The Impact of using Social Network on Academic Performance by using Contextual and Localized Data Analysis of Facebook Groups

Muhammad Aqeel[1], Mukarram Pasha[2], Muhammad Saeed[3], Muhammad Kamran Nishat[4]
Maryam Feroz[5], Farhan Ahmed Siddiqui[6], Nasir Touheed[7]
Department of Computer Science/UBIT, University of Karachi, Karachi, Pakistan[1, 2, 3, 5, 6]
Department of Computer Science, Karachi Institute of Economics and Technology, Karachi, Pakistan[4]
Department of Computer Science, Institute of Business Administration, Karachi, Pakistan[7]

***Abstract*—Social Networks due to their intrinsic nature of being addictive have become an integral part of our civilization and plays an important role in our daily interactions. Facebook being the largest global online network, is used as a primary platform for carrying out our study and hypothesis testing. We built a web crawler for data extraction and used that data for our analysis. Primary goal of this study is to identify patterns among members of a Facebook group using a contextual and localised approach. We also intend to validate some hypotheses using a data driven approach like comparison of student's social participation and activeness with actual class participation and its impact on his/her grades. We have also used user interactions in Facebook groups for identifying close relationships. The polarity of content in a group's comments and posts defines a lot about that group and is also conferred in this paper.**

Keywords*—Social networks; data analysis; data mining; NLP; sentiment analysis

I. Introduction

Social network like Social networking sites like Facebook and Twitter have become ubiquitous in our everyday life. With more than 2 Billion active users Facebook is undoubtedly the most used social networking site. Young students are in majority among these users. Students use it as a platform to share their ideas and collaborate on projects in addition to socializing with other students and people. There have been many studies to understand the impact of social media on the grades of undergrad students [5, 8, 9, 19, 20]. Results of these studies are mixed there is evidence of both positive and negative impact of social media usage on the academic performance of students. For our analysis, we built a web crawler for extracting data from Facebook and used its data throughout our analysis. The data extraction process and building the appropriate schema itself was a challenging task and is also presented in this paper. We extracted profiles of students in an undergraduate class studying computer science in Pakistan. First, we anonymize their data to protect their privacy. Then we study the correlation between their activeness on Facebook and their grades. We applied different data analysis techniques and some well-established correlation test to quantify the impact of social media.

The structure of the paper is as follows. Section 2 explores the related work in this area in detail. Section 3 describes our methodology and our dataset used in this study. Section 4 discusses the results of our study and Section 5 concludes the paper.

II. Related Work

With the advent of online social networking websites, there has been an explosion of user data and information and thus a lot of work has been done in the field of social network analysis. Previous studies involving Facebook data can be subcategorized into following three categories.

A. Social Network Analysis

Social network analysis is the study of analysing social structures using networks and graphs. Networks are built using existing user data available on online social networking sites like Facebook, Twitter etc. Graph data structures are extensively used to model such scenarios. Users are represented in the form of nodes of a graph and their relationships are represented by edges. As Berry Wellman [1] discussed in his paper in 2001, "Computer networks are inherently social networks, linking people, organizations, and knowledge".

Social network analysis is categorized in two types:

1) Sociocentric (whole) network analysis: This involves sociological quantification of interaction among a group of people. It focuses on the identification of global structural patterns. Most SNA research in organizations in recent times concentrates on this sociometric approach.

2) Egocentric (personal) network analysis: This emerged from anthropology and psychology and involves quantification of interactions between an individual (called ego) and all other persons (called alters) related (directly or indirectly) to ego. It is not so difficult to collect data for such studies especially after the advent of online social networking websites. Some generalizations are made if there is some missing data of an ego.

Similar work is in Social Network Analysis.

Many studies have been carried out to understand and better model sociocentric and egocentric networks. J. Ugander et al. [2] in 2011 analysed the global structure of the Facebook

user network graph and characterized the assortative patterns present in the graph by studying the basic demographic and network properties of users. Similar study was carried out by Traud et al. [3] in which they examined the roles of user attributes in a Facebook network of one hundred American colleges and universities.

Many of the work have been done to find relation between social media usage and its effect on academic performance of students Daniel Z. Grunspan et. al. [11] analysed two study networks from a single classroom and study interactions among students at exam time. Nikolaos K Tselios et al. [n1] examines the use of social media site Facebook and its correlation with academic accomplishments. Justyna P. Zwolak et al.[10] used Social Network Analysis to identify patterns of interaction that contribute to taking the advanced course of MI after taking a first course, in this paper they try to prove that social integration of students during course study increase their persistence towards that course. G. Han et al. [12] performed analysis on social network's data to measure student's collaboration in a capstone project, their study suggests the instructor should include structures activities that emphasizes student collaboration to help develop strong information networks in other courses. Eric Brewe et. al. [13] applied sequential multiple regression modelling to evaluate factors which contribute to participation in the learning community PLC (Physics Learning Center - that support the development of academic and social integration), their model indicates that gender and ethnicity were not good predictors of participation in social learning group.

B. Natural Language Processing

Natural language processing can be referred to as a range of computational techniques that attempts to understand, analyse and perform linguistic analysis on naturally occurring text to achieve human like language processing, which is helpful in many tasks. Sentiment Analysis (also known as opinion mining or emotion AI) is an NLP technique used to analyse opinions, sentiments, evaluations, attitudes and emotions from the text. With the growth of social media, individuals and organizations are now extensively using public opinions for making decisions based on sentiment analysis.

Natural language processing can be applied on text, audio, video and voice with some pre-processing. Basically, all formats are usually first converted to text and then the relevant NLP techniques are applied. NLP based on text can lead to discussion analysis, opinion mining, contextual study, dictionary building or corpus building, linguistic study, semantics, ontological study etc.

Machine translation is one of the most studied topics in NLP and significant advances have been made in this field. Machines are now capable of translating almost any human language efficiently and in real time.

Similar work is in NLP.

Substantial work has been done on Facebook's dataset involving NLP techniques. P. Dewan et al. [3] proposed and extensive feature set based on entity profile, textual content, metadata, and URL features to identify malicious content on Facebook in real time. In other work, F Krebs et al. [4] proposed methods for predicting reactions/emotions on user posts on pages of supermarket chains. Their final model predicted the distribution of reactions with an MSE of 0.135. Bharat Gaind [17] et al. also did similar work to detect emotions from the Facebook and Twitter, they applied two-fold approaches comprised of NLP and Machine Learning which yields significantly good results on test data.

SNA and NLP applications are being built using social networks' data in almost every field Adil Rajput. 2019. [14] gives a good overview of sentiment analysis from social network's data of patients and application of NLP to mental health. Hiroki Takikawa [15] conducted both large-scale social network analysis and natural language processing on Japan's political Twitter data, they applied community detection method to identify the five most common communities they also use topic modelling technique their results also showed if topic is solely propagated by left or right wing. Glen Coppersmith et al. [16] used natural language processing and machine learning techniques to detect quantifiable signals around suicide attempts.

C. Data Analysis

There are endless possibilities for data analysis using Facebook's data and people have used it to extract many insights and trends.

Data analysis can be classified into various types

1) Descriptive analysis: Descriptive analysis answers the "what happened" by summarizing past data usually in the form of dashboards. The biggest use of descriptive analysis in business is to track Key Performance Indicators (KPI's). KPI's describe how a business is performing based on chosen benchmarks.

2) Diagnostic analysis: After asking the main question of "what happened" you may then want to dive deeper and ask why did it happen? This is where diagnostic analysis comes in. A critical aspect of diagnostic analysis is creating detailed information. When new problems arise, it is possible you have already collected certain data pertaining to the issue.

3) Predictive analysis: Predictive analysis attempts to answer the question "what is likely to happen". This type of analytics utilizes previous data to make predictions about future outcomes. Business applications of predictive analysis include risk assessment, sales forecasting, etc.

4) Prescriptive analysis: Prescriptive analysis utilizes state of the art technology and data practices. It is a huge organizational commitment and companies must be sure that they are ready and willing to put forth the effort and resources.

Similar work in Data Analysis

Mariam Adedoyin-Olowe et al. [18] provided a survey of data mining and data analysis techniques for Social Media Analysis much of the work is done in this dimension R. Farahbakhsh et al. [6] crawled 479K random user profiles on Facebook and studied the sensitivity of various attributes and listed few which are rarely disclosed by users and are considered private. In another study, S. Mohammadi et al. [7]

explored the popularity evolution for professional users in Facebook. They monitored 8K most popular professional users on Facebook over a period of 14 months and concluded that being active and famous correlate positively with the popularity trend.

III. Methodology

This section includes detailed step by step explanation of our analysis.

A. Data Extraction

We choose to extract data from profile of users whom we can later verify our results with. Data extraction was done in three steps from specialization to generalization and following a localized approach.

1) Same class: Initially, we extracted data from a group of students belonging to the same batch. There were 85 members in that group and had 1207 posts, 4436 comments and 3177 replies.

2) Same major: Our next group had 532 members and all members had the same major. The group had 348 posts, 906 comments and 805 replies.

3) Same university: Lastly, a public group of university including members from all majors was scraped. There were 1593 members in that group and had 323 posts, 1355 comments and 881 replies.

For extracting data from these groups, we made use of python-client of web crawling library - Selenium. As Facebook renders only a small number of posts when a group is opened, we had to automate the scrolling of a page by executing some JavaScript code. We also had to expand all comments and replies in each post as those are collapsed by default. Once the page was completely rendered, we extracted the page source and parsed the HTML using Beautiful Soup library.

B. Schema

We created separate tables for posts, comments and replies and used their URLs as primary keys. We decided to follow this fragmented structure because it would help us in carrying out our analysis independently. Tables can also be merged later when needed.

We designed a separate schema for reactions, because all posts, comments and replies contain reaction_url. In addition, another table for each group was maintained which included information of each user's mentions to other users of that group.

C. Data Cleaning

Data extracted from Facebook was in raw form and had to be cleaned before carrying out any analysis. Cleaning of the data included:

- Text in most of the posts, comments and replies was in roman-Urdu and was converted to English.
- User profile links were URL encoded and had to be formatted correctly.
- An anonymous alias was assigned to each user profile to ensure user privacy.
- Removal of any trailing spaces, punctuations and links in the text so that it would later be used to find out subjectivity and polarity.

D. Polarity and Subjectivity

Polarity can be referred as the negativity or positivity of the text. Its value ranges from -1 to +1. Negative content has polarity less than 0 and positive content has polarity greater than 0, while polarity of 0 denotes the neutrality of the text.

Subjectivity expresses some personal feelings or belief. Content addressing general issue tends to be more objective than texts which contain information about a specific topic. Value of subjectivity lies between 0 and 1. Texts with values below 0.5 tends to be more objective and texts with values greater than 0.5 are interpreted as subjective.

After cleaning and preprocessing of the data, sentiment analysis was carried out on texts of posts, comments and replies to find subjectivity and polarity. For sentiment analysis python TextBlob library was used.

We can clearly see (Table I) that first example is positive and objective as it is addressing public while the second example is more positive than the first one and is also more subjective, because in the second example "Ali" is being addressed by his friend particularly and more positive sentiments are used. In the third example, subjectivity is quite high as it is talking about last class and Friday but is not polar as there are no sentiments involved.

E. User Participation in different Communities

We selected users who were members of all three of the groups we crawled and used their interactions in each group to identify participation patterns in communities of different sizes. We found a total of 14 members who were members of all three groups and counted number of posts, comments and replies of each user in each group.

It is visible (Fig. 1, Fig. 2, Fig. 3) that many of the students are the silent members of other groups except their class's group as they have lesser or no participation there. We can observe the trend as most of the members are very active in their closed class group, less active in their department's group and least active in public group but there is also an exception as one of our subject "Talha" has no post in class's group, but he is active in public group.

TABLE. I. Examples of Polarity and Subjectivity

Example	Sentence	Polarity	Subjectivity
1	Congratulations everyone, what a team work.	0.4	0.16
2	Congratulations Muhammad Ali on winning Trophy you did it once again brother <3	0.75	0.87
3	I forgot my calculator in last class on friday, did anyone find?	0.00	0.06

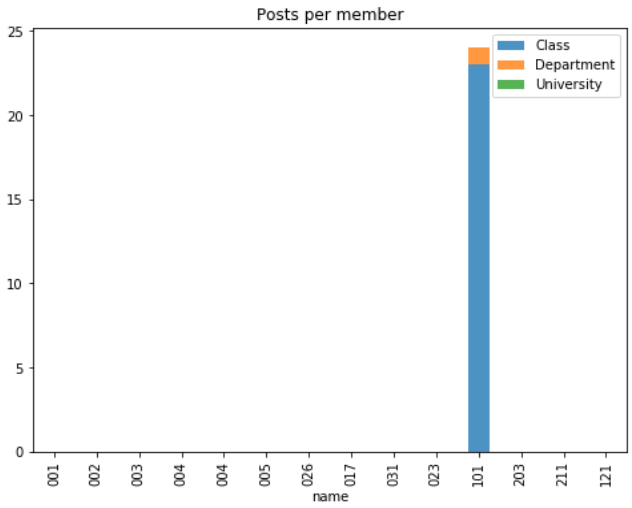

Fig. 1. Number of Posts in Each Group by Common Member.

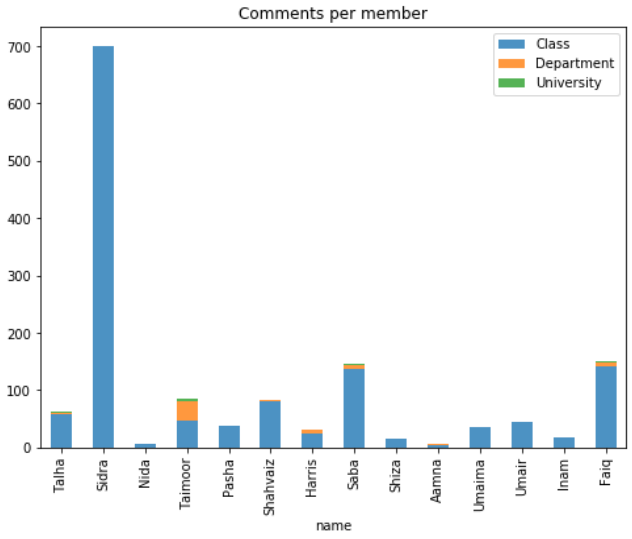

Fig. 2. Number of Comments in Each Group by Common Member.

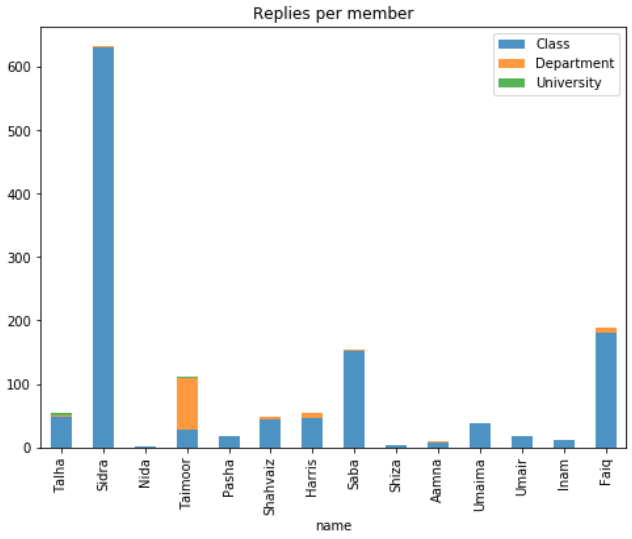

Fig. 3. Number of Replies in Each Group by Common Member.

F. User Interaction Frequencies

Using our existing data tables, we created a frequency matrix determining each user's interaction (mentions) with every other user (Fig. 4). The higher the frequency of user pairs, the stronger is the relationship. Building on these bi-pair user frequencies we can create groups of N members by setting a threshold frequency.

We can clearly see from mention frequency graph that very few users mention every user in the group otherwise most of the users are mentioned by the group of their groups or close friends.

G. Correlation between Grades and Activeness on Social Study Group

As we already have access to user's activity on Facebook class study group; their posts, comments and replies are all available in our dataset, we can make use of this and analyse whether there is any sort of correlation between social network usage and students' grades.

1) Hypothesis: Our hypothesis was that generally students who are active on social study groups should have good grades compared to those who do not participate in discussions happening over the study group.

2) Analysis: We aggregated the student's interaction on the study group by summing their posts, comments and replies. A data frame was built containing all study group members along with their CGPA and total interaction.

We calculated the correlation of CGPA and student's activity on social group and found it to be *0.0065*. Which meant there wasn't any impact of social study group activeness and student's grades. We then decided to dig into this problem a little deeper and segment the data based on some metrics to identify the correlation within the sub groups.

3) Segmenting Students Based on Their CGPA: We decided to create sub groups from the existing pool of study group members based on their CGPA. We partitioned into three separate bins.

a) Good Students - Students with CGPA greater than or equals to 3.

b) Average Students - Students with CGPA greater than 2 and less than 3.

c) Bad Students - Students with CGPA less than or equals to 2.

4) Correlation within Sub Groups: We then decided to analyse the correlation of each sub group separately. Average student's sub group showed no correlation but interesting patterns were identified within subgroups of Good and Bad students.

Correlation Heatmap of good and bad student's sub groups can be seen.

It can be seen clearly from the heatmap (Fig. 5) that there is a strong negative correlation of CGPA with every other social interaction feature. This indicates that those students who have good grades do not participate actively in study

group discussions. For good students, increased participation in an online study group can have a negative impact on their CGPA.

Correlation heatmap of bad students (Fig. 6) indicates an opposite scenario. Bad students who participate actively in online study groups, can have higher CGPA than those bad students who do not. CGPA for bad students is positively correlated, those who participate actively tends to get higher CGPA.

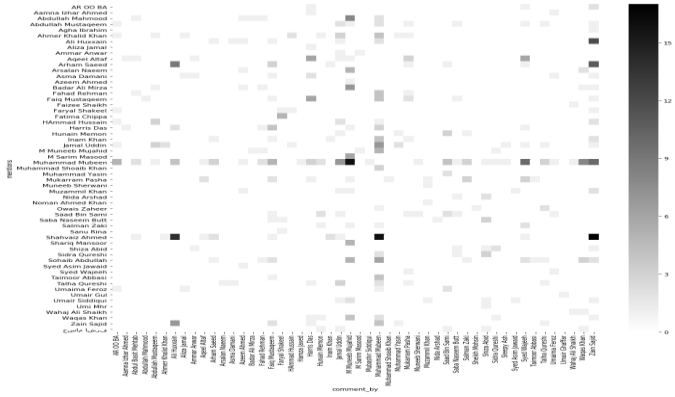

Fig. 4. Heatmap of user Interaction Frequencies.

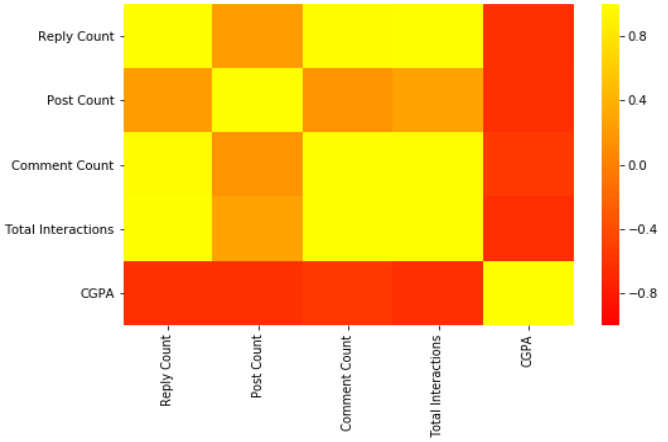

Fig. 5. Correlation Heatmap of Good Student's Sub Group.

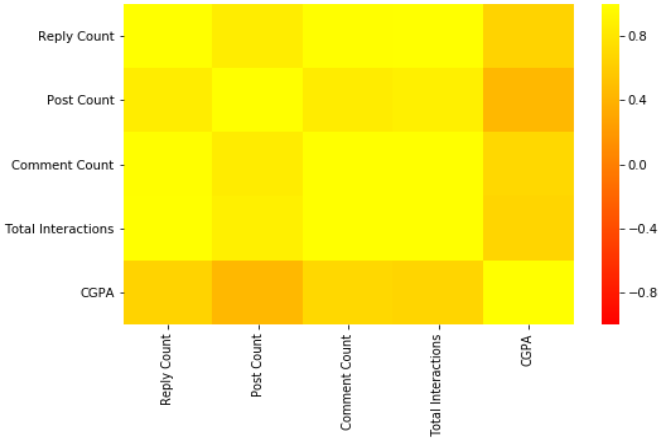

Fig. 6. Correlation Heatmap of Bad Student's Sub Group.

IV. Results

From user participation plots, it's evident that many users are comfortable interacting within their localized communities. This localized approach was helpful in carrying out analysis of student's social participation and was later compared with their actual class participation by their teacher's feedback.

This analysis came up with some good outcomes that students participating on all the forums actively are also vocal in their classes and the students who does not actively participate at all over social media but are active in class have very good grades.

Study of correlation between CGPA and student's interaction in study group generated interesting results as well. Students who are generally good does not need to participate actively in discussions instead the opposite is true for them. One the other hand, students who aren't good have better chances of getting higher CGPA if they participate actively in study groups.

V. Conclusions

Facebook is by far the most popular social networks among students in Pakistan. In this paper we study the relationship between social media usage and GPA score in a Pakistan.

In this study using real profiles and comments we come to a conclusion that mostly extensive use of Facebook on all the forums affects studies negatively in their academic life. On the other hand, actively participating in class group is positively correlated with good grades.

VI. Future Work

From existing data, we can dig out many other interesting insights and prove various hypotheses, few possible options are:

- Analysis on how e-learning using Facebook group is affecting the grades.
- Personality traits using NLP techniques.
- Detection of friend circles and closeness of users.
- We can use a weighted aggregation scheme for scoring student's interaction in study groups.

References

[1] Barry Wellman, "Computer Networks as social networks", 2001.

[2] J. Ugander, B. Karrer, L. Backstrom, C. Marlow, "The Anatomy of the Facebook Social Graph," Arxiv Nov. 2011.

[3] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social Structure of Facebook Networks," Feb. 2011.

[4] P. Dewan and P. Kumaraguru, "Detecting Malicious Content on Facebook," arXiv:1501.00802 [cs], Jan. 2015.

[5] F. Krebs, B. Lubascher, T. Moers, P. Schaap, and G. Spanakis, "Social Emotion Mining Techniques for Facebook Posts Reaction Prediction," Dec. 2017.

[6] R. Farahbakhsh, X. Han, A. Cuevas, and N. Crespi, "Analysis of publicly disclosed information in Facebook profiles," May 2017.

[7] S. Mohammadi, R. Farahbakhsh, and N. Crespi, "Popularity Evolution of Professional Users on Facebook," arXiv:1705.02156 [cs], May 2017.

[8] Mathews Nkhoma, Hiep Pham Cong, Bill Au, Tri Lam, Joan Richardson, Ross Smith, and Jamal El-Den ,"FaceBook as a tool for learning purpose ".

[9] Georgia Sapsani, Nikolaos Tselios , "Facebook use, personality characteristics and academic performance: A correlational study".

[10] Justyna P. Zwolak, Eric Brewer ,"The impact of social integration on student persistence in introductory Modeling Instruction courses".

[11] Daniel Z. Grunspan, Benjamin L. Wiggins, Steven M. Goodreau, "Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research" CBE life sciences education· January 2014.

[12] G. Han, OP McCubbins and T.H. Paulsen,"Using Social Network Analysis to Measure Student Collaboration in an Undergraduate Capstone Course". NACTA Journal, June 2016.

[13] Eric Brewe, Laird H. Kramer, Vashti Sawtelle, "Investigating Student Communities with Network Analysis of Interactions in a Physics Learning Center". Physical Review Special Topics - Physics Education Research, 2012.

[14] Adil Rajput, "Natural Language Processing, Sentiment Analysis and Clinical Analytics".

[15] Hiroki Takikawa, Kikuko Nagayoshi, "Political Polarization in Social Media: Analysis of the "Twitter Political Field" in Japan".

[16] Glen Coppersmith, Ryan Leary, Patrick Crutchley Patrick Crutchley, "Natural Language Processing of Social Media as Screening for Suicide Risk".

[17] Bharat Gaind, Varun Syal, Sneha Padgalwar, "Emotion Detection and Analysis on Social Media".

[18] Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, Frederic Stahl, "A Survey of Data Mining Techniques for Social Media Analysis".

[19] Tariq, W., Mehboob, M., Khan, M. A., & Ullah, F. (2012). The impact of social media and social networks on education and students of Pakistan. International Journal of Computer Science Issues, 9, 407–411.

[20] Al-Khalifa, H. S., & Garcia, R. A. (2013). The state of social media in Saudi Arabia's higher education. International Journal of Technology and Educational Marketing (IJTEM), 3, 65–76.