RESEARCHING PEOPLE'S MOODS ON TWITTER IN ORDER TO EVALUATE THE IMPACT OF IMMIGRATION

D.Siva[1], Dr.Agusthiyar R[2]

[1]Assistant Professor, Department of Computer Science, SRM IST, Ramapuram Campus, Chennai-89.

[2] Professor & Head, Department of Computer Applications, SRM IST, Ramapuram Campus, Chennai-89.

[1]sivad @srmist.edu.in

[2]hod.dca.rmp@srmist.edu.in

Abstract

Moving to a new country necessitates learning a new culture and adjusting to a new environment. They also face culture shock and find it difficult to adjust to their new life in a foreign country. According to experts in the domains of psychology and sociology, there is a larger risk of psychological distress among immigrants during their first few months in the country. We examine the sentiments of immigrants through their tweets in this study. We start by identifying immigrants on Twitter, then look at the types of tweets they send. Based on the messages individuals send through this platform, we investigate their attitudes. We used geo-location meta-data from Twitter to identify immigrants. Using a machine learning approach, we further categorised the tweets into three categories: anxiety, dejection, and aggression

Keywords: *sentiment analysis, mood analysis, natural language processing*

1. Introduction

Studies in Social Science indicate a higher risk of psychological distress for immigrants. This is illustrated by measuring different features such as suicide rate, self-reported psychiatric illness, intake of psychotropic drugs, and desire to leave the new country. With increased popularity ofonline social networks, a new data source is available for researchers to study psychological and behavioral features ofthe users. In this project we focus on Twitter and the tweets made by immigrants to analyze how their psychological features represented in their tweets have changed before andafter their immigration. The psychological featuresinvestigated in this study are dejection, anxiety, and hostility.Twitter is amongst the widely used micro blogging sites where individuals from all corners of the world exchange their thoughts daily in the form of tweets [1]. It is selected asthe base of the study primarily because this micro-blogging service provides publicly available and highly frequent largemessage volume. Hashtags are an important keystone of the tweets that increases their visibility. In terms of information discovery and knowledge creation, this plethora of user created content allows the application of sentiment analysis, which aims to provide an automated mechanism fordetermining the writer's attitude towards the subject or its overall contextual polarity [2].

A. *Problem Statement*

The primary research question in this project can bephrased as: How does immigration impact mental state of theimmigrants shown in their tweets? We break down this question into two sub-problems:

1. How can we identify immigrants on Twitter?

How can we analyze the sentiment of the tweets toextract psychological features?

For the first part we relied on geo-locations of the tweetsto detect the location of the users and see if they havechanged their location for a long time. And for the second part we used a set of manually labeled training data and a Logistic Regression model to analyze the sentiment of the tweets.

B. *Structure*

Section 2 focuses on the related work done in this area. This section first illustrates the traditional approaches done in Social Science and Psychology that studied behavioral and mental features of immigrants. Then it lists the efforts in sentiment analysis of the tweets related to other applications.Section 3 describes our data collection method and how we identified immigrants on Twitter. Section 4 then explains theSentiment Analysis part of the project where we look for themoods expressed on tweets made by the identified immigrants. Section 5 then details our 'Difference in Differences' method and how we evaluated the achieved results. Section 6 then lists the concluding remarks as well as potential future directions for this project.

2. *Related Works*

Research works and studies related to this project can becategorized into three groups;

1. Traditional approaches in Social Science and Psychology that study the impact of

immigration onpeople's mood

2. Research works done in geo-locating twitter usersusing the texts of the tweets and tweets' meta-data

3. Methods to analyze the sentiment of tweets by theircontents

A. *Traditional Approaches*

A study by World Health Organization [3] indicates that the rate of suicide increases between immigrants in the United States. They compared the rate of suicide between people who have immigrated to USA and the people of the native country who have not made the immigration. Figure 1 visualizes the results for each studied country. This gives theinitial idea that immigration potentially impacts people in terms of their psychological features. A similar study [4] hasbeen done for immigrants to Canada and argues that the suicide rate in immigrants to Canada have generallyincreased compared with the suicide rate in their home *country.* The result suggests that the immigration is likely tocause depression, and anxiety.

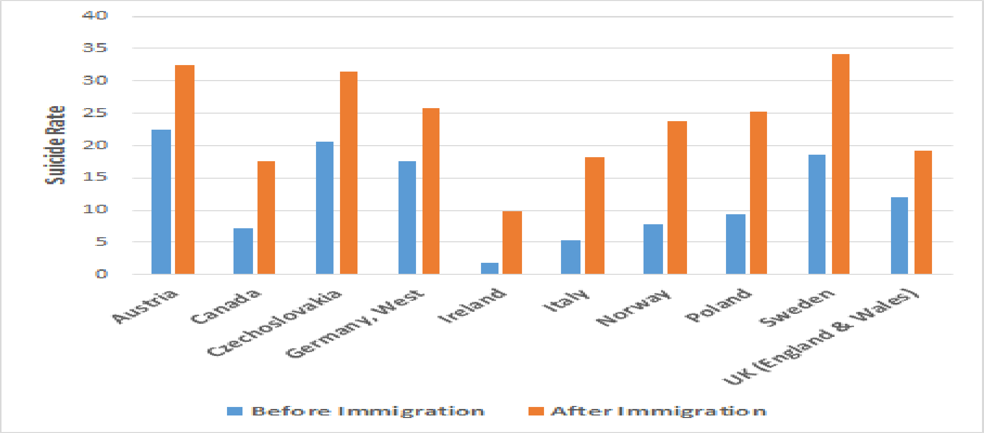

Figure 1. Suicide rate per 100,000 people per year between immigrants to the United States compared with their source country

The authors of role of marital status, duration of residence, and social integration [5] studied immigrants in the state of California between 1998 to 2001 and by comparing the suicide rate between immigrants and native people, showed that immigration increased suicide risk. It is argued that the depression is significantly more in divorced and single immigrants than regular immigrants. Moreover, their results indicate that the suicide risk is more for people who have immigrated recently and have a shorter duration ofstay. In another comprehensive but more recent study [6], therate of suicide and other parameters impacted by immigration are investigated. Globalization, acculturationand acculturation stress, genetic and environmental factors, urbanization and ethnic density, first or second generation ofimmigrants, being in adolescents, and being a refugee and asylum seeker, are among the primary parameters discussed in this research that might be the cause of increased suicide and depression rate between immigrants.

Westman et al. [7] categorized the suicides attempts in Sweden based on the place of birth and gender of the personwho attempted the suicide. The paper divided the number of attempted suicides to the total population of immigrants for each home country in Sweden, to calculate the rate of suicides for in immigrants based on their home country. Theresult illustrated that the suicide rate for immigrants from almost every country is more than the native Swedish people.For example, the suicide rate of Iranian immigrants in Sweden is twice more than the rate of native people ofSweden, in spite of the fact that the suicide rate of native Iranians living in Iran is half the suicide rate of people from Sweden.

Potochnick et al. [8] in yet another study, actively monitored the psychological features of more than 500 Mexican immigrants in the United States and concluded thatAfter arrival in the United States, migrants had a significantly higher risk for first onset of any depressive or anxiety disorder than did non-immigrant family members of migrants in Mexico.

Leersnydera et al. [9] in a psychological research done manually and based on direct interaction with people and might not be very closely related to our project. Nevertheless, the research revolves around a similar research question as what we are targeting, which is the impact of immigration on one’s psychological feature. The research studied Korean immigrants in the US, and Turkish immigrants in Belgium, with this research assumption that the emotional experiences of people who live together become similar. The result show that an emotional acculturation exists in immigrants.

In conclusion, our literature review in the field of Social Science and Psychology illustrate that immigrants have a higher risk of dejection, depression, and ultimately suicide rate.

B. Geo-locating Twitter Users

As part of the project, we need to identify immigrants onTwitter. Therefore, we look into geo-locate tweets and Twitter accounts to see if a user has immigrated or not. This section surveys related work in regard to geo-locating twitteraccounts using either the tweet’s contents or meta-data. Cheng et al. [10] built a model that successfully estimated the location of 51% of their testing twitter accounts with the precision of up to 100 miles only based on the textual contents of the tweets. Their approach focuses on the keywords used in tweets that might be associated with a certain location, such as a particular expression, a name of a place, or a name of a local event. Backstorm et al. [11] assumed that the likelihood of friendship with a person decreases with distance, therefore people tend to be located close to their friends. The idea is used in this research to build an algorithm that can anticipate the location of a socialnetwork’s users based on the location of their friends. It is illustrated that the proposed algorithm outperforms IP-based geolocation.

Amitay et al. [12] in their research made this goal to geotag web-pages based on their contents. The approach is toidentify keywords in the web-pages that might be associated with different places. However, the challenge targeted in this paper is to resolve ambiguities such as those keywords that might represent different locations, and those that represent both location and nonlocation meanings. In order to do so, the whole contents of the page are involved in their proposedalgorithm to give a score to each potential location for the web-page.

Another research [13] on geo-locating web contents on Social Networks studies user's activities on Foursquare social network where users are able to check in their locations. The study attempts to identify a correlation between the location of the users and their activities to build a framework for a potential recommendation system. Geo- locating the users are solely based on the coordination of theusers that they set manually by checking in on the app. This is similar to tweet's location. However, Twitter users tend toturn off their location more than Foursquare users.

Foursquare social network has been the subject of another research [14] to study the locations of social networkusers with a goal of building a human mobility pattern. This study also considers the manually set check-ins from the users in social networks for to identify the location of the users. This is how their approach is different from content-based geo-location systems. In another research [15] in shaping human mobility based on their locations on social networks, the locations are based on what the users manuallyset by checking in on social networks such as Foursquare and Facebook. The result of the paper is a proposed model forhuman mobility patterns in urban metropolitan cities.

A more recent study [16] is closely related to our project in term of identifying cross-border mobility of social network users. The research monitors the location of users on location based social networks to capture cross border movements. This interests us as we identify such mobilities with some additional constraints as immigration, and the user would be important for us as an immigrant. Moreover, the research also analyzes twitter accounts and the location of the tweets as well.

C. Sentiment Analysis on Twitter

As discussed in Section 1.1, the second sub-problem in this project is to find a methodology to anticipate psychological features of a twitter user, in particular depression by analyzing the sentiment of their tweets. Whilethe goal is for twitter accounts, other proposed approaches that analyzes the sentiment and behavior of a user based on their web contents might be interested for us as well. Researches done in the fields of Sociology and Psychology are also considered for this purpose.

In a research done by Carr et al. [17] the goal is to determine if online tweets represent the idea and opinions ofthe person who submitted them on Twitter. For instance in this research they considered the issue of immigration and concluded that by analyzing the tweets one sent, we are ableto determine if they are in favor or opposed to the issue. The analysis and categorizing the tweets however, were done mostly manually to determine if a tweet is positive or negative about the issue.

Kouloumpis et al. [18] investigated the application of Sentiment Analysis and Natural Language Processing in Twitter. It is argued that previous researches in this regard explored the use of Part-of-Speech features and were not significantly successful due to the incredible breadth of topicthat tweets of a user cover. Therefore, they focused on hashtags used in tweets to mine the sentiment of the tweets, besides lexicon, n-gram, and Part-of-Speech features. The combination of all features resulted in the best performance in term of accuracy based on a training data.

Agarwal et al. [19] have done a study in sentiment analysis of tweets. The proposed methodology has threemajor phases; preprocessing, scoring, and design of a Tree Kernel.

The preprocessing is done to organize emoticons, URLs, user mentions, and acronyms such as LOL and gr8. The second phase is to score each word based on being positive or negative using a dictionary of 8000 words that arealready scored and WordNet. Finally, they build a tree of each tweet and score a tweet based on a large list of proposedfeatures.

Bollen et al. [20] in their research to find the mood of tweets, used profile of Mood States psychometric instrument which measures six individual dimensions of mood (Tension, Depression, Anger, Vigour, Fatigue, Confusion). After data preparation and normalizing tweets, the unit mood vector is produced which shows the total score of each tweet based on6 dimensions.

In another study [21], the authors built a sentiment classifier, which specifies negative, positive and neutral sentiment of texts. The first step is to collect required corpus.In this paper tweets are collected with special emoticons. (:) ,:)) or :(, (() which show negative and positive sentiment. Then, TreeTagger is used to find POS tags of terms intweets. After this part they paid attention to the differences between tags distribution. The distribution of POS-tags shows special tags for each positive and negative and neutralsets. Using this information and also construction of n-grams(to handle negation opinion) the sentiment classifier is built using Naive Bayes classifier.

Two closely related researches studied the appearance ofdepression in new mothers. Choudhury et al. in a very closely related research looked for new mothers on twitter and analyzed changes in their emotion and behavior after they gave birth to their babies. They used Queries on TwitterFirehose (available to Microsoft only) to find potential new mother twitter accounts. Then they used Amazon's Mechanical Turk to identify actual new mothers. Then they measured the following parameters to investigate behavioralchanges:

1. Engagement: Number of posts per day, Number of replies per day, Number of retweets per day
2. Ego-network: Number of followers, Number of followees
3. Emotion: Positive Affect (PA) and Negative Affect (NA) both achieved using LIWC lexicon, Activation and Dominance both achieved using ANEW lexicon
4. Linguistic Style: specific linguistic styles achieved by LIWC lexicon

3.FINDING IMMIGRANTS ON TWITTER

The first phase of the project is to identify a set of twitteraccounts who have made an immigration and somehow expressed that in their tweets. Literature review discussed in Section 2.2 showed a number of research works that could successfully geo-locate tweets. What we in fact are looking for is the same concept; determine the location of each tweet of the user and see if they have changed their location.Nevertheless, there existed an alternative approach which was relying on the tweet's geo-tag data provided by Twitter API. Since, a portion of the users disable this tag for their tweets, we had to collect tweets from a large volume of usersto make sure that we have enough users with geo-location enabled.

As a seed for our data collection process, we started by collecting data from the followers of Twitter accounts ownedby Department of US Citizenship and Immigration Service, and department of Citizenship and Immigration Canada. Limited by Twitter API we collected the last 3200 tweets from each user and filtered out those who are private accounts with inaccessible tweets, have less than 100 tweets, whose language were not English, and do not provide geolocation of their tweets.

Table 1. Top 5 destination countries in the identified immigrants

Destination Country	Number of Immigration
United States	312
Canada	210
India	64
United Kingdom	28
Mexico	27

This left us with 22,096 users with total of 116GB of data. The next phase would be to identify immigrants between our collected users, being careful not to consider those who made a trip, as immigrants. In particular we look for Twitter accounts with a change in the country where theymade tweets. We need to consider a timeline of the tweets for each user to specify the date in which the location of the tweets have changed. Moreover, we ignore those users who have moved from location A to location B and then after some time moved back to location A, or only have stayed in location B for less than 3 months. The following algorithm provides a pseudo-code to the algorithm we implemented for identifying immigrants.

```
Result: True if the given twitter account is immigrant let Timeline be List of Tweet objects
containing (Date,Text, Country)
let Country Set be list of unique country namesordered by first appearance date
if len(Country Set) >1 then

if Timeline[ 1].Country == Country Set[ 1]
then
        if Duration of Country Set[ 1] >3 Months then else  else return true
else
return false
end
```

The algorithm enabled us to identify 956 immigrants. Table 1 and 2 shows the top 5 source and destination countries for our identified immigrants, and Figure 2 illustrates their top 10 paths of the immigrations.

Table 2. Top 5 source countries in the identified immigrants

Source Country	Number of Immigration
United States	215
United Kingdom	89
Canada	75
India	56
Mexico	41

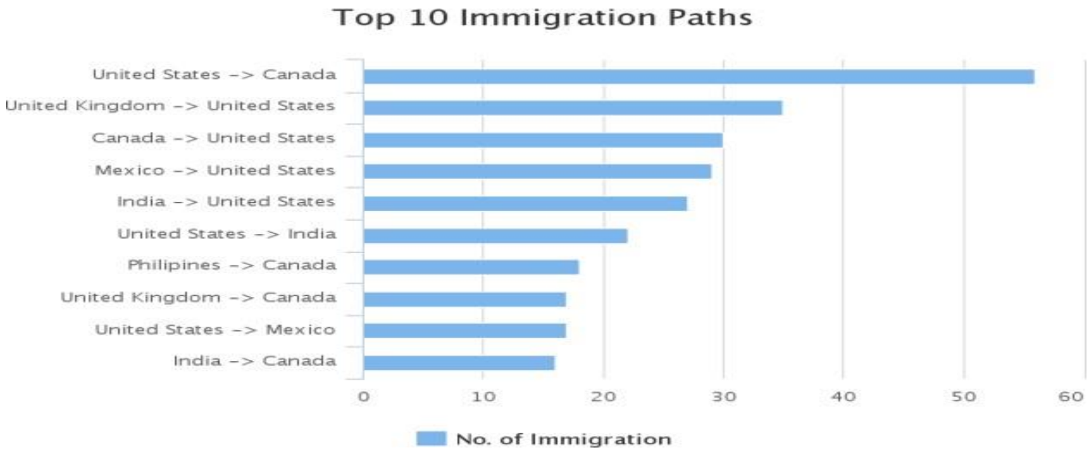

Figure 2. Top 10 paths of immigrations in identified immigrants

Our data collection brings us to two crucial questions;

1. How accurate is our results? Have the identifiedimmigrants actually made an immigration?
2. How does our set of identified immigrants
3. represent the actual demography of the immigrants? Have we made a good sample for our data analysis?

To address the first question, we manually analyzed the profiles of a random selection of ten users of the identified immigrants. We could confirm theimmigration for seven of them as it clearly showed signs of immigration in their profiles and the texts of their tweets. The other 3 did not show immigration sings nor was it clear that they actually have not made an immigration.

Regarding the second question, we compared the demography of the identified immigrants in terms of gender and source country with the demography of actual immigrants in the world. In order to do so we only focused on US immigrants and extracted those identified Twitter immigrant accounts whose destination country was United States. We also used data collected in two studies [22, 23] of demography of immigrants in the United States. Figure 3 showsthe comparison between distribution of our identified immigrants with actual immigrants in the US in term of gender, and Table 3 illustrates the same comparison in term of the rank of the source country. Although this comparison only focused on the immigrants to the United States, the results give a rough estimate on how the distribution of identifiedset of immigrants is compared to the actual demography of the actual immigrants. The comparison could become more accurate if we could get more data from Twitter such as age, education, income, job, etc., however it is not possible.

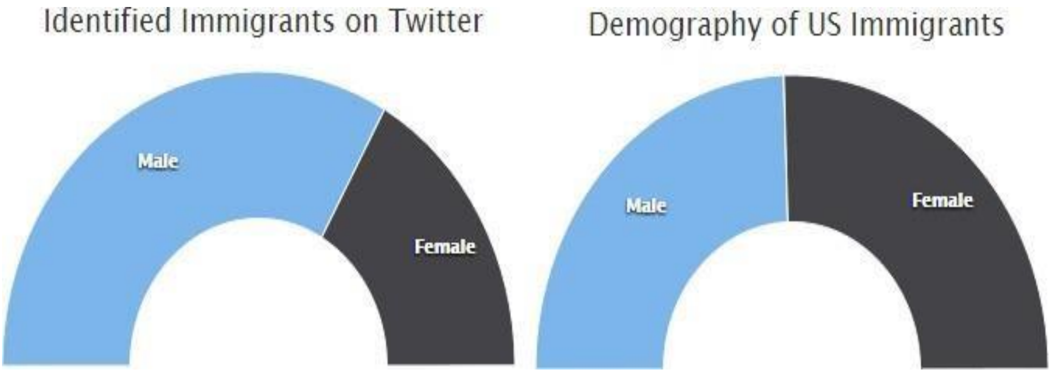

Figure 3. Comparing the diversity of the gender between US actual immigrants and US identified immigrants on Twitter

Table 3. Rank of the source countries of the immigrants to the United States. Comparing the actual demography of the US im- migrants with the US immigrants identified on Twitter

Country	Rank in Identified Immigrants	Rank in Actual Demography
United Kingdom	1	13
Canada	2	11
Mexico	3	1
India	4	3

4.TWEETS SENTIMENT ANALYSISFINDING IMMIGRANTS ON TWITTER

In order to investigate the mood expressed on the tweets we built a classifier trained by Logistic Regression to analyze the sentiment of the tweets in terms of three class labels; dejection, anxiety, and hostility.

A.Training the Classifier

The training data that we used to train the classifier is a set of 2367 manually labeled tweets for the three desired class labels. 31.4% of the tweets in our training data have a positive label for at least one of the class labels i.e., dejection, anxiety, and hostility. Figure 4 illustrates how our training data is distributed between the classes. In order to evaluate the performance of our classifier we used a 10-fold cross validation and Table 4 shows values in regards to the accuracy of the classifier.

B.Results

Using the Logistic Regression classifier, we got the probability of each class label for every tweet of each identified immigrant. Then we focused on the dejection class and calculated the average probability of all the tweets for each immigrant once for the tweets made before his or her immigration and once for the tweets made after the immigration. Figure 5 illustrates the difference between the average probability of dejection after the immigration and before the immigration for those immigrants in which this difference is more than 1%. The result indicates that the dejection expressed in tweets has generally increased in average after the immigration for most number of identified immigrants.

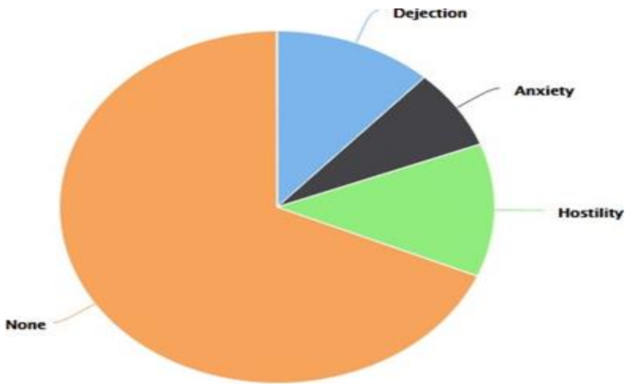

Figure 4. Distribution of the class labels in the training data

Table 4. Accuracies for the classifier trained by each class label

Positive			NegativeClass	Class
Anxiety		Precision	0.9848	0.9424
		Recall	0.6824	0.9979
		F-measure	0.3141	0.9729
		Accuracy	84.62%	
Hostility		Precision	0.9867	0.9811
		Recall	0.6506	0.9990
		F-measure	0.2809	0.9521
		Accuracy	91.04%	

5.*DIFFERENCE-IN-DIFFERENCES ANALYSIS*

Difference-in-differences (DD) analysis is a popular method in economics, statistics, and quantitative research to estimate causal relationships. This section describes how we utilized this technique to verify our findings. We identified one similar user account who has not made any im- migration for each identified immigrant that has the same gender and is from exactly the same country and has a similar features set that contains rate of tweets 200 days before the immigration, rate of replies (user mentions) 200days before the immigration, and date of account creation. In

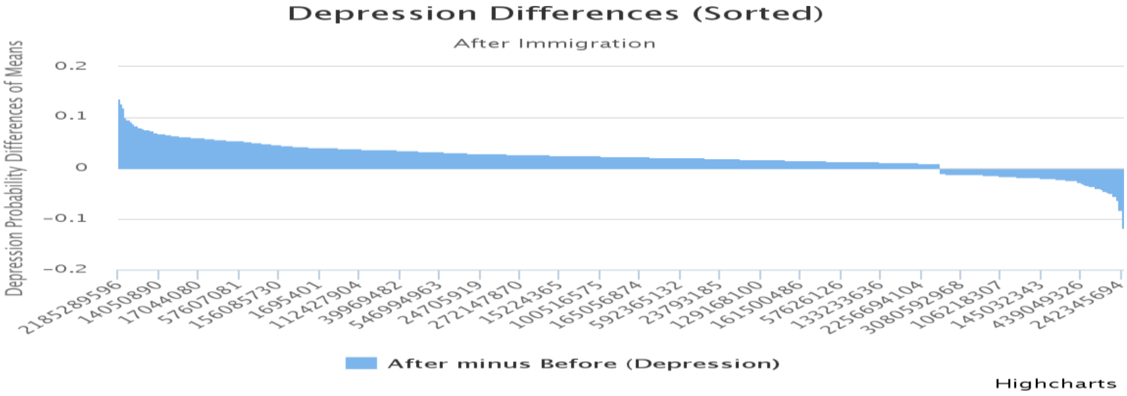

Figure 5. Distribution of the class labels in the training data

order to evaluate the similarity we used co-sine similarity. Asa result we have a set of similar non-immigrant twitter accounts as many as the identified immigrants. The next step is to use our classifier to measure the probability of every tweet from 200 days before the immigration to 200 days afterthe immigration for each user and calculate the average trend for all the users once for immigrants and once for similar non-immigrants. Figure 6 illustrates the comparison of the trends for each class label between the immigrants and their corresponding similar non-immigrants. The result shows a slight increase in hostility, dejection, and anxiety inimmigrants while the similar users experienced a flat trend.

Finally, we get the mean value for all the probabilities of all the tweets made before and after immigration by all users.

Table 4. Accuracies for the classifier trained by each class label

		Positive Class	Negative Class
Dejection	Precision	0.9843	0.9445
	Recall	0.6829	0.9979
	F-measure	0.3778	0.9226
	Accuracy	86.27%	

Figure 7 shows that the difference between the values achieved for the immigrants is not significant with the valuesachieved for their similar non-immigrants. The difference value is 0.000793393 and does not seem to be significantly different from zero (given the size of the variance).

6.CONCLUSION AND FUTURE WORK

In this paper we successfully identified a set of Twitter accounts who have made an immigration from one country to another. Using sentiment analysis with a Logistic Regression model we classified their tweets basedon three class labels; dejection, anxiety, and hostility. Our results indicated a slight increase in all class labels after theimmigration. Potential future work for this project includesbut are not limited to:

1. Focus on individual countries and investigate howimmigrants to that particular country has changedtheir mood expressed on their tweets.
2. Include more meta-data of the tweets such as number of retweets and URL share. This could potentially indicate sense of sociability on Twitter by users. We can then study how more or less socialized the person became after the immigration.
3. Repeat the process on data collected in socialnetworks such as Facebook.

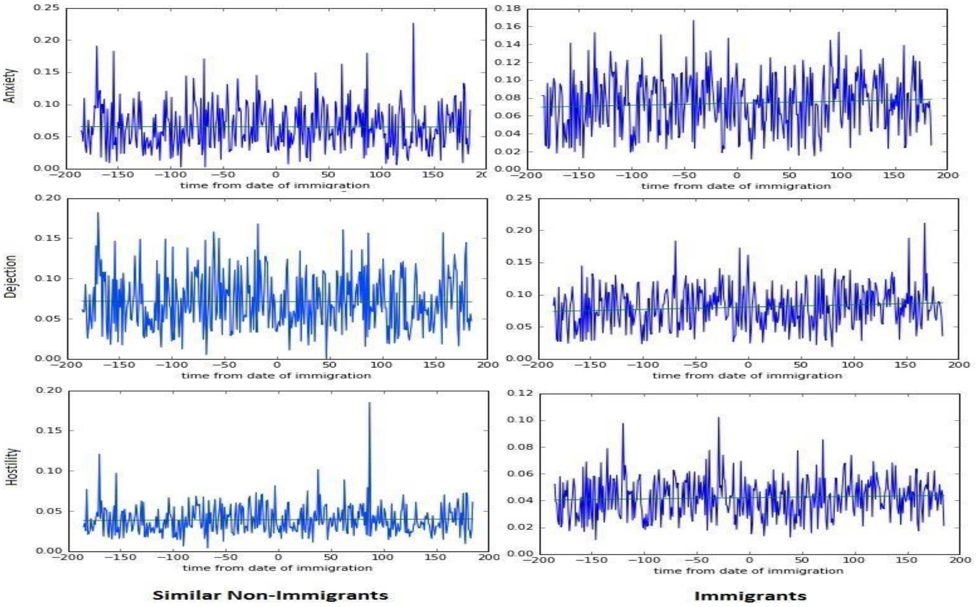

Figure 6. Difference in Differences analysis: The trend for each class label before and after the immigration in both immigrants and their similar accounts.

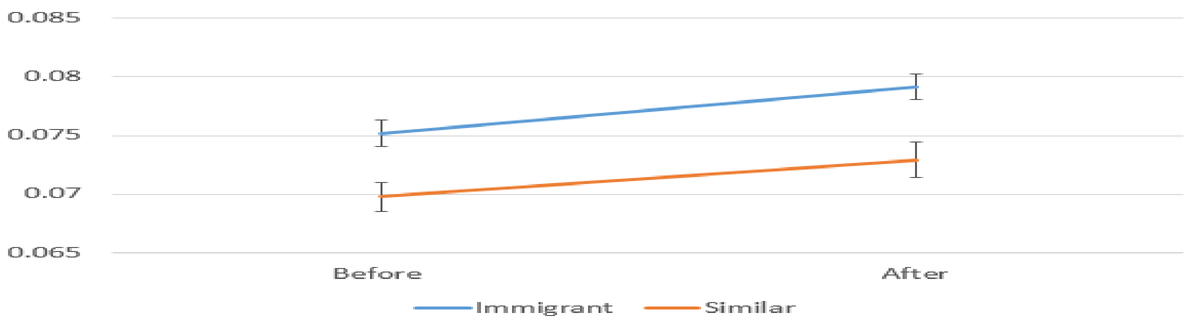

Figure 7. Total Mean of dejection probablity for All before and after immigration

ACKNOWLEDGEMENT

We thank the anonymous referees for their useful suggestions.

REFERENCES

[1] Abhilash Mittal and Sanjay Patidar. 2019. Sentiment Analysis on Twitter Data: A Survey. In Proceedings of the 2019 7th International Conference on Computer and Communications Management (ICCCM 2019). Association for Computing Machinery, New York, NY, USA, 91–95. DOI:https://doi.org/10.1145/3348445.3348466

[2] Jour Khan, Atif Khattak, Asad Masood Batool, Rabia Satti, Fahad Ahmed Hussain, Jamil Khan, Wajahat Ali Khan, Adil Mehmood Hayat, Bashir. 2020. Tweets Classification and Sentiment Analysis for Personalized Tweets Recommendation. 1076-2787 https://doi.org/10.1155/2020/8892552 10.1155/2020/8892552

[3] Organization, World Health. Suicide rates among immigrants to the us and in their former country, 1996. URL http://suicidemethods.info/ tables/immgrant.htm.

[4] Malenfant, Eric C. Suicide in canada's immigrant population. *Health Reports*, 15(2):9, 2004.

[5] Kposowa, Augustine J., McElvain, James P., and Breault, Kevin D. Immigration and suicide: The role of marital status, duration of residence, and social integration. Archives of Suicide Research, 12(1):82–92, 2008. doi: 10.1080/13811110701801044. URL http:// dx.doi.org/10.1080/13811110701801044. PMID: 18240038.

[6] Ratkowska, Katarzyna Anna and De Leo, Diego. Suicide in immigrants: an overview. 2013.

[7] Westman, Jeanette, Hasselstrom, Jan, Johansson, Sven- ¨ Erik, and Sundquist, Jan. The influences of place of birth and socioeconomic factors on attempted suicide in a defined population of 4.5 million people. Archives of general psychiatry, 60(4):409–414, 2003.

[8] Breslau, Joshua, Borges, Guilherme, Tancredi, Daniel, Saito, Naomi, Kravitz, Richard, Hinton, Ladson, Vega, William, Medina-Mora, Maria Elena, and AguilarGaxiola, Sergio. Migration from mexico to the united states and subsequent risk for depressive and anxiety

disorders: a cross-national study. Archives of General Psychiatry, 68(4):428–433, 2011.

[9] De Leersnyder, Jozefien, Mesquita, Batja, and Kim, Heejung S. Where do my emotions belong? a study of immigrants emotional acculturation. Personality and Social Psychology Bulletin, pp. 0146167211399103, 2011.

[10] Cheng, Zhiyuan, Caverlee, James, and Lee, Kyumin. You are where you tweet: a content-based approach to geolocating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 759–768. ACM, 2010.

[11] Backstrom, Lars, Sun, Eric, and Marlow, Cameron. Find me if you can: improving geographical prediction with social and spatial proximity. In Proceedings of the 19th international conference on World wide web, pp. 61–70. ACM, 2010.

[12] Amitay, Einat, Har'El, Nadav, Sivan, Ron, and Soffer, Aya. Web-a- where: geotagging web content. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 273–280. ACM, 2004.

[13] Noulas, Anastasios, Scellato, Salvatore, Mascolo, Cecilia, and Pontil, Massimiliano. An empirical study of geographic user activity patterns in foursquare. ICwSM, 11: 70–573, 2011.

[14] Cheng, Zhiyuan, Caverlee, James, Lee, Kyumin, and Sui, Daniel Z. Exploring millions of footprints in location sharing services. ICWSM, 2011:81–88, 2011.

[15] Noulas, Anastasios, Scellato, Salvatore, Lambiotte, Renaud, Pontil, Massimiliano, and Mascolo, Cecilia. A tale of many cities: universal patterns in human urban mobility. PloS one, 7(5):e37027, 2012.

[16] Blanford, Justine I, Huang, Zhuojie, Savelyev, Alexander, and MacEachren, Alan M. Geo-located tweets. enhanc- ing mobility maps and capturing cross-border move- ment. PloS one, 10(6):e0129202, 2015.

[17] Carr, Jason D. Measuring twitter sentiment and implications for social psychological research. Available at SSRN 2499736, 2014.

[18] Kouloumpis, Efthymios, Wilson, Theresa, and Moore, Johanna D. Twitter sentiment analysis: The good the bad and the omg! Icwsm, 11:538–541, 2011.Carr, Jason D. Measuring twitter sentiment and impli- cations for social psychological research. *Available at SSRN 2499736*, 2014.

[19] Cheng, Zhiyuan, Caverlee, James, and Lee, Kyumin. You are where you tweet: a content-based approach to geo- locating twitter users.