Comparing the Interests of People in Different Cities Using Topic Analysis and Social Media

Final Report

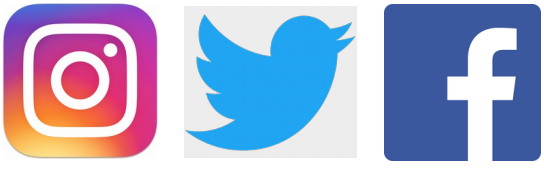
Jake Lasley
Computer Science
The University of Texas at El Paso
El Paso, Texas
jlasley@miners.utep.edu

Alan Motta
Computer Science
The University of Texas at El Paso
El Paso, Texas
amotta3@miners.utep.edu

I. ABSTRACT

Social media and social media applications are hotspots for collecting information about individuals. These websites and their component applications are an outlet for people to share their interests, details about themselves, and react to other individual's posts. The nature of these websites brings us closer to the people and business that we know and love. We add our friends online, we follow the artists we like to listen to, and we check in to the places we love to go to. Furthermore, they are great places for users to express their interests, not only through words, but also through pictures. With the given information about user's interests (including their follower's interests), we then make recommendations to users for potential new interests.

With the information provided by users of the social media platforms, we then can compare users' interests in certain areas we choose. The idea is to find common interests in cities, despite their location and way of life. From rural to urban areas, we expect habitants to share things in common, nonetheless.



For the last decade Instagram, Twitter, and Facebook have become some of the most common/used social media platforms

II. STATEMENT OF PROBLEM

Currently, websites help businesses find their target audience, and artists discover their fanbase. However, these are services that you must pay for that make use of people's private information. There are applications that help you discover new people, but these applications are specialized solely for that one purpose and rely on private information, too. There are currently not many methods that help online profiles make these connections using publicly available data, much less any that give the people the power to do this themselves. There is a need for such methods and that also protect user's online privacy

III. OBJECTIVES AND GOAL

Four our project we decided to investigate profile users' interests. We thought it would be interesting to see, and then, compare interests across the following cities:

- New York City, New York
- Denver, Colorado
- San Diego, California

Next, we will retrieve user information:

- Gather user profiles from the cities: We chose cities whose habitants were likely to have interests in things that might not be found in other regions
- Compare profiles to user interests.
- Compare interests between the cities: What interests do these cities share in common?

The topics of interest we chose were films, music, and sports.

IV. OBJECTIVES BREAKDOWN

Gathering Data.

To gather data, the first option was to investigate Instagram, to see how much data and what kind of data could be retrieved. However, the application program interface (API) proved to be highly private and restrictive, as well. Furthermore, the requests rates of the API were low – less than a hundred per minute. Not getting the data that we needed, we turned to Twitter instead. Compared to Instagram, we found Twitter's API to be less restrictive, allowing us to use its multiple search API endpoints to get information such as users, Tweets, topics, and more. Also, requests rates were much higher (hundreds per minute).

Our two main API endpoints where:

- Users/search: It returns the top 1000 results across approximately fifty pages of users.
- Statuses/user_timeline: given a user ID or username, this API returns 200 tweets from their timeline (A *from* or *since* time can optionally be specified, but the api will only return a tweets from the last 3,200 on their timeline.

Then, we used Twurl, which is a cURL-like application, tailored specifically for the Twitter API. Twurl abstracted away the authentication process for API requests. For our purposes, we used the following commands to get JSON encoded strings full of results (parsed by using Python):

- **$ twurl /1.1/users/search.json?q=Denver;**
- **$ twurl **
- **/1.1/statuses/timeline.json?id=user_id;**

You can see the shell scripts that we used to grab multiple pages on our website.

From each city, we grabbed the top 1,000 user accounts and then, from each user we collected 200 tweets.

Cleaning Data.

Once we collected the data we were looking for, we went ahead to "clean it". The first step was to manually remove business accounts. This is due to users/search API endpoint returning public accounts that include unwanted business accounts. Second step was to consider only users whose accounts self identified as living in our cities of interest.

After completing this process, we were left with 260 users from San Diego, 167 users from New York, and 186 users from Denver.

Processing Data with SVM

After the data was cleaned, we started to process it. To predict the topic of tweets we used a linear Support Vector Machien (SVM). Linear SVMs are a form of supervised machine learning. We train the SVM on feature vectors of documents from different topics.

Our SVM categorizes a document into one of 3 categories, Film, Sports, or Music. To train it, we compiled a corpus of 60 documents, 20 per category. From these documents we compiled feature vectors, and calculated their tf-idf scores.

With the linear SVM we were able to achieve a prediction accuracy of 91.666% vs 51% accuracy training with a Naive Bayes model.

V. RESULTS

For each city, we predicted the most probably interest of the each user (out of film, sports, and music) based on an aggregation of all of their tweets that we had collected, and predicted the most probable topic category of each tweet (out of our there categories).

The following illustrations show the number of users from each city in each category.

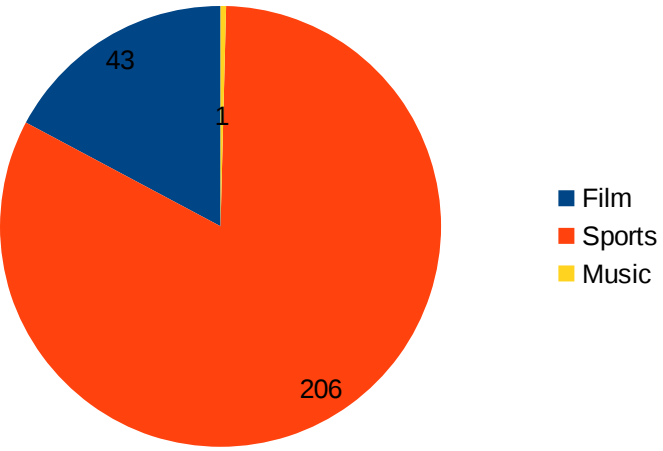


Illustration 1: San Diego Users Categories

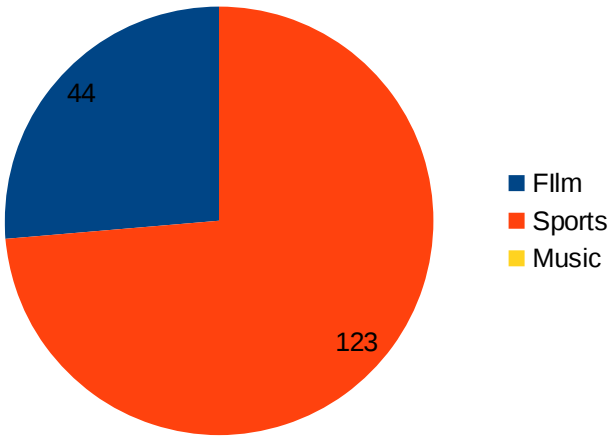


Illustration 2: New York Users Categories

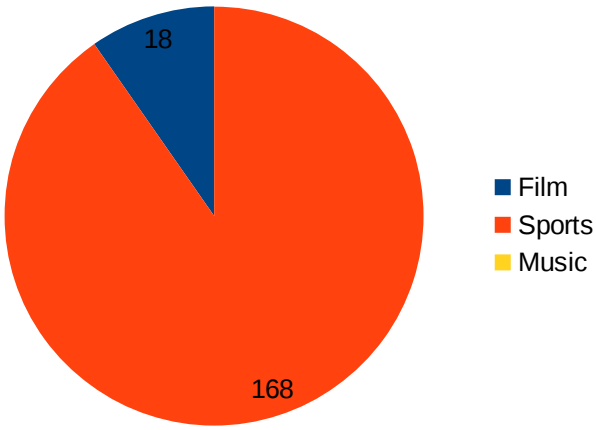


Illustration 3: Denver User Categories

A majority of the users from all three cities main tweet topic was categorized as sports, and only one user from all 3 cities was given a categorization of music. Looking at the categorization of tweets alone paints a slightly different story.

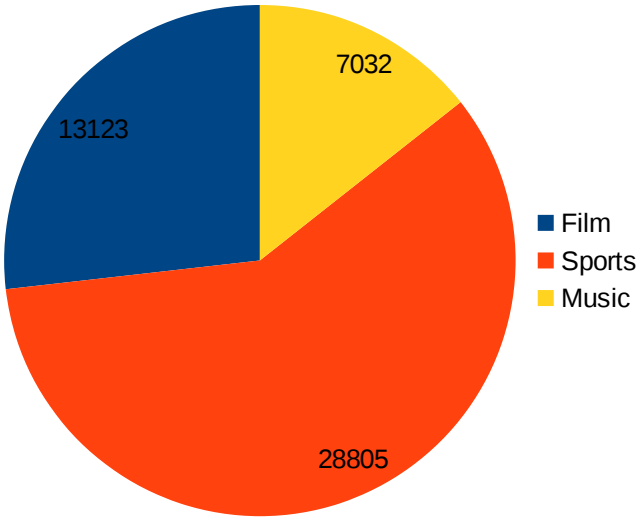


Illustration 4: San Diego Tweet Categories

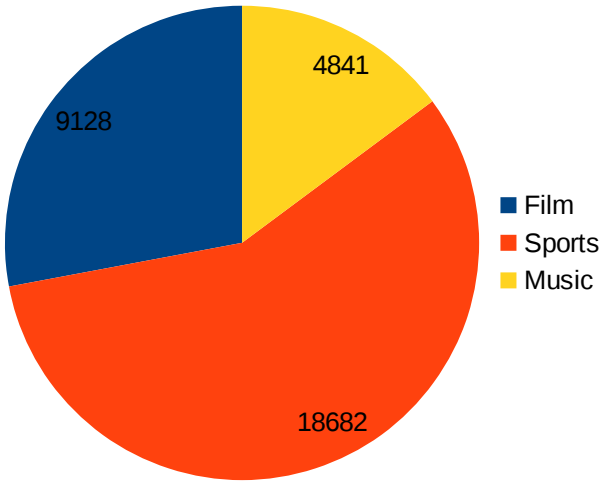


Illustration 5: New York Tweet Categories

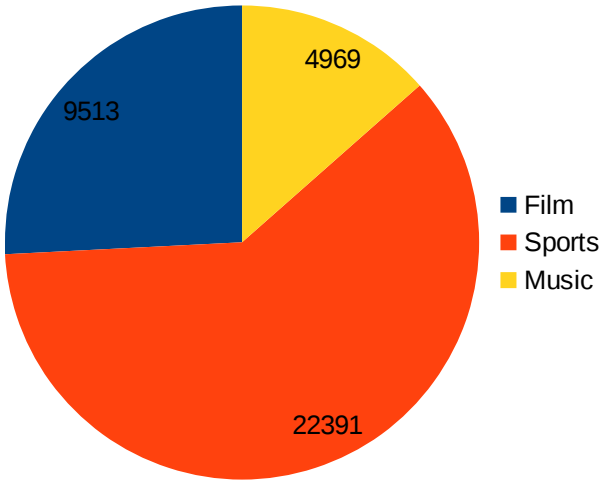


Illustration 6: Denver Tweet Categories

The categorizations of the tweets from each city show that there are individual tweets which were predicted to be about music, but that the users themselves were more likely interested in sports or film.

The three cities show a similar distribution of interest amongst tweets. Approximately 14% of the tweets in each city were classified under music, ~26% film, and ~58% sports.

In each city, sports was the category with the most tweets and users, which is probable given that all three of these are home to major league sports franchises like the Denver Broncos, and the New York Jets, and the San Diego Padres.

VI. PROBLEMS

The difficulty of working with text classification is that the category predictions are restricted to the categories in which you train your SVMs. If you only train an SVM with 3 categories, everything belongs to one of the three categories when there are likely infinite categories. SVMs are better suited to predict demographics like age because there are a discrete number of age groups of equal sizes. So while our users might not have been tweeting about sports, film, or music, their tweets are classified here as such.

VII. WORK BREAKDOWN

Jake's tasks:

+ Data gathering: Investigate social media API's, their restrictions and features.
+ Data cleaning: Noisy data needs to be cleaned by removing business accounts.
+ SVM training: Training the supervised machine learning on feature vectors of a topic

Alan's tasks:

+ Topic Feature vector construction
+ User Feature vector construction

VII. FUTURE WORK

Analysis of text was the first major task to perform on the project. Now that the task was done, the next step would be to add image

analysis to users posts in combination with purely text-based analysis. Pictures are generally more eye catching than mere text. If we manage to analyze both pictures and text, chances are we will find matching results faster than by only using purely text-based analysis.

VIII. REFERENCES/RELATED WORK

[1] How to Find and Reach Your Target Audience on Instagram, https://sproutsocial.com/insights/instagram-target-audience/

[2] 6 WAYS TO CREATE A MORE ENGAGED INSTAGRAM AUDIENCE, https://www.businessinsider.com/Instagram demographics-2013-12

[3] Top Instagram Demographics That Matter to Social Media Marketers, https://blog.hootsuite.com/instagram-demographics/

[4] Twitter user profiling based on text and community mining for market analysis

www.elsevier.com/ locate/knosys

[5] Towards User Personality Profiling from Multiple Social Networks

Kseniya Buraya,1 Aleksandr Farseev,2 Andrey Filchenkov,3 Tat-Seng Chua4

[6] Text Categorization with Support Vector Machines: Learning with Many Relevant Features by Joachims