets to Detect Users' Opinions about Government Services in Saudi Arabia: Ministry of Education as a case study

Mashael M. Alsulami [1], and Rashid Mehmood[2]
[1] Computer Science Department, King Abdulaziz University, Saudi Arabia
[2] High Performance Computing Center, King Abdulaziz University, Saudi Arabia
malsulami0485@stu.kau.edu.sa[1] , RMehmood@kau.edu.sa[2]

Abstract:

With the rapid growth of social media usage among citizens, the amount of daily data has substantially increased. Using these data to gain knowledge of citizens' experiences with different services is a potential task for many organizations to improve their services and to discover emerging requirements. Twitter is one of the most widely used platforms as a source of big data to predict, detect, or recommend various actions based on the available information. The government of Saudi Arabia invests a lot of resources in terms of time, human capital and money to enhance the performance of all government ministries in the country. The movement towards evaluating the performance of provided services is essential for all ministries in Saudi Arabia.

This study considers the Ministry of Education as a case study to design a sentiment analysis model to examine tweets related to services provided by the ministry. The aim of this work is to inform decision-makers, the issues related to the performance of the ministry by exploring the users' opinions about their services. In particular, one of the most recent hashtags related to the Ministry of Education is examined which is #New_University_System. The sentiment analysis model aims to categorize the users' attitudes about the new university system that has been launched recently. It will classify the Arabic tweets regarding the posted hashtag into three categories: Positive, Negative, and Neutral. Also, the study investigates the users' expectations about the new service. The results will help decision makers to find the gap between users' needs (reflected by their opinions and expectations) and what the government's expectations in applying the new university system.

Keywords- *Sentiment analysis, Arabic sentiment analysis, SAP HANA platform, predicting social media analytics, and social networks.*

I. INTRODUCTION

A. Motivation

With the increasing demand to provide better governmental services to citizens in Saudi Arabia, the need to explore their need is substantially necessary. Recently, Saudi Arabia moves towards the vision of 2030 where every information counts and could be beneficial in the development process of the country. Ministries in Saudi Arabia are responsible of serving citizens and provide services that are suitable to each individual no matter what his or her gender or social level. From that perspective, leaders in Saudi Arabia have pay a very close attention to the quality of services provided to citizens in different areas such as health, education, and housing. Evaluating the services of such areas may result in two main benefits: discovering new needs, and enhancing existing services. People used to chat about these governmental services in their family/friends gathering. However, with the new world we live in, these informal chats move from our houses, coffee shops, schools, and places of work to the virtual world. Social media becomes the source of information to explore the individual's point of view and opinion about certain topics. Thus, the motivation of this study is to examine if Twitter can be used as a source of information to explore citizens' needs, opinion and expectations about services provided by different ministries in Saudi Arabia. Ministry of Education is considered in the scope of this study by examining a new launched service provided to Saudi universities and it is called New University System. This service is still under investigation by decision makers. This study aims to address twitter users point of view about this new system. Also, it shows what users are expected from this new service. We

believe that this kind of information could be beneficial for decision makers in Ministry of Education if they consider them carefully.

B. Background information

Recently, Twitter data has gained so much attention from researchers as a source of information. It is mentioned in [1] that number of tweets per day has been increased to reach over five hundred million messages. Twitter users use twitter daily for many reasons such as socializing with others, navigation for information or simply for fun. Moreover, a wide variety of topics are discussed daily in Twitter ranging from simple sleeping note to a more complicated topic regarding politics, technology…etc.

Many companies and organizations invest in analyzing tweets about their products to gain a clear vision of what their customers need and think about their current products. Reviewing products is one of the tasks that use Twitter as a source of information [2]. Tweets about a coffee taste, a t-shirt brand or even a software application may consider as insightful and valuable piece of information to be used by the relevant company or organization. However, little information is known about how users interact with governmental services provided by different ministries in Saudi Arabia. Thus, this study is one of the first to investigate the possibility and benefit of analyzing Arabic tweets about certain governmental services to extract user's point of view and expectations about these services.

In this study, Ministry of Education is considered and more precisely we are examining the new university system that has been launched recently and has not activated yet. The idea behind this service is to classify the Saudi universities into three categories. This classification will help both students and faculty members to customize their specialists and fit in the right places that match their capabilities. A formal agenda has been released and made accessible by public to understand the proposed system. Also, Ministry of Education allows faculty members in Saudi Arabia to register their recommendations and comments on the new system. Citizens in Saudi Arabia have very ranged opinions where some have strongly agreed and other disagreed. Also, some citizens have not read the agenda and others have not understood it correctly. These different expectations are investigated in this study along with the sentiment polarity about service under consideration.

C. *Organization of this paper*

This paper is structured as follow. Section 2 is a literature review that consists of three main subsections: social media analytics, sentiment analysis techniques, and Arabic sentiment analysis. Section 3 discusses the proposed methodology that is used in the analysis. Section 4 shows the main results of the analysis. Finally, a conclusion with future work is given in section 5.

II. LITERATURE REVIEW

In this literature review, three main directions are investigated: social media analytics SMA, sentiment analysis techniques, and Arabic sentiment analysis. In the first subsection, social media analytics is discussed including several studies that prove the effective use of social media data as a resource for information. Also, applications and tools that are used to perform social media analytics are explored. The second subsection addresses the existing techniques to perform sentiment analysis on social media data. Finally, the last subsection highlights the work towards Arabic sentiment analysis.

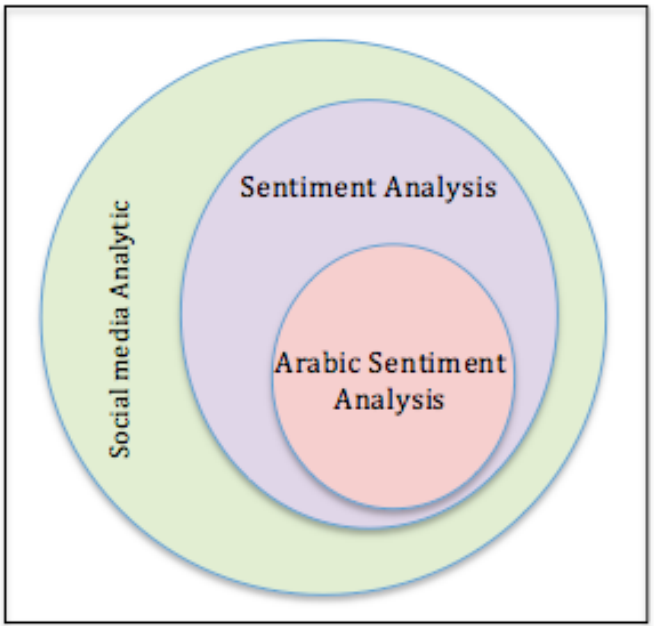

Figure 1. Overview of SMA techniques

A. Social Media Analytics (SMA)

Social media analytics is a term that includes several concepts and techniques such as social media filtering, social media classification, social network analysis, sentiment analysis, and engagement tracking [3]. The main goal of social media analytics is concentrating on social contents and how they can be used to provide a qualitative description for specific issues and a meaningful analysis in different domains [3].

The following studies are examples of using social media data to provide either one of the previously mentioned goals or both of them.

Authors in [4] have proposed an automatic monitoring system to detect patterns of abuse to specific medications. They use social media as a resource for their proposed system. Basically, their results prove that Twitter data can be used effectively in detecting signals of abuse. Also, authors in [1] have used Twitter data to record tweets about certain software applications. The analysis results can be used as valuable input for relevant companies to categorize each issue addressed by the analysis and deal with it. Another example to use social media data as a resource of information is a work proposed by [5]. They investigate how the emotion of customers can be used to impact the stock market. Mood extraction has been studied to find a relationship between the stock market and the general mood states in Twitter. In their study, they consider Twitter data as their primary dataset. The results of their study help increasing the portfolio of stock market up to 36% within six months period.

All the pervious works and more prove that social media data especially twitter data have been used primarily in analytics process in different areas and domains. Huge amount of daily data that have been published through social media are valuable contents in many cases.

The reset of this section focuses on applications and tools that are used to perform such analysis to social media contents.

Tools and applications that are used for that purpose can be classified according to [3] into two main categories:

business intelligence (BI) tools, and social media monitoring (SMM) tools. An example of a BI tool is SAP HANA platform, which will be discussed shortly while there are many SMM tools available such as TweetReach and Hotsuite. Both categories have certain capabilities to handle the nature of social contents. They are sharing some processing tasks in order to analyze social contents. These intersected processing tasks are: data extraction, data transformation and data analysis. In term of data extraction, social media account is required to use social media APIs while SMM tools do not require that kind of restriction. Another interesting feature regarding data extraction is that BI tools can retrieve social contents up to few days while SMM tools allow delivering social contents up to few years. In term of data transformation and load, the work is harder for BI tools since user-defined data load and no automated monitoring are available while SMM tools have a predefined database schema. In term of data analysis, classical BI tools need to set up the analysis manually along with definition and configuration of algorithms used in analysis. In SMM tools, data analysis is much easier since analysis algorithms are predefined for social contents [3].

B. Sentiment analysis techniques

As shown in figure 1, sentiment analysis is one of the main techniques used to analyze social contents. Many attentions have been given to sentiment analysis in the past few years under the context of natural language processing and text mining. Sentiment an be defined as an opinion, point of view, attitude, or feeling to a thing including topics, movies, products, or any other object or idea [6]. Sentiment analysis is the process of detecting certain words or phrases that indicate polarity of certain topics or objects. It is beneficial to perform sentiment analysis in marketing and research. Valuable information can be influenced from analyzing or reviewing user's opinion about certain topics, products or even ideas [6].

In order to perform sentiment analysis, there are two main approaches to do so: machine learning approach, and semantic orientation approach [7]. Machine learning approach is based on the fact of extraction features from an input text, which are converted later to certain vectors machine that used to represented useful information from the input text. There are several mechanisms for feature extraction based on machine learning approach. Some of these mechanisms are Part-of-Speech (POS) tags, Bag-of words, and n-grams [7].

In term of semantic orientation approach, there are two main techniques used in the literature: corpus-based technique, and dictionary or knowledge-based technique [8]. In term of corpus-based semantic orientation approach, polarity of terms is detected through the analysis of a large dataset to a selected domain. The main limitation of this approach is the need to have a training dataset where polarity terms are restricted to that corpus. On the other hand, the main strength for this approach is its simplicity [8]. In term of dictionary or knowledge-based approach, public linguistic databases are available on the Internet. These databases store information about each word, concept, and phrase along with the connection between them [9]. In order to show that connection, several representations have been used such as semantic networks [9].

Generally, to perform sentiment analysis on any social media APIs, there are four main phases must be considered as sentiment analysis workflow. These phases are: data collection, data preprocessing, data classification, and data visualization. Data collection phase is responsible of gathering tweets by using predefined keywords that are related to the target topic. In data preprocessing, a cleaning process is performed by identifying each word in the retrieved tweets from the first phase as a token and then delete duplicate tweets, remove emotions and stop words. In data classification phase, a classifier is selected to analyze user's sentiment about certain topics. Finally, visualization is presented to illustrate the results of the analysis [10].

C. Arabic sentiment analysis

According to [10] and recent reports, Arabic language has been growing so fast in Twitter. However, the studies that are conducting to investigate Arabic text mining and natural language processing for Arabic text are not compatible with the wide use of Arabic language in social media. The main reason of that is due to the complexity of

the structure of Arabic language with its special characteristics. Most of sentiment analysis model have been proposed are regarding other languages such as English [10].

There are some good researches that have studied Arabic sentiment analysis and find techniques to deal with its complexity. The importance of the study towards Arabic language is that it considers a mother language for almost 22 countries with more than 300 person specks it [11]. The Arabic language is classified into three categories as: classical Arabic, modern Arabic and dialectical Arabic [12]. Basically, the workflow of Arabic sentiment analysis follow the same workflow mentioned in the previous section. However, to create analysis of Arabic tweets, a customized dictionary must be created [12]. Most of the sentiment analysis approaches that consider Arabic as language for the analysis contents are based on the supervised approach [13]. Another approach that has been used with Arabic sentiment analysis is corpus annotation [14]. In term of annotation approach, there are three main techniques to annotate Arabic text: manual annotation technique, crowdsourcing annotation technique, and automatic annotation technique. In general, annotation can be done at sentence level, phrase level, or word level [14].

III. PROPOSED METHODOLOGY

An overview of the proposed methodology is shown in figure 2. The proposed analysis method consists of seven main phases:

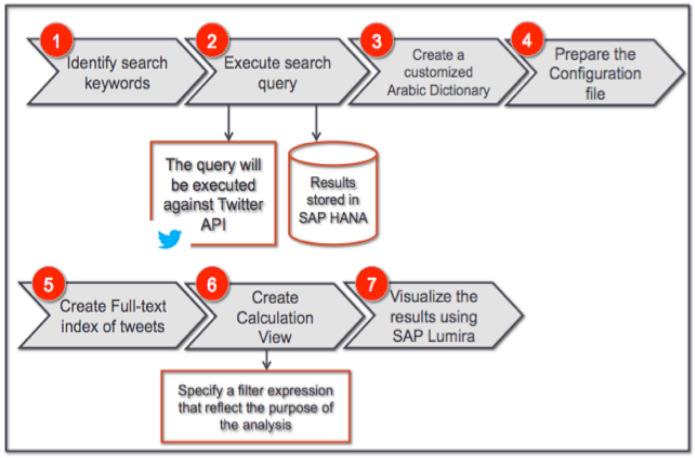

Figure 2. Proposed methodology

1. In order to extract twitter data regarding the selected topic, certain keywords have been identified such as #نظام-الجامعات-الجديد، مسودة الجامعات، مشروع الجامعات الجديد، مسودة نظام الجامعات

2. After identifying the previous keywords, they are included in the query that is executed against Twitter API.

3. The result of the previous query is a set of all tweets that include any of the identified keywords. The result is stored in a table in a predefined database schema inside SAP HANA platform.

4. To analyze Arabic tweets, a customized dictionary must be created. There are two main dictionaries have been created for the purpose of this project:
   - A dictionary that includes set of words categorized as positive, negative and neutral to show the sentiment polarity about the selected service.
   - A dictionary that contains a classification to show user's expectations of the new system. There is three categories that show what kind of changes the new system brings: changes related to faculty, changes related to the structure of administration, and changes related to students.

5. Each dictionary is added to a separate configuration file that is used to create a calculation view.

6. Full-text indexes of tweets are created with respect to the two configuration files.

7. Calculation views are created according to each configuration file. In each calculation view, a filter expression that reflects the purpose of that view is written and executed.

8. Finally, SAP Predictive analytics tool is used to visualize the results of the analysis.

A. SAP HANA Tool

It is a popular, an in-memory, and column oriented database server. It is mainly used as a database-sever to store and retrieve data from and/to on-premise, cloud, and hybrid applications. Also, it performs other functions such as reporting and analysis, data modeling, and provisioning. The datastore used in most in-memory database are column-oriented datastore. The idea of columnar-based database is to store data in columns rather than rows. There are several advantages of columnar-based databases such as better I/O bandwidth utilization, higher cache efficiency, faster data aggregation, high compression rates, and column-based parallel processing. On the other hands, it suffers from some disadvantages such as load times can be slow, less efficient for transactional processes, and possibly slower relational interfaces. SAP HANA is a platform more than a database. It can be described in term of its application services, processing services, integration/quality services and database services. In the bottom of SAP HANA architecture, several database services are provided such as data modeling, administration and security. In term of application services, SAP HANA provides services to applications in form of JAVA scripts, Web servers and others. In term of processing services, SAP HANA support wide range of processing services such as Spatial data processing, predictive analysis, text analysis, and kinds of data analysis and processing. Also, SAP HANA supports different types of integration services such as Hadoop and Spark integration, and data virtualization [15].

As being a database, SAP HANA allows to create different views from tables with a data-provisioning feature supported underneath. Data that are stored and processed by SAP HANA platforms could be pulled out from different data sources such as ERP and SCM. There are different tools with different functionalities supported by SAP HANA. Also, it provides different offerings in form of side-by-side scenario such as sales analysis for retail or integrated scenarios such as business one on HANA. There are different tools in SAP HNAN platform such as SAP NetWeaver BW, SAP NetWeaver BWA and SAP Lumira [15].

Business intelligence (BI) applications can potentially benefit from SAP HANA. Generally, SAP HANA Lumira can handle computations required by BI systems by performing the following tasks. First, preparing data by importing data from SAP HANA or and DB, transforming data, and merging data. Second, visualizing trends from big picture to details and with real-time support. Third, composing storyboards by composing stories/dashboards, adding dynamic filters, and including texts. Finally, sharing results by sending visualizations by email, exporting as file, and publishing to HANA, Business Objects Explorer, and Lumira Cloud/Server [15].

IV. RESULTS

Basically, the following charts show two main analysis results. Figure 3 show sentiment polarity of Twitter users about the new university service that has been launched recently by the Ministry of Education. Figure 4 shows the users' expectations about the new system. Basically, three categories have been identified: changes related to the structure of administration, changes related to students, and changes related to faculty.

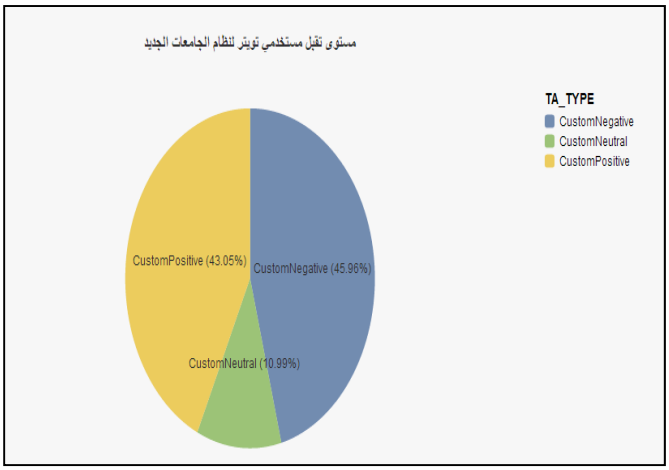

Figure 3. Sentiment polarity distribution about how Twitter user's accept #New_University_System

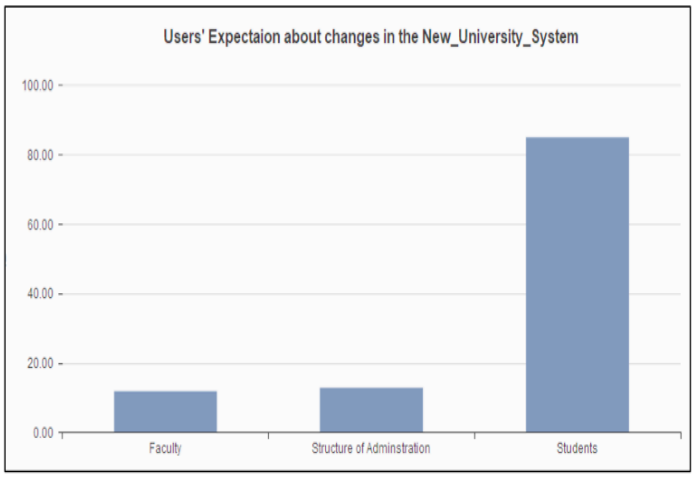

Figure 4. User's classification based on their expectations about #New_University_System

V. Conclusion and Future Work

In this study, we explore and investigate the applicability of Twitter data to be used as a resource of information to report users' opinions and expectations about certain governmental services and more precisely, the agenda of the new universities system. The goal of this study is to use Twitter data effectively to detect what users think and expect (or need) from the new university system, which launched by Ministry of Education. The results of this analysis can be useful for decision makers in the Ministry of Education to make further analysis regarding the new agenda. One of the main benefits of social media analysis is that it is not expensive process to perform. In this study, we found that Twitter is used to chat about different governmental services provided by different ministries in Saudi Arabia. Results shown that in our particular case of study, most of tweets were posted by official Twitter universities accounts. Also, accuracy of the analysis is one of the main limitations when using social media data as source of the analysis. Even with identifying a set of positive and negative words, it is hard to detect the sentiment of the whole tweet by the appearance of these identified words. We believe that more efforts need to be done to find automatic analysis approaches that take into consideration Arabic sentiment analysis techniques. Our future work will be focusing on enhancing the sentiment analysis of our study. The enhancements will involve engaging more vocabulary in the configurable dictionary file to cover more possibilities. Also, we will consider wide range of different governmental services in different ministries.

REFERENCES

[1] E. Guzman, R. Alkadhi, and N. Seyff, "An exploratory study of Twitter messages about software applications," *Requir. Eng.*, vol. 22, no. 3, pp. 387–412, 2017.

[2] P. C. Vinh, "Nature of Computation and Communication," *Mob. Networks Appl.*, vol. 21, no. 1, pp. 391–396, 2016.

[3] M. Wittwer, O. Reinhold, R. Alt, F. Jessen, and R. St, "Business Information Systems Workshops," vol. 263, pp. 252–259, 2017.

[4] A. Sarker *et al.*, "Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from twitter," *Drug Saf.*, vol. 39, no. 3, pp. 231–240, 2016.

[5] S. H. Archana and S. Godfrey Winster, "Drugs categorization based on sentence polarity analyzer for Twitter data," *2016 2nd Int. Conf. Sci. Technol. Eng. Manag. ICONSTEM 2016*, pp. 28–33, 2016.

[6] D. Barapatre, M. J. Meena, and S. P. S. Ibrahim, "Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC – 16')," vol. 49, pp. 363–368, 2016.

[7] N. Zainuddin, A. S. B, and R. Ibrahim, "Trends in Applied Knowledge-Based Systems and Data Science," vol. 9799, pp. 269–279, 2016.

[8] B. Agarwal and N. Mittal, "Prominent Feature Extraction for Sentiment Analysis," no. i, 2016.

[9] B. Schuller and T. Knaup, "Learning and knowledge-based sentiment analysis in movie review key excerpts," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6456 LNCS, pp. 448–472, 2011.

[10] M. A. B and M. Hadzikadic, "Social Computing and Social Media. Applications and Analytics," vol. 10283, pp. 191–202, 2017.

[11] E. Refaee, "Social Computing and Social Media. Applications and Analytics," vol. 10283, pp. 275–294, 2017.

[12] M. El-Masri, N. Altrabsheh, and H. Mansour, "Successes and challenges of Arabic sentiment analysis research: a literature review," *Soc. Netw. Anal. Min.*, vol. 7, no. 1, pp. 1–22, 2017.

[13] C. Xiao, Z. Qin, and X. Luo, "Web Information Systems Engineering – WISE 2016," vol. 10042, pp. 105–120, 2016.

[14] L. Almuqren, A. Alzammam, S. Alotaibi, and A. Cristea, "Social Computing and Social Media. Applications and Analytics," vol. 10283, pp. 215–225, 2017.

[15] C. Lorraine R. Gardiner, California State University, "Introduction to In-Memory Databases for Analytic Applications," *Sap Ua - Sap Hana*, vol. Chapter 1, no. August, 2016.