Harvesting tweets for a better understanding of Engineering Students’ First-Year Experiences

Dr. Sherif Abdelhamid, Virginia Polytechnic Institute and State University

Dr. Sherif E. Abdelhamid served as an Assistant Professor of Computer Science at the College of Computing and Information Technology (CCIT). Before joining CCIT in 2018, he was an Infrastructure Software Engineer at the Center for Open Science, Virginia, USA.

He obtained his Ph.D. and MSc in Computer Science from Virginia Tech and his current research work lies at the intersection of computation and education. In particular, he is interested in building and evaluating software systems that enrich the students’ learning environment and promote their success.

Mrs. Mona Aly, Arab Academy for Science, Technology and Maritime Transport

Mona obtained her MSc and BSc in Electronics and Communication, Arab Academy for Science, Technology and Maritime Transport. Her research interests include social networks, and algorithms, wireless and mobile communications, wireless sensor networks, and QoS provisioning in wireless networks.

Dr. Andrew Katz, Virginia Polytechnic Institute and State University

Andrew Katz is an assistant professor in the Department of Engineering Education at Virginia Tech.
Abstract

Twitter, a popular social networking and microblogging platform, harvests and stores large amounts of data about myriad topics through millions of short messages (tweets). Among this array of topics, some tweets can contain valuable information related to engineering education and first-year engineering experiences. Unfortunately, despite the existence of such related tweets, the engineering education community writ large typically does not have adequate background and statistics on their number and content in order to glean information from this corpus of tweets. In general, data from tweets can be very useful for both qualitative and quantitative studies focusing on first-year engineering experiences. By incorporating data collected from Twitter, we can have the opportunity to discover interesting patterns and themes. In this paper, we report on the results of a study in which we collected and analyzed tweets related to engineering education and first-year engineering experiences. Additionally, we present the implemented pipeline used in our study. The pipeline uses the Twitter application programming interface (API) to pull tweets that contain specific key terms related to our topic of interest and then extracts the tweet content along with other metadata before storing the information in a central online database. Researchers can have access to a web-based interface where they can use the harvested tweets in their studies and get the latest tweets and news feeds.

1 Introduction

1.1 Background and Motivation

Every day millions of tweets are sent all over the world, carrying large amounts of data on various topics. Some of these tweets are related to engineering education in general, and first-year engineering students specifically. These tweets can be created by students, universities, governments, policymakers, among others. Tweets may contain information about daily activities, important announcements, learning content or resources, discussions on a specific topic, locations, and much more. Additionally, tweets can show interactions between tweet creators and other users or accounts who retweet, comment on, or like the tweet. This type of information creates networks of interactions or collaboration that can reveal patterns or themes of interest. These tweets are already archived and maintained by Twitter and provide a valuable source of knowledge for engineering education researchers to understand different topics related to engineering education in general and first-year engineering experiences in particular. This motivated us to dig into the Twitter archive, harvest, and analyze tweets relevant to our topic of interest. In addition to this analysis, we made the retrieved tweet datasets publicly available to other researchers for further analyses. Our approach is a multi-method one that involves both quantitative and qualitative data analysis.

1.2 Contributions

Our major contributions include: 1) an automatic software pipeline to harvest relevant tweets that contain relevant keywords to engineering education and first-year experiences, 2) a descriptive
and content data analysis of the harvested tweets, and 3) implementation of a web-based chatbot that researchers and students can use to get access to the latest tweets, news feeds, and the harvested tweet dataset from our implemented pipeline.

2 Literature Review

Researchers from different fields have been using tweets to gain insight into their subject domains including marketing [1], healthcare [2][3], activism [4], cybersecurity [5], athletics [6], and natural disasters [7]. This trend is also present in engineering education. For example, in one study [8], researchers used Twitter to trace the participation and conversations about a campaign geared towards the promotion of STEM learning and engagement among the public. In another study [9], researchers used Twitter to support students’ design thinking and to help students express their ideas with greater focus and higher emotional content.

Our research work involves descriptive and content analysis. The analysis of tweets is close in concept to document analysis. According to Bowen [10] document analysis is a process for reviewing or evaluating documents both in printed and electronic format. Documents (tweets in our case) can include text and images that have been created without a researcher’s intervention. According to [11], document analysis results in passages that are then organized into themes, categories, or case examples.

3 Data Collection Pipeline

The tweet collection pipeline consists of three main phases: 1) harvesting, 2) archiving, and 3) analysis. The harvesting step is performed through the Twitter application programming interface (API). We developed a python script that connects to Twitter through the site’s API and searches for all the tweets that contain the keywords ‘first-year engineering’. The search covered the period from 2016 to 2019 and resulted in a sample of 10000 tweets. The goal of the search is to find tweets relevant to first-year engineering experience, however, it is likely to retrieve some tweets that are not relevant to the intended search query. Such tweets are called false positives.

4 Descriptive Data Analysis

Descriptive analysis is used as an exploratory analysis to get a better understanding of the data in the study and provides the measures and numbers that can help to reveal interesting patterns and themes. The collected tweets span over four years from 2016 until 2019 as shown in figure 1.a. A cap has been put, due to technical limitations with the used tools. The total number of unique users/accounts is 7420 with an average of 1.35 tweets per user. The most active account within the selected sample is “educlashco” with 154 total tweets. The “educlashco” account provides the latest news related to Mumbai University/I.T. with associated hashtags #mumbaiuniversity and #mu. This account joined Twitter in January 2016. The most active account based on the number of retweets and replies is “MicroSFF” with 6133 retweets, 85 replies, and 14662 favorites. “MicroSFF” account writes very short science fiction and fantasy stories on Twitter and the tweets are also cross-posted to Facebook and other social media platforms. The account’s use of the storytelling approach might be one reason for the high user and student engagement. The most mentioned account is “YouTube” with 115 mentions. The number of used URLs in the
tweets is 3834. The most used URL is for the University of Mumbai, the office for first-year engineering students. Most of the URLs we found point to social media posts/status (e.g. Twitter, Facebook), news websites (e.g. Wall Street Journal, CBC Canada, CBS), or university websites (e.g. Georgia Tech, Mumbai University), see figure 1.b. Some of the URLs are in shortened formats using domains such as .ly, .gl, .tt, among others. The total number of shortened URLs is 832, with most of them using the .gl domain. Correlation analysis of the number of favorites, replies, and retweets showed that the three variables are positively correlated. The number of favorites and retweets were strongly correlated, $r(998) = .51$, $p < .001$. All correlation coefficient values are shown in figure 1.c. Kendall’s tau statistic was used to estimate the rank-based measure of association.

Figure 1: (a) The distribution of tweets through the four years period, (b) The Percentage of social media platforms used by tweeters in the URLs, (c) A correlogram showing the correlations between the number of favorites, replies, and retweets.

5 Content Analysis

5.1 Codebook Analysis
Codebook analysis is employed to categorize and analyze the tweets. According to [12], this technique is defined as a qualitative and sense-making effort that can digest a large volume of data and identify key patterns and themes. One of the main reasons to use this qualitative analysis is because tweets may contain casual colloquial language, sarcasm, and misspellings. This type of content may lead to ambiguity in meaning and subjectivity in interpretations [13]. According to [14], during the analysis of social media content (e.g. tweets), erroneous assumptions are likely to occur if automatic algorithms (e.g. machine learning) are employed without incorporating a human inspection. The lens we used in the content analysis was to identify: 1) the first-year engineering students’ experiences with the intention to identify issues and challenges, 2) how students and faculty use Twitter to communicate, announce for events, and share learning resources. The tweets were sorted into the appropriate categories by two independent raters with an agreement of over 92% of the tweets, indicating a high level of reliability. When raters disagreed, a discussion was utilized leading to a 100% agreement at the end. Researchers read a random sample of 200 tweets and developed initial categories that were used to code another random sample of 200 tweets. Codebook analysis revealed eight main categories representing different topics and uses of Twitter by various users (students, educators, universities, among others). Most of the tweets were describing the first-year experience, not
only from students’ perspective but also from educators and teachers. Various emotions were captured in the tweets, see figure 2.

![Figure 2: (a) The distribution of codes/categories extracted from the tweets, (b) A detailed distribution of different emotions as expressed through the tweets.](image)

5.2 Sentiment Analysis

Part of our content analysis involved a sentiment of the tweets. According to [5][15], sentiment analysis can strongly reveal opinions, emotions, and feelings, and therefore can be very useful to understand first-year engineering experiences. MeaningCloud [16] has been used for text analytics purposes in the development/testing/validation of this part of the research. MeaningCloud provides support for short sentences, so it can be convenient for tweet analysis. The tweets’ text is analyzed to determine if it expresses a positive/negative/neutral sentiment; to do this, the local polarity of the different sentences in the text is identified and the relationship between them evaluated, resulting in a global polarity value for the whole text. Figure 3.a shows the breakdown of the sentiment across a sample of 1,000 randomly selected tweets. We set a cap due to technical limitations and we used this sample as a proof of concept.

5.3 Automatic Tweets Emotion Categorization

In this part of the analysis, we conducted an automatic emotion categorization using an in-depth rule-based method. We assigned one or more categories to the tweets, based on a combination of morphological, semantic, and text rules. We used the emotion recognition categorization model in MeaningCloud. As proof of concept, we randomly selected and analyzed 730 tweets. Out of this number, 365 tweets were identified with “no category”. The rest of the tweets were categorized into eight categories as shown in figure 3.b.

![Figure 3: (a) The distribution of the tweets’ polarity. Polarity ranges from P+ (very positive) to N+ (very negative). (b)The results of the automatic emotion categorization of the tweets.](image)
5.4 Network Analysis

Researchers may also be interested in studying networks as they reveal the structure of the social network of the users and the interactions between them. In network analysis, people are typically represented as nodes and connections as links or edges. Researchers are usually interested in identifying key network properties including hubs (important nodes), brokers (nodes connecting two or more groups together), isolated nodes, communities, and connected structures.

In this part of the analysis, we constructed a social network from the sampled tweets where each node represents a Twitter account (e.g. user, business, or organization), and each edge represents the relationship through the mentions. The number of nodes and edges is 3285 and 2601, respectively. The average number of connections of each node (mention interactions) in our network is 1.584. ‘YouTube’ account has the highest number of connections (degree = 92) which reflects the high number of mentions along with YouTube URLs used in the tweets. Following the YouTube account, we have a ‘BrianD_VIUEng’ account with a degree equal to 25. The nodes with the highest degree in our social network are the accounts that highly mention or get mentioned by most of the other accounts. These nodes are often referred to as hubs, and calculating degree is the quickest way of identifying hubs. Betweenness centrality is another network measure that is useful to capture broker nodes that stand between groups and give the network connectivity and cohesion. ‘ChiragVariawa’ account has the highest betweenness centrality. This account belongs to Prof. Chirag Variawa, Director of First-year Curriculum at the University of Toronto. The network density equals 0.00048 which is very low. Small dense networks can indicate the possibility of ‘exploratory group thinking’ development where the resemblance of ideas is highly valued and normatively enforced [17].

6 User Interface

Researchers and students can use our chatbot to get access to the latest tweets, news feeds, and the periodically harvested tweets datasets from the implemented pipeline. New datasets will be added periodically and can be used for further analyses by other researchers. Users can visit the page at fb.me/fyeechatbot and send messages to our chatbot at m.me/fyeechatbot

Conclusion and Future Work

In this paper, we presented a pipeline that involved collecting, archiving, and analyzing harvested tweets. Both quantitative and qualitative data analyses were used. Results have shown that Twitter can provide a valuable source of knowledge for engineering education researchers to understand different topics, themes, and issues related to engineering education and first-year engineering experiences specifically. Results also revealed how students and faculty use Twitter to communicate, announce for events, and share learning resources. In the end, we provided a web-based chatbot that allows access to the latest tweets, news feeds, and our harvested tweets datasets. Our future work will involve the utilization of more data collection techniques. We aim to increase the sample size to hundreds of thousands of tweets. We plan to start an initiative where Twitter users can use the same hashtag #fyeet when posting tweets related to their first-year engineering experience. This will allow the harvesting of more relevant tweets.
References


