

Analyzing Tweets for Better Decision-Making using Machine Learning

Hazzaa N. Alshareef and Imran Usman

h.alshareef@seu.edu.sa i.usman@seu.edu.sa

College of Computing and Informatics, Saudi Electronic University,
Riyadh, Saudi Arabia

Summary

The adaptive advancement in internet technology convince individual to exchange the information. One of the most influential platforms to share contents over the World Wide Web is the social media. Social media includes Blogs, content forums, social networking sites and virtual environments. These are quickly becoming one of the most persuasive sources of news, economies, businesses, thoughts, emotions, feedback and reviews. The ceaseless swift development of electronic Arabic contents in social media diverts and in Twitter especially represents a chance for opinion mining analysis. This work represents a novel Naïve Bayes classification framework to classify and analyze positive, negative and neutral Arabic tweets regarding Saudi Electronic University.

Key words:

Arabic Tweets, Naïve Bayes Classifier, Sentiment Analysis

1. Introduction

In recent time, micro blogging is the new trend of communication. This phenomenon is primarily attributed to advances in computing power, internet bandwidth, and mobile devices. People all over the world exhibit their experience, views, opinion on social media sites so that folks get feedback or assistances from it. The interaction in the social media websites between the users around the world provided good and valuable information on 24/7 basis. This information can be used by any interested organizations, data scientists or any society to do their data analyzing, researches, and feasibility studies to have the correct understanding about the customers and targeted audiences. Such data mining procedure make businesses and organization able to have the right decision making and provide satisfactory services for the beneficiaries.

In the Arabic world, it is realized that there are huge communications and interaction between the users of social media networking websites, which become a way of life for many people. Twitter is one of the influential social media platforms in gulf countries where people share their opinion, views and reviews in short discourse about 280 characters. It had 290.5 million active users worldwide and expecting to increase up to 340 million users by 2024 [1].

A large part of users from the gulf countries are contributing and significant markets for Twitter. In the Arab world, Saudi Arabia has over 10 million active Twitter users. It is the huge market for Twitter with ranked 8th worldwide and 1st in gulf world respectively.

In Twitter, a special type of digital relationship ‘following-follower’ makes people transfer information rapidly. In order to receive messages, user has to subscribe the other users and become his/her ‘follower’. Large numbers of users express their sentiments over an issue, about an organization or product. Therefore it is necessary to analyze these views for better gauging the public perception.

In this paper, we will apply the Naive Bayes Classification to classify and analyze twitter tweets about Saudi Electronic University, which have been tweeted in *الجامعة السعودية الالكترونية* #. Since this hash tag is used by all formal twitter accounts of the Saudi Electronic University (SEU), and it has a lot of communications either between SEU accounts and the other twitter users, or between the interested twitter users about the SEU at different subjects. These tweets provide a good understanding of what the public thinks and wants from and about the SEU. Microsoft Azure platform is used for storing the tweets before and after classification in the cloud by using Blob storage. To visualize the data, Microsoft Power Bi is connected with the Blob storage which contains the classified tweets.

The rest of the paper is classified as: Section 2 includes literature studies till now. Section 3 describes the methodology of the proposed system. Section 4 discusses the result and analysis of the proposed study using the visual analysis. Section 5 includes the conclusion of the proposed work.

2. Literature Review

The past two decades have seen an exponential growth of influential internet based social media platforms such as Facebook and Twitter. As internet penetration continues across the globe, multitudes of people are using these platforms to generate, disseminate, and consume information in large volumes. Liu et al.[2] identify that social media contains big data with the potential to shape

human lives in profound ways. Social media has the potential to generate highly rich information both in terms of quantity and quality. Although it is arguable that there is still a significant percentage of social media data that does not meet the standards levels of quality, the use of improved extraction methods makes it possible to sift through the large volumes of data and derive information that meets a wide range of applications including social, economic, and political. It is proved in [3] that a knowledge management framework is effective at leveraging big social media data and derives valuable knowledge for business decision making. Corporate institutions, political, and non-government organizations are taking advantage of the immense potential of social media. In [4], it is analyzed that billions of active users, automation bots, artificial intelligence, and analytics collectors continuously scanned through the sites and gather immense volumes of data that is used to guide marketing schemes and enhance user experience. The findings are similar with those of [5], who argue that when big data and data analytics are used effectively, they can offer organizations with real-time actionable information that is critical to identifying problems, needs, and offer feedback on the effectiveness of policy implementation. Queiroz et al.[6] proposed a novel framework which is based on twitter data to provide efficient decision making for sustainable environment.

In recent time, textual document production increased rapidly in social media. For instance, up to march-2021, Twitter's users generating 500 million tweets per day in different languages [7]. In these tweets or messages, people all around the world share their opinion and feeling. The technical study of these messages/tweets is known as sentiment analysis [8]. Ruz et al. [9] proposed a novel methodology of Bayesian classifiers for sentiment analysis in critical events using two Spanish datasets. In [10], Arabic tweets are used for sentiment analysis using big data approach (SAP HANA) and lexicon-based classifier. A study by Elhadad [11], proposed a mechanism for sentiment classification of Arabic and English tweets. The scheme includes vector space model and TFIDF algorithm for sentimental text representation and classification. Alsudias et al.[12] suggested a methodology for Arabic tweets classification which is composed of three machine learning approaches i.e Naïve Bayes, SVM classifier and Logistic Regression. It is witnessed in [12] that machine learning algorithms achieved accuracy up to 84%. In [13], a hybrid approach is adopted to analyze Arabic tweets by selecting best features for improving performance of SVM classifier. Alqurashi et al.[14], proposed a HITS[15] and PageRank[16] based scheme in order to identify information super reader using Arabic tweets. A comprehensive study by Ghallab [17] critically analyzed about 108 articles for Arabic sentiment analysis. Since the morphological complexity of the Arabic language make it less explored contents as compared to English language.

Still researchers trying their level best to achieve high accuracy using complex structure algorithms. It is evident that machine learning algorithms especially deep learning algorithms accomplish complex tasks. Mohammad et al. [18] investigated three different deep learning architectures i.e. CNN, LSTM and RCNN for Arabic sentiment analysis. It is witnessed that LSTM outperformed the rest of the DL algorithms by achieving 81.31% accuracy.

A comprehensive research study is conducted by Nassif [19] to explore the strength of DL algorithms to analyze Arabic Subjective Sentiment Analysis (ASSA). It is observed that CNN, RNN and LSTM neural networks were the most used algorithms for ASSA. Ombabi et al.[20] proposed a hybrid approach of CNN-LSTM and SVM for Arabic sentiment analysis. The best features from the tweets are learned using CNN and LSTM algorithms. The features map is passed to the linear classifier SVM for sentiment classification. Unstructured Arabic text is growing rapidly in social media and analyzing such data for sentiment analysis become even more difficult if the data is very limited for DL algorithms. In order to improve the performance of Neural Networks, data augmentation techniques are used to achieve up to mark accuracy [21]. Oussous et al. [22] proposed CNN-LSTM based deep learning architecture for Arabic sentiment analysis. It is demonstrated in [22] that the performance of deep learning algorithms can be improved using efficient preprocessing Arabic tweets. The strength of Neural Networks specifically Recurrent Neural Network (RNN) and Long-Short-Term-Memory (LSTM) is investigated in [23] for Arabic sentiment analysis using Arabic Tunisian dialect.

The preprocessing of text is a cumbersome task for improving the performance of deep learning algorithms. It can be avoided if the refined datasets are fed to the neural networks. Elnagar et al.[24] introduced single and multi-label preprocessed datasets for Arabic text analysis task. It is witnessed in [24] that neural networks can achieve high accuracy up to 96% using preprocessed text. Kwaik et al.[25] suggested bi-directional LSTM with CNN deep learning architecture for polarity prediction of the text. The proposed DL architecture is trained using three Arabic Sentiment corpora the LABR [26], ASTD [27] and Shami-Senti [28]. Bdeir et al.[29] proposed a multi-label classification hybrid frame work for Arabic sentiment analysis. The hybrid scheme of word embedding and DL neural algorithms including CNN and RNN achieved accuracy about 90% with hamming loss 0.02.

Arabic language sentiment analysis is full of challenges due to the complex linguistic syntactic structure of the language. A study by Alsayat [30] explored all the challenges and how to deal with. The challenges of the ASA include phonology, lexicon study, morphology and semantic knowledge.

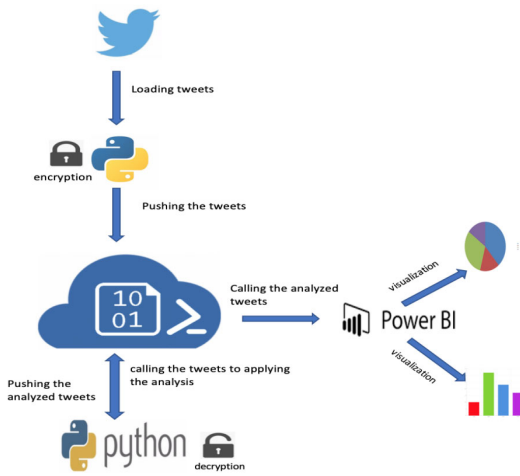


Fig. 1 Block diagram of the proposed system

3. Proposed Machine Learning based Methodology

Twitter data is ideal because it contains spatial and temporal information and can be easily used. Figure 1 shows the block diagram of the proposed scheme.

An API is used to access the Twitter platform. While supporting a large number of functions for interacting with Twitter, the API functions most relevant for acquiring a Twitter dataset include: Retrieving tweets from a user timeline (i.e., the list of tweets posted by an account) and Searching tweets also filtering real-time tweets (i.e. the tweets as they are passing through the Twitter platform upon posting). There are two APIs that can be used to collect tweets. First for doing one-time collection of tweets, REST API is used. For doing a continuous collection of tweets for a specific time period, then use the streaming API. After collecting the tweets, the collected data is stored in the cloud using Azure platform. Azure Cloud computing platform tends to be less expensive and more secure, reliable, and flexible than detected servers.

Azure Blob storage is Microsoft's object storage solution for the cloud. Blob storage is optimized for storing massive amounts of unstructured data. Unstructured data is data that doesn't adhere to a particular data model or definition, such as text or binary data. Some of the advantages of the blob storage include data consistency, object mutability and supporting multiple blob types.

For analyzing and classifying the collected tweets we decided to use the naive Bayes algorithm, which is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks. Also, it is easy to implement, requires a small amount of data to train the model and efficient results are obtained in most of the cases. The Bayes' Theorem is defined by the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

The components of the above statement are:

- $P(A|B)$: Probability (conditional probability) of occurrence of event A given the event B is true
- $P(A)$ and $P(B)$: Probabilities of the occurrence of event A and B respectively
- $P(B|A)$: Probability of the occurrence of event B given the event A is true.

Table 1. Percentage of classified tweets

Classification	Percentage %
1 (positive)	40 %
-1 (Negative)	30 %
0 (Natural)	30 %

4. Research Results and Discussion

The tweets are loaded from Arabic *الجامعة السعودية الالكترونية* #, since it is containing most of the tweets about Saudi Electronic University (SEU). Based on the tweets which downloaded, saved on the azure-Blob storage, are classified as Positive tweets (have labeled as 1), Negative tweets (have labeled as -1) and Natural tweets (have labeled as 0). Table 1 and figure 2 show the percentage of the classified tweets used in the proposed system.

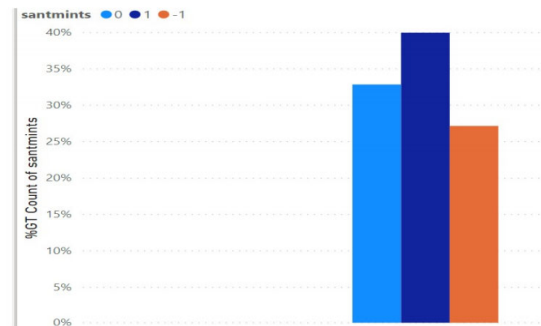


Fig. 2 Percentage of classified tweets

The locations of the tweets which have been generated by the owners of twitter accounts, and which we were able to load it, were from Saudi Arabia, Tanzania and Egypt. The tweets from Egypt and Tanzania were advertisements tweets. The other tweets were from different cities in Saudi Arabia or from Saudi Arabia without mentioning the city. Table2 depicts the locations from where the tweets are generated.

Table 2. Percentage of tweets generated from different locations

Location	Percentage %
Al Baha, Saudi Arabia	2.5%
Jeddah, Saudi Arabia	20%
Kingdome of Saudi Arabia	42.5%
Riyadh, Saudi Arabia	30%
Egypt	2.5%
Tanzania	2.5%

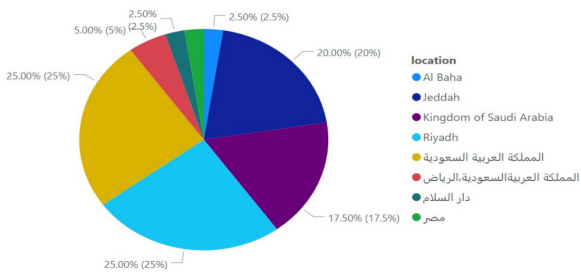


Fig. 3. All Tweets Locations

Figure 3 presents a pie chart depicting the total number of tweets by location. As can be seen from the figure, the highest number of tweets by location is from Egypt and the rest of Saudi Arabia excluding the mentioned cities. Both of these share a percentage of 25%. In order to make a useful analysis of the proposed system, it is highly judicious to classify between the types of tweets. Such a classification will not only enhance the effectiveness of the proposed work, but also make it feasible to be adopted for the real world applications.

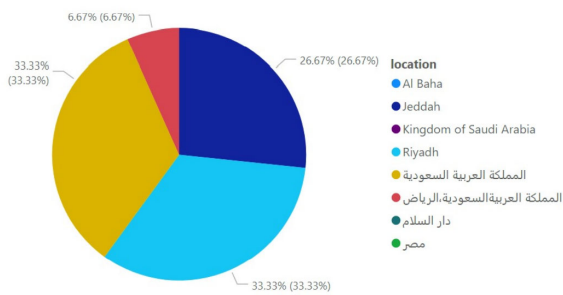


Fig. 4 Positive Tweets Locations

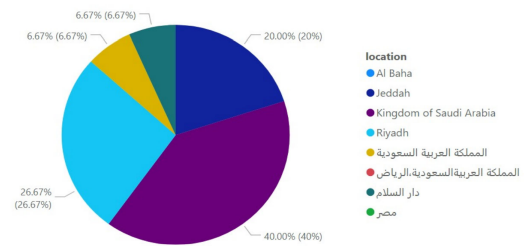


Fig. 5 Negative Tweets Locations

Figure 4 presents the division of positive tweets, whereas figure 5 displays a pie chart of the negative tweets. It can be observed from figure 4 that the highest number of positive tweets are from the Riyadh district followed by other cities in the rest of Saudi Arabia. The second major city with the highest number of positive tweets is Al Baha. The country with the highest number of positive tweets outside Saudi Arabia is Dar es Salaam in Tanzania. When it comes to the negative tweets we can observe from figure 5 that the highest number of negative tweets are from the other small cities from the rest of Saudi Arabia excluding the major cities. Both the positive and the negative tweets are of prime importance when it comes to analyzing the public opinion or, the social opinion in general.

In addition to the positive and the negative tweets, there is a valuable importance of analyzing the neutral tweets as well. These tweets are presented in figure 6. It can be observed again that the highest percentage of the neutral tweets are from the rest of the Saudi Arabia with a total percentage reaching to 40%. All the major cities including Riyadh, Al Baha, and Jeddah share the second highest place with an individual percentage of 10% each. Egypt and Tanzania also share the second place in neutral tweets with a percentage of 10% each. Neutrality in an opinion is considered as a major contributory force in social opinion because with proper steps take, it can be converted in to either the positive side or the negative side of the opinion.

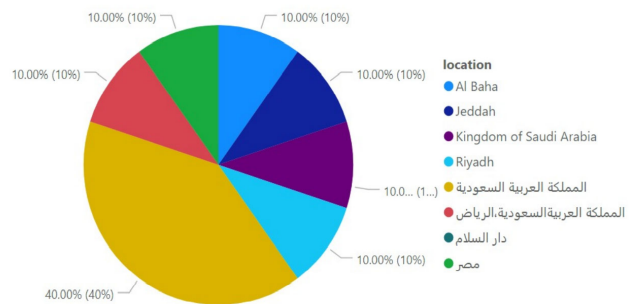


Fig. 6 Natural Tweets Locations

5. Conclusion

Social media platforms are one of the most durable sources of information and feedback around the world. The Twitter web site is one of the most platforms used in Saudi Arabia and because of that we used it to classify and analyze the tweets which have been written about the Saudi Electronic University (SEU). For storing the data we used Microsoft Azure Blob storage which is a cloud storage solution for storing the unstructured data. Also, we used the Naïve Bayes Classifier which is a classification technique based on Bayes' theorem with an assumption of independence between predictors to classify the data to Positive (labeled as 1) which was 40%, Negative (labeled as -1) which was 30% and natural (labeled as 0) which was 30%. then we did the visualization of the data by connecting the Blob storage to Power Bi and we realized that most of the tweets are from Saudi Arabia since the SEU is a Saudi University, also there are few of the tweets were from outside of the Kingdom of Saudi Arabia, such as Egypt and Tanzania.

Acknowledgments


The authors highly acknowledge the support provided by Saudi Electronic University, Saudi Arabia for this research.

References

- [1] Accessed on 20-March-2021 [Online] Available <https://www.statista.com/statistics/303681/twitter-users-worldwide/>
- [2] Liu, Qian, Jingsi Ni, Jing Huang, and Xiaochuan Shi. "Big data for social media evaluation: a case of WeChat platform rankings in China." In 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), pp. 528-533. IEEE, 2017.
- [3] He, Wu, Feng-Kwei Wang, and Vasudeva Akula. "Managing extracted knowledge from big social media data for business decision making." *Journal of Knowledge Management* (2017).
- [4] Manca, Stefania, Luca Caviglione, and Juliana Raffaghelli. "Big data for social media learning analytics: potentials and challenges." *Journal of e-Learning and Knowledge Society* 12, no. 2 (2016).
- [5] Hasnat, Baban. "Big data: An institutional perspective on opportunities and challenges." *Journal of Economic Issues* 52, no. 2 (2018): pp.580-588.
- [6] Queiroz, Maciel M. "A framework based on Twitter and big data analytics to enhance sustainability performance." *EnvironmentalQualityManagement* 28, no. 1 (2018): pp. 95-100.
- [7] Accessed on 20-March-2021 [Online] Available. <https://www.internetlivestats.com/twitter-statistics/>
- [8] C. Diamantini, A. Mircoli, D. Potena, E. Storti, Social information discovery enhanced by sentiment analysis techniques, *Future Generation Computer System* 95 (2019): pp.816-828.
- [9] Ruz, Gonzalo A., Pablo A. Henríquez, and Aldo Mascareño. "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers." *Future Generation Computer Systems* 106 (2020): pp. 92-104.
- [10] Alomari, Ebtesam, Rashid Mehmood, and Iyad Katib. "Sentiment analysis of Arabic tweets for road traffic congestion and event detection." In *Smart Infrastructure and Applications*, pp. 37-54. Springer, Cham, 2020.
- [11] Elhadad, Mohamed K., Kin Fun Li, and Fayez Gebali. "Sentiment analysis of Arabic and English tweets." In *Workshops of the International Conference on Advanced Information Networking and Applications*, pp. 334-348. Springer, Cham, 2019.
- [12] Alsudias, Lama, and Paul Rayson. "COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?." In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. 2020.
- [13] Alassaf, Manar, and Ali Mustafa Qamar. "Aspect-Based Sentiment Analysis of Arabic Tweets in the Education Sector Using a Hybrid Feature Selection Method." In *2020 14th International Conference on Innovations in Information Technology (IIT)*, pp. 178-185. IEEE, 2020.
- [14] Alqurashi, Sarah, Abdulaziz Alashaikh, and Eisa Alanazi. "Identifying Information Superspreaders of COVID-19 from Arabic Tweets." *Preprints* (2020).
- [15] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.
- [16] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." *Stanford InfoLab, Tech. Rep.*, 1999.
- [17] Ghallab, Abdullatif, Abdulqader Mohsen, and Yousef Ali. "Arabic sentiment analysis: A systematic literature review." *Applied Computational Intelligence and Soft Computing* 2020 (2020).
- [18] Mohammed, Ammar, and Rania Kora. "Deep learning approaches for Arabic sentiment analysis." *Social Network Analysis and Mining* 9, no. 1 (2019): pp. 1-12.
- [19] Nassif, Ali Bou, Ashraf Elnagar, Ismail Shahin, and Safaa Henno. "Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities." *Applied Soft Computing* (2020): pp.106836.
- [20] Ombabi, Abubakr H., Wael Ouarda, and Adel M. Alimi. "Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks." *Social Network Analysis and Mining* 10, no. 1 (2020):pp. 1-13.
- [21] Beseiso, Majdi, and Haytham Elmousalami. "Subword attentive model for Arabic sentiment analysis: A deep learning approach." *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, no. 2 (2020): pp. 1-17.
- [22] Oussous, Ahmed, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. "ASA: A framework for Arabic sentiment analysis." *Journal of Information Science* 46, no. 4 (2020): pp. 544-559.
- [23] Jerbi, Mohamed Amine, Hadhemi Achour, and Emna Souissi. "Sentiment analysis of code-switched tunisian dialect: Exploring RNN-based techniques." In *International Conference on Arabic Language Processing*, Springer, Cham, 2019. pp. 122-131.

- [24] Elnagar, Ashraf, Ridhwan Al-Debsi, and Omar Einea. "Arabic text classification using deep learning models." *Information Processing & Management* 57, no. 1 (2020): pp. 102-121.
- [25] Kwaik, Kathrein Abu, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. "LSTM-CNN deep learning model for sentiment analysis of dialectal Arabic." In *International Conference on Arabic Language Processing*. Springer, Cham, 2019. pp. 108-121
- [26] Mohamed Aly and Amir Atiya. Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, 2013 pages 494–498
- [27] Mahmoud Nabil, Mohamed Aly, and Amir Atiya. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015 pages 2515–2519
- [28] Chatrine Qwaider, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. Shami: A corpus of levantine arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*
- [29] Bdeir, Abdullah M., and Farid Ibrahim. "A Framework for Arabic Tweets Multi-label Classification Using Word Embedding and Neural Networks Algorithms." In *Proceedings of the 2020 2nd International Conference on Big Data Engineering*. pp. 105-112. 2020.
- [30] Alsayat, Ahmed, and Nough Elmitwally. "A comprehensive study for Arabic sentiment analysis (Challenges and Applications)." *Egyptian Informatics Journal* 21, no. 1 (2020): pp. 7-12.



Hazzaa N. Alshareef  holds a Ph.D in Computer Science - Mobile and Internet Systems from University College Cork, Ireland in 2016. He also received his M.Sc. in Software & Systems for Mobile Networks from University College Cork, Ireland in 2011 and his B.Sc. in Computer and Science and Engineering from Taibah University, Saudi Arabia in 2007. He is currently an assistant professor at the College of Computing and Informatics at the Saudi Electronic University, in Saudi Arabia. His research areas of interest include Cloud &

Mobile Computing; Wired & Wireless Networks; Vehicle Network & Unmanned Aerial Vehicle (UAV); Data Mining & NL; System Analysis & Design; IoT; Cybersecurity; and Artificial Intelligence. He has published many papers in international journals and conferences. He can be contacted at email: h.alshareef@seu.edu.sa.



Imran Usman received his BE degree in Software Engineering from Foundation University, Pakistan in 2003 and MS Computer System Engineering from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan in 2006. He joined Pakistan Institute of Engineering and Applied Sciences as a research scholar and

received his PhD degree in 2010. From 2009 to 2010 he served at Iqra University Islamabad, Pakistan as an Assistant Professor in Department of Computing and Technology. From 2010 to 2012 he served as Assistant Professor and Senior In-charge Graduate Program in the Department of Electrical Engineering at COMSATS Institute of Information Technology Islamabad, Pakistan. He is presently serving as Associate Professor in College of Computing and Informatics, Saudi Electronic University, Kingdom of Saudi Arabia. His present research interests include machine learning, digital image processing, evolutionary computation and digital watermarking. Dr. Usman has a number of research papers to his credit and has supervised many BS, MS and PhD students. He is also a Senior Member of IEEE.