Two Tier Analysis of Social Media Collaboration for Student Migration

Ronak Razavisousan and Karuna P. Joshi
Information Systems Department
University of Maryland Baltimore County (UMBC)
Baltimore, MD 21250
{Ronak2, Karuna.joshi@umbc.edu }

Abstract – **Global adoption of Social Media as the preferred medium for collaboration and information exchange is increasingly reshaping social realities and facilitating new research methodologies in various disciplines. Social Media applications are collecting a large amount of User-Generated Content (UGC) and web data that contains knowledge about novel approaches of global collaboration between people. We have done a detailed study of the factors that lead to student migration, as espoused by social scientists, and compared it with factors observed by analyzing over 10 million Twitter posts. Using the gravity model as our baseline, we built a novel methodology to identify the features and facts that twitter posts offer for studying human collaboration during migration. We leveraged methods from Natural Language Processing (NLP) to extract contents specific to migration from social media posts. We used topic modeling- Latent Dirichlet Allocation (LDA) to extract the topics from tweets and word embedding- Word to vector (W2V) to find the correlation and similarity between UGC and socioeconomics theories. In this paper, we present our methodology in detail, along with the results of our analysis.**

Keywords—Student migration, Student mobility, Educated immigrant, student flow, Gravity model

I. INTRODUCTION

Over half of the world population is currently online, and a large volume of data is getting generated by mobile users, social media applications, and sensors. Researchers increasingly recognize that these "digital traces" present an enormous opportunity to complement traditional sources of migration data and improve knowledge of various aspects of migration. [1]. Social media platforms capture the geo-location of their users and so provide a useful dataset to study mobility patterns and the collaboration approaches of migrants.

United Nations (UN) report [2] indicates that human mobility has increased in recent decades all around the world. Knowing more about other geographical points and traveling facilities accelerate this phenomenon. Accessing the internet provides more data for immigrants, and people share their knowledge, experiences, and concerns through the web.

Demographical changes and economic influences are the main reasons that immigration receives attention from governments [3]. Labor force growth, entrepreneurship by immigrants, and human capital are three critical areas where immigrants influence economic growth. Mixing the cultures and providing the sociological changes are evaluated as positive effects of accepting immigrants. Therefore, knowing the flow of immigrants and forecasting about human immigration is valuable for societies, governments, and all involved parts [4]. Gravity model [5] is the most prevalent model used for modeling the flow of immigration [6]. Applying this law to human behavior concepts, like migration, was proposed by Ravenstein [7], while Q. Stewart [6] pioneered the use of this law for studying human geographical movement.

The goal of this research is to determine the novel collaboration approaches employed by potential migrant students—people who move to other countries for education—while using social media applications, like Twitter and Facebook, for determining factors facilitating migration. We compare these factors with existing socioeconomic models on migration, like the gravity model. For this study, we collected and analyzed over 10 million tweets. Techniques like Natural Language Processing (NLP), Topic Modeling, and Semantic similarity, were used to build a novel methodology to analyze critical terms in the migration theory research and compare them with the topics discussed in the tweet posts related to student migration. In this paper, we present our approach along with the results of our analysis of the Twitter feed related to student migration. The main contribution of this study is leveraging the economic model of migration to analyze social media data about student mobility.

The remainder of this paper consists of five sections: section II lists the related works. Section III describes the methodology of our study, and section IV describes the experiment and the results. We conclude in Section V.

II. RELATED WORKS

Immigration is defined as an international movement of people into a destination country of which they are not natives or where they do not possess citizenship to settle or reside there [1]. Immigration has been an active topic of research by the social scientists who have studied the reason why people migrate, the effect on a country's resources due to this migration and the consequences of brain gain or drain on societies. UN report [2] considers three principal roles for immigration: host, origin, and immigrant, and we have examined all the three roles in our study. Host refers to locations or countries of immigrant destinations. Origin points to the countries from which the immigrants are coming. Immigrants are the people who change the location of leaving from origins to the hosts.

A. Gravity Immigration Model

Immigration, as a research topic, has been studied from different perspectives. Economic and sociological assessments are common approaches for analyzing the concept of immigration, and we use these as the ground truth for our research. There is a rich list of studies, and we select five of them in the qualitative part of our research [6,8-11]. These papers mainly focused on studying the students' motivation for migration or used a gravity model to describe human migration. Findlay *et al.* [8] have studied the motivation, situation, and adaptation of international students. Michel Beine *et al.* [10] developed a theoretical model to identify different motivations of student mobility from the origin and host point of view [10]. Having a rich list of motivating factors for student mobility helped us to get more accurate results.

The network of immigrants in the host country and the role of social media for supporting this network has been studied by Lee Komito [9]. Another research by Karemera *et al.* [11] examined the migration in North America considering the gravity model [11], and we have referenced it while comparing the factors of human migration based on the gravity model to the factors extracted from social media.

The gravity model is expanded for all types of human migration in research by Simini *et al.* [6]. It is inspired by Newton's law of Universal Gravitation that defines the gravitational force between two objects as

$$\boldsymbol{f = G\frac{m_1 \times m_2}{r^2}} \quad (1)$$

In equation 1, f is the gravitational force between two objects, *G* is the gravitational constant, m_1 and m_2 are masses of the two objects, and *r* is the distance between two objects.

A gravity model of migration, inspired by Newton's law estimates and predicts the forces which influence human immigration. It is evident that the human reactions and process of decision making affect the accuracy of the results, but the model and its results on human migration are reasonably acceptable [12-13]. When the Gravity model is applied to other sciences, it is not as static as the law in physics. For instance, the population of the origin and destination can be simulated as m_1 and m_2[14], in some cases, for using this law to model the trade between two areas. Elsewhere the gross national product (GDP) of the origin and destination replace the m_1 and m_2, and the geographical distance between these two points is selected as *r* [14]. We chose the gravity model as a baseline and wanted to determine if the social media data can validate it.

B. Social Media Analysis for Social behavior

The second group of related works refers to the researches who have used web data and social media to analyze the social factors of migration. Hughes et al. [15] presented a wide range of data sources to study immigration from traditional methods to modern ones. They also provide a list of data sources and their features. We used the feature list of data sources to choose the most appropriate one. This research focuses on comparing the differences between traditional and modern methods. In our research, we use the results of the traditional method as the ground truth to compare with the data from modern methods. McGregor et al. [16] have studied the feasibility of studying migration from social media. They studied the role of social media and its influences on the concept of human migration from four perspectives. They categorize and studied the role of social media in four groups including 1) Influence of social media on migration; 2) migration integration and the role of social media in this process; 3) immigrant networks and the role of social media; 4) the role of social media in studying migration [16]. The main distinction of this work from our research is that we explicitly focus on student migration. Moreover, the core concept of their research highlights the role of social media in the field of migration. In contrast, we focus on the migration model and use social media as a data source.

Researchers have also modeled human mobility. Hawelka *et al.* [17] leveraged the geolocation of tweets to extract the mobility pattern. They also considered the seasonal timing and community network in this pattern. Another example is the study conducted by Pinto *et al.* [18], which examined user behavioral activity and social media activity to justify migration between cities. They also consider the tweets' geolocations. Geo-located social media activity, such as posts on Twitter and LinkedIn, have been used to infer international migration flows [19]. The main difference between these previous approaches and ours is in the type of data we have analyzed. These studies have accessed the geolocation of users and identified the location changes in their approach, while we extract the pulling and pushing factors for movement based on the social media content shared by users.

Grover and Kar [20] extracted a practical solution for promotional marketing based on the social media activities of users. They built a twitter engagement model for understanding user dynamics. Hong and Davison [21] studied topic modeling for microblogging datasets such as twitter. McCallum *et al.* [22] proposed a model to identify groups and topics among the text. Mikolov *et al.* [23] worked on the approaches to find a similarity between words, and they proposed a model to compute a vector for each word. Maas *et al.* [24] built a model that uses a mix of unsupervised and supervised techniques to learn word vectors that capture semantic terms. Our novel approach differs from these studies, as it combines topic modeling, clustering, and word embedding to determine the closeness of topics from different data sources. The main contribution of our work is to study student mobility based on the gravity model of migration with the help of content analysis of UGC on Twitter.

III. METHODOLOGY

Migration is typically studied as a sociology and economic phenomenon. Thus, we referenced five socioeconomic articles in [6, 8-11] as ground truth for our study. We developed a hybrid framework for this exercise by combining qualitative and quantitative approaches.

A. Qualitative method

For qualitative analysis, we studied the gravity model proposed for student migration based on research in [6, 8-11], where the focus is either on student's migration or gravity model. The work by Bergers *et al.* [27] inspired us for this approach. The qualitative part gave us a list of factors related to student migration. Qualitative methods helped us to identify the

sentences and paragraphs which contain immigration factors. The research collection contained information about student's migration and the three roles of migrations (origins, hosts, and immigrants). Then the collected sentences were classified into different groups. Finally, we obtained a list of factors used in the mapping process discussed in detail in section III. *C.* We used manual text annotation and content analysis for this qualitative analysis.

B. Quantitative method

We collected posts from Twitter, focusing on user discussions about migration (see section IV A). Social media provides unstructured data, and we extracted the contents of the user-generated posts using existing NLP techniques. First, we used probabilistic topic modeling, Latent Dirichlet Allocation (LDA) [29] for extracting the standard topics in the users' posts. We also ran LDA on selected articles [22-24] from the social science field. We next used word embedding approaches (W2V) [24] to find the similarities between topics from the datasets.

C. Research Methodology

Our methodology, illustrated in Figure 1, consists of three main steps – data collection and preparation, clustering, and mapping, and proofing. The steps are described in detail below:

Step 1: Data collection and preparation:

In this step, the social media's framework and the method of collecting data were identified. Moreover, the socio-economic sources, which include models, rules, and articles about those rules, were selected. After identifying our source articles [6, 8-11], we manually extracted the sentences and paragraphs from these articles. Not all parts of the text in these articles were related to migration, such as research techniques and research processes. Pieces of the texts extracted from these articles contain the facts and information that will be related to the selected model and rules. The collected text was used in the mapping process.

Step 2: Two-level of clustering: In the second step, we performed two levels of clustering. In the first level, we have a content analysis by extracting the topics. While LDA is a known technique for topic modeling topics in large documents, LDA has limitations for microblogging posts due to its short length. Hence, we used the Twitter-LDA model [27]. At the end of the first level, we obtain a list of topics that are popular for each dataset. To identify the similarity and closeness of the extracted topics, the W2V-LDA [28] inspired us. By leveraging W2V, all the topics from social media and articles were converted to the vectors in a unique vector space to find topics with the highest similarity.

Step 3: Factors Extraction: In the third step, we produced a list of topics from social media with the highest similarity to the topics from articles. The extracted factors, in the first step, guided us to map each topic to the migration factors that socioeconomic articles listed. Then we determined which social media topics validated model factors.

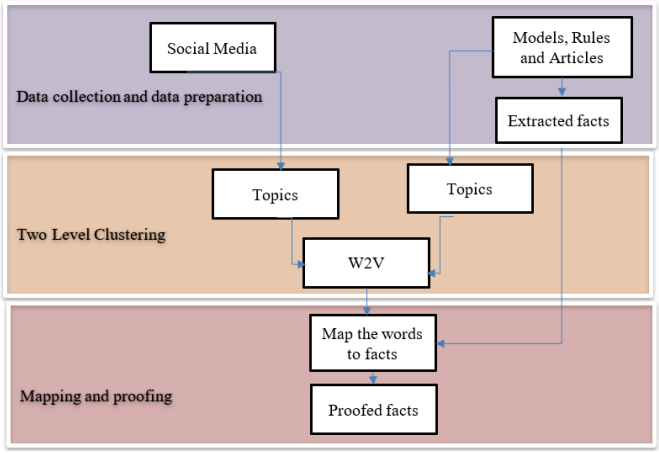

FIGURE I: THREE-LEVEL ARCHITECTURE OF THE RESEARCH METHODOLOGY

IV. EXPERIMENT

In this section, we describe our experiments based on the methodology.

A. Data collection

As we mentioned earlier, we choose Twitter as a source of social media data because of the accessibility, popularity, and number of active users. Moreover, Twitter has various ways to extract data to study users' mobility [18]. Twitter APIs used for collecting data. Sample data shows that people who talk about the migration in social media have a vast range of points of view. For narrowing the scope of the research, we limit the topics to student mobility and student migration.

To compile a comprehensive list of keywords to extract from Twitter, we conducted interviews with international students and staff in the International Student Office at the University of Maryland, Baltimore County. The main goal of the interviews was to determine the concerns, worries, questions, and opinions of international students pertaining to student migration.

The results of the interviews are categorized below:

1. Common questions topics:
   - Work permission (Permit) for post-graduation
   - Tend to stay or find a way to stay
   - CPT- Duration – extension- employer
   - Permanent resident (PR) – green card
   - OPT – STEM OPT – traveling
   - Carrier
   - Internships – coups
   - International faculty _ EB1 / EB2
   - Passive working – other sources of income
2. Visa related questions:
   - Embassy for an interview– tips for the interview- get visa in a first place
   - Visiting home – returning visa- traveling (vacation – conference)- documents- safety
3. Updating the immigration record:
   - Updating /Changing position / employer
4. Applicant questions:
   - University ranking
   - location: diversity in the area/ Job hob

- living cost
- safety

5. Countries that the process of getting a visa may take longer (7 countries in Muslim Ban and specific people of Venezuela).

The list of keywords identified from the interviews and the categories are presented in table I. The process of data collection from Twitter was based on this keyword list, and we collected all tweets that had any of the keywords. We could not confirm if the writers of the collected posts were actual students or not. However, this did not affect our analysis in determining the topics of the content and so will not influence the results of the research. Twitter API was used for this data collection, and the collection has done from Nov 2018 to Feb 2019 (four months). We collected almost 10 million tweets.

TABLE I: LIST OF KEYWORDS USED FOR DATA COLLECTION FROM TWITTER

Immigration	F2visa	GRE
Visa	Student migration	Student green card
Student visa	Foreign student	US university
Educated immigrant	International education	Student work permit
Graduate student immigrant	International student	OPT
F1visa	TOEFL	CPT

B. Data cleansing

One of the critical steps of data preparation was cleaning the data before the analysis process. For data cleansing, we performed these steps: 1) URL removal, 2) retweets removal, 3) stop words removal, 4) Stemming, 5) Tokenizing, and de-tokenizing.

Although the contents of URLs that people share for each topic have valuable information about the users' concern, in this work, we decided to remove all URLs and focus on the user-generated content. For the same reason, we removed the retweet posts because the duplication of the posts is not a focus of this study, and we focused only on the content of the tweets. The original tweets saved, but the retweeted posts removed.

Stop words removal and stemming are primary steps for text analysis to omit the words that do not add semantics and turning the words into its roots.

Tokenizing is a common task in text mining to break strings into words. It was also necessary for our cleaning process to prepare the input data for topic extracting using LDA (next step in C).

After running the LDA, there were some topics in our results for which de-tokenization was required. Some phrases or bonded words that should come together were separated due to the tokenizing process, and we rejoined them by de-tokenizing.

C. Data processing

Figure II illustrates our approach to process data from the two sources – academic articles and twitter feed. Topic modeling using LDA [29] helped us identify the common topics from a list of theoretical factors and Twitter data. We selected the top 50 topics in the two datasets and compared them using word-embedding techniques (word2vector) in the same vector space. We identified the similarity between the topics from Twitter and the academic articles. The similar topics were mapped to the original factors by using the ontology and factors' list.

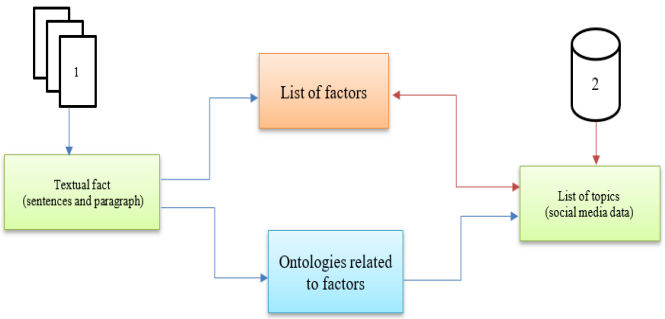

FIGURE II: DATA PROCESSING MODEL FOR TWO SETS OF DATA. THE FIRST DATA SET SHOWED BY NUMBER 1 REFERS TO THE ARTICLES, AND THE SECOND DATA SET SHOED BY NUMBER 2 REFERS TO SOCIAL MEDIA DATA.

D. Factors Extraction

Two types of factor extraction steps were used for our study: 1) manual extraction and 2) automated extraction.

1) Manual Extraction

This process had four steps:

1. Manually annotate and extract all sentences and paragraphs which contain the facts and information about student mobility or student migration.
2. Categorize the extracted items into groups.
3. Provide the ontology for each group. The ontology is based on the article and extracts text in step 1. We should be aware that the existing ontology under the topic of migration, such as what Papadopoulos and Tsianos [27] propose, does not cover the specific area about student migration, which is the scope of this work. Consequently, we had to create the ontology for student mobility using the list of the extracted factors.
4. Assign each group of factors at the list to one of the immigration roles. (Host, Origin, Immigrant)

Table II, III, and IV present the extracted factors which were grouped into the three immigration roles identified by the UN. The process of extraction requires mapping the data from two sets of data sources, so having a comprehensive ontology helps us in the mapping process.

TABLE II: LIST OF FACTORS RELATED TO ORIGINS COUNTRIES

Origin	Skills and quality of education
	Job opportunities and a better life situation
	Globalization and ease of migration and access to the education
	Mobility Culture
	Educated labor
	Population
	Brain drain
	Brain gain
	Economic Factor (distance and GDP)
	stress level
	Unemployment rate
	Fact: rate of returning international student

TABLE III: LIST OF FACTORS RELATED TO HOST COUNTRIES

Host	Skills and quality of education
	Cost of migration (living cost, distance, language, and network.)
	The fact of common host list countries
	Fact: rate of returning international student
	Job opportunities and a better life situation
	Globalization and ease of migration and access to the education
	facilitate immigration
	Mobility Culture
	Educated labor
	Job hunt by immigrant
	Living Cost
	Population
	Rules and policies
	Brain gain
	Colonial power
	Economic Factor (distance and GDP)
	financial benefit
	Students and education provider background

TABLE IV: LIST OF FACTORS FOR IMMIGRANT

Immigrant	Skills and quality of education
	Cost of migration (living cost, distance, language, and network.)
	Immigrant background (Gender-race, Socio-economic situation, language component, personality)
	Unemployment rate
	Job opportunities and a better life situation
	Globalization and ease of migration and access to the education
	Family and financial history
	Facilitate immigration
	Mobility Culture
	Educated labor
	Job hunt by immigrant
	Life skills and personal experiences
	Rules and policies
	Education cost
	Economic Factor (distance and GDP)
	Student financial support
	Students motivation and background
	The success of the previous generation
	A transformation that student faced

2) Automated extraction

The automated process had the following steps:

1. We ran topic modeling (LDA) on collected tweets and selected the top 50 topics. LDA was also used to identify the common words from a list of theoretical factors in the articles, and the top 50 topics from this list were selected.
2. The next step was to find the similarities between the topics of tweets and the articles' topics. For this purpose, all topics were converted to the vectors and the similarities between vectors were assessed using Word embedding W2V technique. Each topic from datasets converted to a vector in a single vector space. The goal was to find the vectors of twitter's topics that were close to the vectors of articles' topics. We can thus determine the specific area of the article's topics that tweets also confirm.

V. RESULTS

In this section, we describe the results obtained by clustering the topics in the academic articles and the twitter data collected by us. As mentioned earlier, W2V is a technique to find the semantic similarity and closeness of the items in each dataset. Results of W2V is an *N*-dimensional matrix, which *N* is the total number of comparing items. To be able to plot the results, we used t-distributed Stochastic Neighbor Embedding (t-SNE) [30] to reduce the dimension of the vectors into two. This reduction makes it possible to visualize the vectors.

Figures V, VI, VII, VIII, and IX show the results' plots. We separately compare twitter's topics to the topics from each article. Consequently, we have five plots drawn in a two-dimensional graph. Each topic is presented by a circle in the plot (blue circles are representing Twitter's topics, and pink circles are representing the articles' topics). For example, figure III shows the two-dimensional form of Twitter's topics and topics from the article by Findlay et al. [8].

Since we considered the articles as the ground truth [6, 8-11], we must find the closeness of twitter data to the extracted data from these articles. Different approaches exist to draw the plots. Creating a bounding box, as presented in figure III or highlighting the cluster group, as shown in figure IV, had been tried to find the best way of visualization. As shown in figure III, boxes cannot help in our clustering approach. The red circles surrounded by the blue circle are the target, while the grid breaks this meaning when a red circle is in one box and very close blue circles are in the other boxes. The second approach (figure IV) does not have this drawback, but it suffers from different spacing for a 360 ° view. All the pink circles are not in the center of the highlighted area, so there is a considerable risk of ignoring the blue circles which are close enough to the target words.

The two mentioned approaches are useful when all the instances have the same level of priority, while in our study, we want to find the closeness of one group of words (twitter's topics) to another group (article's topics). So, we decided to create a threshold for target words (article's topics) to find all the twitter topics which are in the same distance to them. We only show the article's topics with the 0.005 thresholds, and it is the reason that the pink circles are shown in the plot with the bigger diameter. Then, we consider the blue circles, which are in the area of pink circles, as the topics close to the central topic.

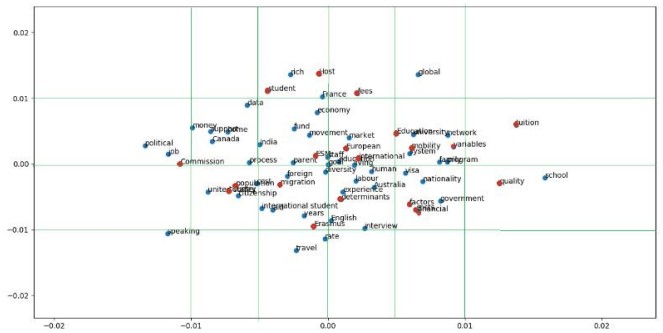

FIGURE III: BOUNDING BOX FORMAT FOR CLUSTERING THE RESULTS.

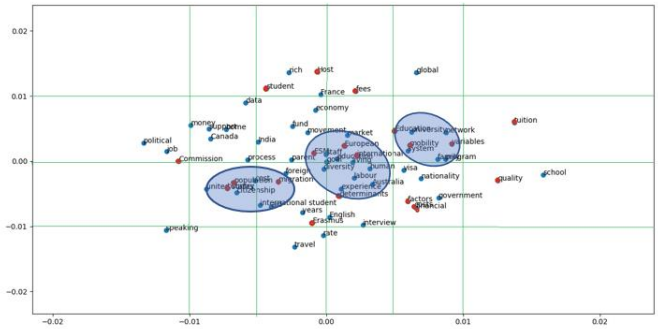

FIGURE IV: IDENTIFYING THE CLUSTER IN THE PLOT BY HIGHLIGHTING THE AREA OF EACH CLUSTER

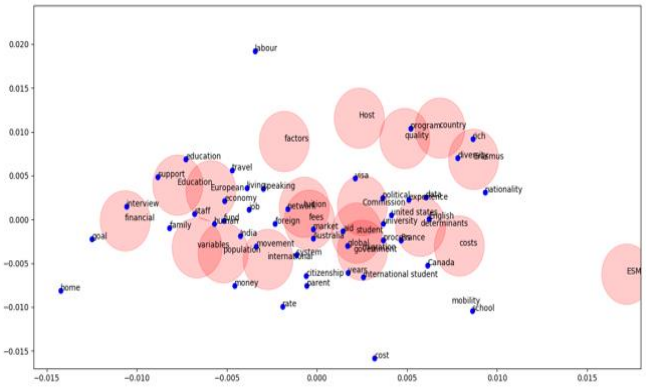

FIGURE V: PLOT OF VECTORS FOR THE FIRST ARTICLE [8] AND TWITTER TOPICS THE PINK CIRCLES REPRESENT THE TOPICS FROM THE FIRST ARTICLE. THE BLUE CIRCLE REPRESENTS THE TOPICS FROM TWITTER. ALL ARTICLES' TOPICS SHOW WITH THE 0.005 THRESHOLDS, SO THEY SHOWED IN THE PLOT BY THE BIGGER CIRCLE. ALL THE BLUE CIRCLE WHICH IS IN THE ARE PINK CIRCLE IS CONSIDERED AS TWO CLOSE TOPICS WITH CLOSE VECTORS. THERE IS THE SAME INFERENCE FOR PLOT VI, VII, VIII, AND IX.

In some cases, when the article's topics are very close to each other and have an overlap, the topics are considered as a set. For example, 'tuition' and 'fees' in figure VI are presented in the same cell in table VI.

For each article, all the blue circles which are close to the pink circles with the distance ≤ 0.005 are related to the article's topics and presented in related tables below each plot.

To have a clear view of the results, we use the topic-factor table (table XI). With the help of this table, we map each topic to the factors that come from; consequently, we can say which factors were observed in twitter's topics.

This work serves as a proof of concept for how a wide range of behavioral features linked to the economic model can be inferred from the digital traces data that are left by publicly-available social media. We demonstrate that behavioral features related to immigration can be validated from the microblogging network Twitter.

TABLE V: THE ARTICLE- TWITTER TOPICS -THE LIST OF TWITTER TOPICS THAT ARE IN THE TARGET AREA OF THE FIRST ARTICLE'S TOPICS [8]. EXTRACTED FROM FIGURE V

Article's topics	Twitter's topics
Financial	interview
Education	Support, education
European	Staff, economy
Variables	Human
population	India, human, fund
International	movement
Tuitions, fees	Networks, foreign, market, Australia
quality	program
Erasmus	Diversity, rich
Commission	Visa, Political, University
Student	Aid, global
government	Aid, global, years, international student
determinants	English, France, experience, data

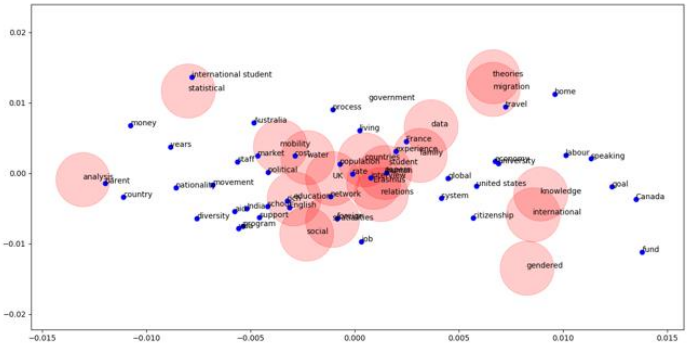

FIGURE VI: PLOT OF VECTORS FOR THE SECOND ARTICLE [9] AND TWITTER

TABLE VI: THE ARTICLE- TWITTER TOPICS -THE LIST OF TWITTER TOPICS THAT ARE IN THE TARGET AREA OF THE SECOND ARTICLE'S TOPICS [9]. EXTRACTED FROM FIGURE VI

Article's topics	Twitter's topics
analysis	parent
Statistical	International student
Mobility	Cost, political, market
Education	Rich, English, school
UK	Population, network, rate
Country	Population, rate, interview
Family	France, experience
migration	Travel

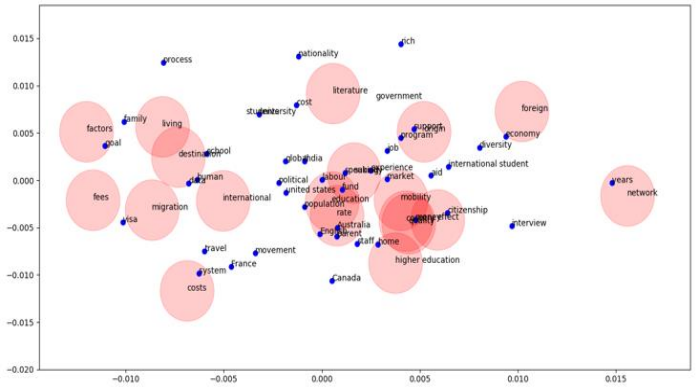

FIGURE VII: PLOT OF VECTORS FOR THE THIRD ARTICLE [10] AND TWITTER TOPICS

TABLE VII: THE ARTICLE- TWITTER TOPICS -THE LIST OF TWITTER TOPICS THAT ARE IN THE TARGET AREA OF THE THIRD ARTICLE'S TOPICS [10]. EXTRACTED FROM FIGURE VII

Article's topics	Twitter's topics
network	Years
foreign	Economy
origin	Support, program
Quality, effect	Citizenship, country,
mobility	Market
Higher education	Home
Education, rate	Australia, English, parent, population, fund, labor
cost	system
Destination	Human, data, school
factors	goal

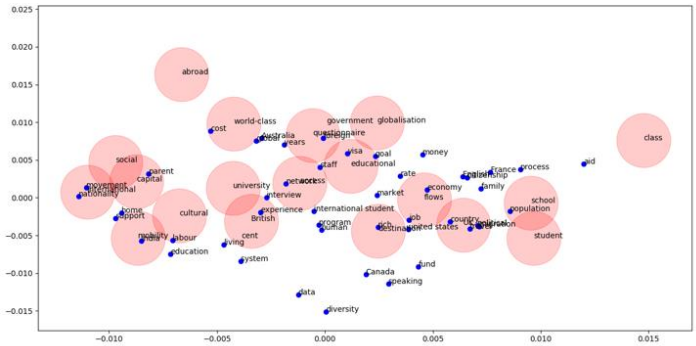

FIGURE VIII: PLOT OF VECTORS FOR THE FOURTH ARTICLE [11] AND TWITTER TOPICS

TABLE VIII: THE ARTICLE- TWITTER TOPICS -THE LIST OF TWITTER TOPICS THAT ARE IN THE TARGET AREA OF THE FOURTH ARTICLE'S TOPICS [11]. EXTRACTED FROM FIGURE VIII

Article's topics	Twitter's topics
International	Movement, nationality
Capital	Parent
Mobility	India
Cultural	Labor
British	Living
Network	Access
questionnaire	Foreign, government, years, visa, goal, staff
Destination	Rich
flow	Economy, job
UK	Country, political, travel, education
school	Population
World class	Cost

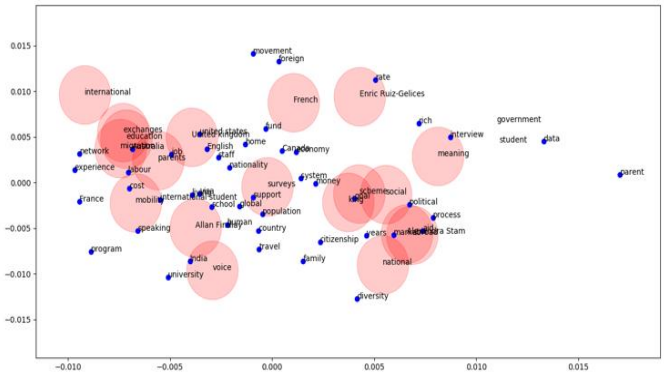

FIGURE IX: PLOT OF VECTORS FOR THE FIFTH ARTICLE [6] AND TWITTER

TABLE IX: THE ARTICLE- TWITTER TOPICS- THE LIST OF TWITTER TOPICS THAT ARE IN THE TARGET AREA OF THE FIFTH ARTICLE'S TOPICS [6]. EXTRACTED FROM FIGURE IX

Article's topics	Twitter's topics
Alexandra Stam	Aid, market, political
Social	Political
King, scheme	goal
meaning	Interview
Enric Ruiz-Gelices	Rate
surveys	Support, population
voice	India
Mobility	Speaking, cost
United Kingdom	United states, English, job
parents	Australia, job
Exchange, education, migration	Australia, labor

At the end of the process, we identified a list of article's topics which has twitter's topics in their close neighbors. Some topics are repeated more than once, which means these topics come from more than one article. Table X illustrates these topics and the number of articles that contain them. Having a quick look at the topics reveals that some of these topics are very close to the keyword list (table I), such as migration, International, and education. Moreover, having the United Kingdom or the UK in this list shows that the selected articles studied the British university more than others.

TABLE X: THE TWITTER TOPICS THAT REPEATED IN THE ARTICLE. THE NUMBER IS THE NUMBER OF ARTICLES WHICH HAVE THIS TOPIC IN THEIR TOPIC LIST

Topics	Number of Articles containing the topic
Mobility	4
Education	4
UK, United Kingdom	3
International	2
Network	2
Quality	2
Migration	2

The closeness of topics "population" and "India" in table V, support the gravity model of migration. The population of each society simulated as a mass in the gravity equation, a populated country like India, can have more emigrants and also accepts more immigrants. Additionally, having a topic "European" and "economy" can relate big economies of European countries to the fact that they the more international students to support the gravity model. Comparing the two topics process has a drawback that it needs manual assessment for inferring the relation. For a more accurate evaluation, we built a mapping table (Table XI). This table identifies the facts that relate to each topic. Based on these facts, we follow the content to relate twitter's topics to the gravity model.

TABLE XI: TOPIC- FACTOR TABLE- THE LEFT COLUMN SHOWS THE LIST OF THE ARTICLES' TOPICS HAVING TWITTER TOPICS IN THEIR NEIGHBORS. THE RIGHT COLUMN SHOWS THE LIST OF RELATED FACTORS FOR EACH TOPIC EXTRACTED FROM THE ARTICLES.

TOPIC	FACTORS
Cost	— Job opportunities and a better life situation – Family and financial history – facilitate immigration – Job hunt by immigrant — Life skills and personal experiences — Living Cost — education cost — Student financial support
Country	– Cost of migration (living cost, distance, language, and network.)
Education	– Cost of migration (living cost, distance, language, and network.) – Immigrant background (Gender-race, Socio-economic situation, language component, personality) – Job opportunities and a better life situation – Globalization and ease of migration and access to the education – Family and financial history – Life skills and personal experiences – Living Cost
Family	– Immigrant background (Gender-race, Socio-economic situation, language component, personality) – Family and financial history – Population
Foreign	– Skills and quality of education – Immigrant background (Gender-race, Socio-economic situation, language component, personality) – Job opportunities and a better life situation – Fact: rate of returning international student – Life skills and personal experiences – Colonial power – Cost of migration (living cost, distance, language, and network.)
Government	– Cost of migration (living cost, distance, language, and network.)
International	– Students and education provider background – Skills and quality of education – Cost of migration (living cost, distance, language, and network)
	– Fact: rate of returning international student – Job opportunities and a better life situation
Migration	– Cost of migration (living cost, distance, language, and network.) – Immigrant background (Gender-race, Socio-economic situation, language component, personality) – Globalization and ease of migration and access to the education – Facilitate immigration – The success of the previous generation — Skills and quality of education
Network	– The success of the previous generation
Population	– Immigrant background (Gender-race, Socio-economic situation, language component, personality) – Students motivation and background – Skills and quality of education
Rate	– Job opportunities and a better life situation – Globalization and ease of migration and access to the education
School	– Educated labor
Social	– Immigrant background (Gender-race, Socio-economic situation, language component, personality) – language component, personality
Student	– Job opportunities and a better life situation – Globalization and ease of migration and access to the education – Facilitate immigration – Mobility Culture – Life skills and personal experiences – Population – Education cost – Student financial support

Table XI shows the list of topics that has the factors for mapping in the topic-factors' table. Some topics do not exist in this table. We use the ontology for mapping this group. We searched the topic in the Ontology list and fetched all factors which mapped to the word. Table XII shows this list.

There is also another list for the words that appear in both common topics but are not mapped to any other factors. This list includes Alexandra Stam, analysis, British, Commission, determinants, Enric Ruiz-Gelices, Erasmus, European, King, meaning, origin, parents, questionnaire, scheme, statistical, surveys, voice, World-class.
Canada, family, money, and process are twitter's topics that never been in any threshold of article topics.

Five related articles were reviewed, and 49 factors extracted manually. The immigration ontology was built based on the factors and related text. The list of topics from twitter data that were semantically similar to the article topics in the two-dimensional vector space was mapped to the factors with the help of factor's list and ontology. Table XIII lists all the student migration factors extracted from twitter data and the frequency percentage of each one.

TABLE XII: TOPICS THAT ARE NOT IN THE FACTOR-TOPIC LIST. THIS GROUP IS MAPPED BASED ON THE ONTOLOGY.

Cultural	– Skills and quality of education – Cost of migration (living cost, distance, language, and network.) – Mobility Culture
Destination	– Cost of migration (living cost, distance, language, and network.) – Job opportunities and a better life situation – Life skills and personal experiences – Living Cost – Rules and policies
Effect	– The success of the previous generation
Exchange	– Skills and quality of education
Factors	– Immigrant background (Gender-race, Socio-economic situation, language component, personality)
Financial	– Cost of migration (living cost, distance, language, and network.) – Family and financial history – Facilitate immigration
Flow	– Skills and quality of education – Cost of migration (living cost, distance, language, and network.) – Fact: common host list countries – Job opportunities and a better life situation
High education	– Skills and quality of education – Immigrant background (Gender-race, Socio-economic situation, language component, personality) – Family and financial history – Education cost
Mobility	– Mobility Culture – Students and education provider background
Quality	– Skills and quality of education – Cost of migration (living cost, distance, language, and network.) – Job opportunities and a better life situation
Tuitions, fees	– facilitate immigration – Living Cost – education cost
UK, United Kingdom	– The fact of common host list countries – Population
Variables	– Job opportunities and a better life situation

As shown in Table XIII, 22 factors out of the total 49 factors mapped from twitter data. We analyzed these factors from Origin-Host-Immigrant point of view and observed that 78% of the immigrant's factors (Table IV) and 83% of the host's factors (table III) were present in the social media data. We were able to validate only 58% of the factors related to origin (table II). As seen in table XIII, the topics related to job opportunities were discussed more frequently for student migration compared to education factors like cost or financial support.

TABLE XIII: FACTORS- FREQUENCY TABLE

Factors	Frequency %
Cost of migration (living cost, distance, language, and network.)	12.5
Job opportunities and a better life situation	11.36364
Skills and quality of education	10.22727
Immigrant background (Gender-race, Socio-economic situation, language component, personality)	9.090909
Facilitate immigration	5.681818
Family and financial history	5.681818
Life skills and personal experiences	5.681818
Globalization and ease of migration and access to the education	4.545455
Living Cost	4.545455
Mobility Culture	3.409091
Population	3.409091
Success of previous generation	3.409091
Education cost	4.545455
Educated labor	2.272727
The fact of common host list countries	2.272727
Fact: rate of returning international student	2.272727
Student financial support	2.272727
Students and education provider background	2.272727
Colonial power	1.136364
Job hunt by immigrant	1.136364
Rules and policies	1.136364
Students motivation and background	1.136364

VI. CONCLUSION AND FUTURE WORK

Social Media applications are collecting a large amount of User-Generated Content (UGC) that contains knowledge about novel approaches of global collaboration. We have developed a hybrid framework using qualitative and quantitative approaches to analyze the collaborative approaches used by student migrants and potential migrants when sharing information about student migration strategies on social media, specifically on Twitter. We reviewed the current academic models proposed on migration and identified the key topics discussed on migration. We also collected over 10 million tweets based on student migration topics that were identified by our domain expert. We created two layers of clustering methods for mapping our Twitter data and extracted factors from socioeconomics articles. Topic modeling (LDA) and word embedding (W2V) were the two methods used for clustering.

Our analysis was able to validate that a large percentage of Immigrant related factors, identified by research scholars, was actively discussed in social media. However, only about 58% of factors influencing the country of origin were a point of discussion by the collaborators. These numbers indicate that immigrants do not rank those factors as high as social science researchers do.

As part of our ongoing work, we are further analyzing the factors and refining the main points of discussion by merging semantically similar topics. We will also be collecting similar

datasets from other social media sites, like Reddit and Facebook, to determine if the same factors are discussed on those sites as well. We also plan on releasing our dataset to the research community.

REFERENCES

[1] "Migration data portal," last retrieved 8/15/19. [Online]. Available: https://migrationdataportal.org/themes/big-data

[2] U. N. DESA, "United Nations, Department of Economic and Social Affairs, Population Division (2017). International Migration Report: Highlights(ST/ESA/SER.A/404).," 2017.

[3] B. Edmonston, and J. P. Smith, The immigration debate: studies on the economic, demographic, and fiscal effects of immigration, *Washington, D.C.: National Academy Press*, 1998.

[4] M. Harvey, T. Kiessling, and M. Moeller, "Globalization and the inward flow of immigrants: *Issues associated with the repatriation of global managers," Human Resource Development Quarterly*, vol. 22, pp. 177-194, 2011.

[5] J. E. Anderson, "The gravity models," *Annu. Rev. Econ., vol. 3, pp. 133-160.*, 2011.

[6] F. Simini, M. C., González, A. Maritan, and A. L. Barabási, "A universal model for mobility and migration patterns," *Nature, vol. 484, p. 96–100*, 2012

[7] W. Tobler, "Migration: Ravenstein, Haythornthwaite, and beyond," *Urban Geography, vol. 16, pp. 327-343*, 1995

[8] A. M. Findlay, A. Stam, R. King, and E. R. Gelices, "International opportunities: *searching for the meaning of student migration," Geographica Helvetica, vol. 60, pp. 192-200*

[9] L. Komito, "Social media and migration: Virtual community 2.0," *Journal of the American Society for information science and technology, vol. 62, pp. 1075-1086*, 2011.

[10] M Beine, R Noël, and L Ragot, "Determinants of the international mobility of students," *Econ. Educ. Rev., vol. 41, pp. 40-54*, 2014.

[11] D. Karemera, V. I. Oguledo, and B. Davis, "A gravity model analysis of international migration to North America," *Appl. Econ., vol. 32, pp. 1745-1755*, 2000.

[12] A. Crymble, "Introduction to Gravity Models of Migration & Trade," *in The Programming Historian*, 2019.

[13] J. Poot, O. Alimi, M. P. Cameron, and D. C. Maré, "The gravity model of migration: *the successful comeback of an aging superstar in regional science,"* 2016.

[14] Deardorff, Alan, "Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World?" The Regionalization of the World Economy,1998

[15] C. Hughes, E. Zagheni, G. J. Abel, A. Wisniowski, A. Sorichetta, I. Weber, and A. J. Tatem, "Inferring Migrations: Traditional Methods and New Approaches based on Mobile Phone, Social Media, and other Big Data: Feasibility study on Inferring (labor) mobility and migration in the European Union from big data and social media data," 2016.

[16] E McGregor, and M Siegel, "Social media and migration research," *United Nations University-Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT)*, 2013

[17] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located Twitter as a proxy for global mobility patterns," *Cartogr Geogr If. Sc., vol. 41, pp. 260-271*, 2014.

[18] A. L. Pinto, M. Garcia-Herranz, M. Cebrian, and E. Moro, "Social Media Fingerprints of Unemployment," *Plos One, vol. 10*, 2015

[19] B. State, M. Rodriguez, D. Helbing, and E. Zagheni, "Migration of professionals to the US," *in International Conference on Social Informatics*, 2014.

[20] P. Grover, and A. K. Kar, "User engagement for mobile payment service providers–introducing the social media engagement model," *J. Retail. Consum. Serv., p. In Press*, 2018.

[21] L. Hong, B. D. Davison, "Empirical study of topic modeling in twitter," *in Proceedings of the first workshop on social media analytics, Washington D.C., ACM* 2010.

[22] A. McCallum, X. Wang, and N. Mohanty, "Joint group and topic discovery from relations and text," *in In ICML Workshop on Statistical Network Analysis, Berlin, Heidelberg*, 2016.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv, p. 1301.3781*, 2013.

[24] W.Zhao, J.Wayne X., J Jing,J. Weng Jianshu,J. He Jing,E. Lim. 2011. "Comparing Twitter and Traditional Media Using Topic Models. " *In Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 1821, Proceedings. Berlin: Springer Verlag*, 2011.

[25] Nguyen, Dat Quoc, et al. "Improving topic models with latent feature word representations." *Transactions of the Association for Computational Linguistics 3*,2015

[26] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," 2011.

[27] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist., vol. 22, pp. 39-71*, 1996

[28] D. Papadopoulos, and V. S. Tsianos, "After citizenship: the autonomy of migration, organizational ontology, and mobile commons," *Citizsh. Stud., vol. 17, pp. 178-196*, 2013.

[29] D.M. Blei, A Y. Ng, M. I. Jordan,"Latent Dirichlet Allocation", *Journal of Machine Learning research, 3(Jan):993-1022*, 2003.

[30] van de Maaten, L.J.P.; Hinton, G.E. "Visualizing Data Using t-SNE" (PDF). *Journal of Machine Learning Research. 9: 2579–2605*, 2008