CIS 400 Senior Project
Spring 2018 Final Report

1. Student Information

List the names and Penn email addresses of all members of the team. Also, please indicate your team or project name.

Team #14: Get2KnowUs
Samantha Caby
Matt Cohen
Vivian Ge
Natasha Narang

2. Advisor Information

*List the name and Penn email address of your CIS faculty advisor. Also, please document **all** the meetings you have had with your advisor this semester.*

Chris Callison-Burch

January 26th:
We talked to CCB about our plans for the semester and what direction we wanted to take our project in. We confirmed what types of analysis we would be doing (performing sentiment analysis on posts containing different product names). He suggested using the LIWK (Linguistic Inquiry and Word Count) program to perform the sentiment analysis. We discussed the next steps we should take in our project to be implemented by our next meeting with him (user classifications)

February 2nd:
We spoke with CCB about how we would be structuring our app going forward and have agreed on a layered analysis approach, where we provide the user with multiple levels of query and analysis customization based on an initial user segment being targeted. The different analysis options we present are still being finalized, but we have narrowed

February 9th:
We demoed how we allow users to decide if posts are classified correctly for CCB. We also showed him some basic analysis on the frequency of self identification phrases within the database in general. We talked more about what APIs we should be using for the sentiment analysis. We also decided to start setting up the different analysis tabs for the following week.

February 16th:
This meeting was cancelled because of scheduling conflicts. However, he requested that we manually classify posts from different groups, create more queries, and train a logistic regression classifier for groups of users using all of their posts before our next meeting with him (2/23).

February 23th:
We set up a feature spec to go over with CCB and finalized the overall direction of our project. We talked more about the different types of analysis we'd be conducting and the way we'd set up the analysis within our web app. We also talked about the features of our classifier and made sure that they were appropriate. Finally, we discussed the analysis we'd be adding to our app - LIWC and log-odds ratios.

March 2nd:
This meeting was cancelled because of Spring Break.

March 16th:
We discussed the different ways of feeding data into our sentiment analysis functions to determine the best way to aggregate and analyze data. We also discussed the progress of our classifiers and the types of analyses we'd be developing for the deep-dive and comparison tabs.

March 23rd:
This meeting was cancelled because CCB was away at a conference.

3. Summary

Describe your project in one sentence.

We developed a web-interface to gather information on previously hard to isolate populations by performing language analytics and generating profiles based on certain self-identified traits.

4. Overview of Problem and Approach

Briefly describe the problem you set out to solve and how you approached solving it.

We developed an interface to provide language analytics and generate profiles of groups of individuals based on certain self-identified traits. Currently, sets of users are segmented based on a few simple traits such as places they have shopped and clearly expressed interests. But, with language analytic tools, there are many more insights to be gained about a group of people based on their text patterns. Our project serves to uncover the nuanced language patterns of groups of individuals in order to better understand them as well as to provide a tool that will classify a random user. This is specifically geared towards helping researchers and marketing teams, understand their subjects and audience better.

Groups of authors are originally segmented based on self-identification phrases provided by the end-user. A user gives a phrase (such as "I'm a dad") and we query a database to find the authors that have used the phrase. Then, based on the all the posts / comments from authors in this group we perform language analysis. We use NLP methods to determine unique / identifying factors about specific segments as well as contrastive analysis between groups. We improved our classifications using Bayesian and SVM classifiers trained on human labeled data.

We believe this is an important problem to solve because it will enable a better understanding of demographically and otherwise segmented groups of people, an understanding that can be used to improve people's lives (one such example is helping individuals quit smoking). Our app is a good solution because it is generalizable and makes it easy to understand the displayed results.

5. Implementation

Describe what you built/implemented in your project, including the system architecture, implementation details for each component, and how components communicate.

The major components of our project are:

Database

Our database is currently being hosted on Google BigQuery. We optimized for speed of queries and ability to link to the database through an API. The data stored in the database is the entire Reddit dataset provided to us by Professor Chris Callison-Burch. It's split into two tables: Reddit_posts and Reddit_comments. We are focusing on the text of the comments generated in response to Reddit posts across all subReddits and, more specifically, using only active users (with at least 10 comments) and filtering out site-identified bots.

Text Classifiers & NLP Analysis

We developed our text classifiers to perform language analytics with Gaussian Naive Bayes and SVM classifiers. The responsibility of these classifiers is to first determine if a given phrase correctly identifies the user segment we are trying to determine (for example, the difference between "*I'm a dad* who…" and "when *I'm a dad*, I will").

There are two types of data used in the text classifiers. The first is the list of all posts in which users self-identify as being part of the user segment we are interested in analyzing. We took this list and sanitized it to produce only the results that definitely self-identify a given Reddit poster. The initial list itself is received from querying the database as described above. The sanitized data from this classifier is then used to retrieve the other posts written by this set of verified users across Reddit (we will again query the database with the names of the specific Reddit posters whose content we are interested in), and perform language analytics on the text we receive. We used the bag-of-words model to actually classify the text written by different authors and determine differences in language trends between user groups. We used Python and multiple NLP and machine learning-focused modules, including scikit-learn's Bayesian, SVM, and testing modules, as well as IBM Bluemix and LIWC to perform further analysis.

Web Application

We built an Express application to allow a user to see our language analytics results for a given population segment the user is interested in. There are three different ways the user can analyze a group.

1. Generalized: The user firsts specifies a phrase that they believe encompasses the population they would like to analyze. The application will then, on the back-end, query the database for

the phrases specified by the user, retrieve the other posts written by those authors across Reddit, classify the text of their posts (using the Python classifiers we've built) in order to determine common word trends, key phrases, sentiment and language analysis, and then return this information to the user in an intuitive visual form (provided by D3.js).

2. Deep Dive: The deep dive is to allow a user to learn more about how a group of authors talks about a specific word and phrase. The application will search through the already retrieved posts and filter for the ones that contain the new word. Again, the application will perform sentiment and language analysis on these texts and display the results to the user.
3. Cross-Group: The user provides two different phrases that will describe two different groups of authors. We again query the database for the posts from these groups and use the comparative log-odds ratio to compare the language used by both groups.

The web application will not perform any calculations or analysis itself, but rather will transfer the data to and between the database and text classifiers. It will first take the phrase input provided by the users and generate a query that is sent to the database. The database will then send those results to the text classifier to sanitize input and after the valid inputs are received and the classifier produces the language analytics, these computational insights will be sent back to the web application in various forms (for the different graphs and visual aids we want to display).

6. Evaluation

Describe how you determined that your solution solved the problem that you have identified.

Our evaluation of our project was composed of two main parts: user surveys and accuracy testing. As we proposed first semester, the three axes of evaluation for our project were **Usability, Speed,** and **Accuracy**. For usability, we sent out user surveys to families and friends. We asked users to try various interesting queries that would be pertinent to their lives and see the results. While we recognize the limitations of this group of individuals, they used our application and provided genuine feedback. In our survey, we asked questions that included whether the UI was intuitive, whether the speed was fast enough, as well as questions regarding perceived accuracy of results. Once we got these results, we looked at the queries that our users searched for and looked into the results they retrieved. This gave us insight into how well our application identified individuals and how well it provided a platform that users would want to use.

With regard to accuracy testing, we used cross-fold validation to estimate the test accuracy of our classifiers across a range of popular self-identification phrases (using hand labeled classification data). We found our medium classifier had an accuracy of 94% (test accuracy) which was above expectations, and that as we increased the number of training samples, test accuracy trended upwards:

Threats to Validity / Assumptions:

- **Validity of self-identification -** Our project is heavily reliant on the assumption that we can use self-identification and key phrases as a thorough starting point for classifying author-groups. While will train classifiers to distinguish between correctly identified authors and those who do not fall in the desired subset, the pool we are pulling from will only include authors who matched the self-identification or key phrases tests. As such, our assumption over valid self identification has two parts:
    - Firstly we are assuming that on average, if people clearly self-identify as belonging to a group, they are telling the truth. Given a phrase such as '%m a dad%', our

classifiers will be used to weed out comments including "when i'm a dad," "thats trustworthy coming from a dad's perspective," or "I'm a daddy, 24 years old, looking for an 18 year old…", all of which are clearly not showing the user is a dad. That said if a user is just lying about being a dad, he will be included in our author-group.

    - Secondly, we are assuming that a sufficient number of individuals within the groups that we segment self-identify. While our app will return information on the number of matches to any given key phrase (so that an end-user is able to improve their search), the only authors that will be considered are those who directly match the phrase. As such, if there are only a small number of matches relative to a large number of Reddit users who actually fall within the category, we will be more reliant on the assumption that it is a representative sample.

- **Accuracy of human labeled, classification data on key groups** - as part of our classification of Reddit users as in or out of the desired self-identifying group, we rely on human labeled data to train the SVMs. As such we are making the assumption both that people who use our app to help classify are doing so honestly, and that said individuals are able to correctly interpret whether or not a given post is representative of the group of interest. We do not believe this to be an unfair assumption because as mentioned above, we are already assuming self identification to be a reasonable means of delineation; the difference here is that we are trusting that our users can differentiate whether or not the identification phrase is representing a different group entirely or if it is qualified by words like 'when', 'until', 'as soon as' or any form of negation.
- **Whether individuals posts on Reddit are representative of their 'true' self** - this threat to validity is largely a result of the way in which we choose to present our results. We are claiming that our app provides an analysis of the user specified groups' language, but in fact we are actually providing an analysis of how user specified groups write on social media (currently only pulling from Reddit). As such, one could make the argument people are less filtered or more aggressive given anonymity, or that they represent themselves differently in any number of ways. That said, there is precedent for using social media posts to analyze individuals' language use.

***Note: With regard to the self-identification concerns / assumptions, there have been multiple peer reviewed papers on the Reddit data set, and others that have used these same assumptions and have been successful (both in number of users identified and in achieving results that agree with other types of studies).*

Legal and Ethical Considerations:

We have no legal requirements for our projects as the Reddit data set is public and the users, while visible by username, have no data attaching directly to their real identities. That said, we understand there are ethical concerns associated with out projects subject, and have considered some implications:

- **Privacy** - throughout our project we will remain mindful and vigilant of how our searching, segmenting, and analysis affects both individuals privacy and 'privacy of groups' (a term associated with possible right of demographically or otherwise defined groups to not be profiled in certain manors).
- **Quantity and Specificity of Data Returned** - while again usernames are not directly linked to real identities, given specific enough searches or even analysis of sparsely used sub-Reddits it may be possible to ascertain information about individuals. We will aim to not facilitate such misuse of the public Reddit data.
- **Offensive Misrepresentation or Oversimplification** - more broadly speaking, we will be cautious of our presentation of results that may negatively profile or offend given demographic segments or groups of authors. It's not our intent to encourage discriminatory actions in any way nor to perpetuate such ideas / cultural assumptions, so while we plan to stay rigorous to the NLP methodology, we will be vigilant in understanding the results we present.

7. Individual Contributions

Describe each team member's primary contributions to the project.
Natasha primarily worked on the Node.Js web application, developing the EC2 instance, and our analysis for comparing multiple groups.
Matt developed the classifiers we used to accurately place a Reddit user into the group they identify with, built word cloud analysis, and helped with the LIWC analysis.
Vivian worked on the visualizations across the webapp (including general CSS); she also worked with the BlueMix API to do sentiment analysis on the first and second tabs of the web page.
Sammi generated suggested queries based on the user input and worked on some of the queries to BigQuery; additionally, she helped build out the three tab layout and did most of the work for the deep dive tab.

NOTE: Although we have listed specific aspects of the project that we each worked on, all of our work sessions were together and we all contributed to all parts of the app.

8. Business Plan

Problem/Need

At the moment, there is not a straightforward way to gather information on people who self identify with a group. Because of this, there is a lack of understanding of how people feel about certain ideas or products. Therefore, it is difficult for marketing firms and researchers to fully understand the people they are selling to or working with.

Value Proposition

Our product will provide users with a way to search for a specific group of people to learn more about how the specified group speaks in general. We also perform analysis to compare different groups of people and how a single group talks about a provided topic or product. With this analysis, a marketing or advertising company can make more informed decisions and researchers will have an efficient way to study groups of people.

Stakeholders

Collaborators: We use the services of Amazon, Google BigQuery, and IBM but we have no direct collaborators. We will have a direct to consumer strategy and no intermediate organizations.
Consumers: Our consumers would be marketing firms and everyday individuals that would likely use the free service of our product as well as researchers that would use the paid version of our product.
Competitors: This is currently a nonexistent market so we cannot tell who our competitors would be. Marketing firms that provide research for advertising agencies is a potential competitor but our product offers a very different service.

Market Opportunity

The market we would like to target is very broad. Essentially, we created a project that allows anyone to analyze the reddit posts of a group of people. This type of service, one that allows an average person to retrieve text analyses, does not exist currently so this is a completely nonexistent market. The difficulty of creating a market is that there is likely very little demand as is meaning we would have to show potential customers the benefits of our platform. A key advantage though would be that we are the first movers into this market segment.

Customer Segments (& Growth Estimates)

***Note**: Our target market is solely composed of US based institutions currently. This is due to our source data being entirely from English speaking, primarily US based individuals. With regard to both of our primary customer segments, we estimates 10% of the total customer bases will make use of our platform in some capacity as there are not many tools that provide a similar service. The percentage of those that we will retain as subscribers is discussed in more depth in the Cost and Revenue section.*

Advertising/Marketing Firms

There are 14,000 such establishments in the United States with expenditures exceeding 180 billion in 2017. With regard to the industries growth, different media platforms have been experiencing vastly different rates of change, but the industry as a whole has consistently seen revenue growth of 2.1-2.3 percent; the same rates are forecast to be consistent through 2020. We anticipate these firms to use the subscription section of Get2KnowUs to better understand specific customer segments that were previously hard to isolate. Such information will be of use in both reaching and engaging their target markets.

Researchers

As we are in the introductory phase we anticipate that half of the top 100 NLP programs will try Get2KnowUs, and that within each of these programs, an average of 3 professors / grad students will engage. As discussed in other sections, finding well formatted raw data as well as the additional analytic features we generate is a difficult task and a bottleneck for many doing NLP research. As there is no existing tool that provides these, Get2KnowUs has unique appeal to such potential users.

Market Research

To gain some insight into whether or not our product would be useful, we allowed many different people to use our product and provide feedback as to what they liked, the effectiveness of the results, and features that they would like to be incorporated.

Competition

While NLP has been heavily researched in the last several years, there is virtually no platform for the average person who is not a researcher to analyze texts by certain groups of people. Because of this, our product offers a unique solution for both advertising agencies as well as researchers to gain key insights into different demographics. Through both our research as well as surveys collected, there seem to be no real competition as this is an untapped service market.

Marketing researchers are not direction competition but we anticipate advertising agencies to use these in conjunction with analyses provided by marketing firms or as substitutes, depending on the advertising agency. Our product provides a new lens of analysis for groups of people while advertising agencies will conduct analyses such as surveys and focus groups thus we provide very different information and are not considered competitors.

Cost/Revenue

As this is a software product, the only tangible costs of our product are the serverside costs of maintaining a running instance. We are currently using Amazon EC2 to support our project and expect the upfront costs per year of our t2.micro instance to be $118. This should support tens of thousand requests per year and, of course, can be easily scaled up at a similar cost scale once our product picks up over time.

In terms of the revenue model, our product will follow a freemium model. Get2KnowUS provides three types of analytics - general (which is designed for users without any experience in NLP who are looking for generalized information about a given user segment), deep-dive (which provides LIWC calculations that can be used by NLP experts to further their research), and comparative analytics (which conducts an IPLO analysis to provide detailed statistics on cross-group comparisons). We will provide the first, general analytics tab for free, and provide the remaining two tabs as part of a subscription model for $12.99/month. For subscribed users, we will also offer the ability to plug in our platform directly to their current NLP projects and provide beta access to our newest features.

Looking at the standard statistic of 3% of freemium product users converting to paid subscriptions, we can expect a similar rate for our product as well. From the Customer Segments section above, we anticipate about 1,400 marketing firms to request to use our product as subscribers. This will generate $218,232 in revenue. We also expect about 150 NLP research contracts, which will generate an additional

$23,382 in revenue, giving a total of $241,614 in expected revenue in Year 1. Taking out our costs as calculated above, we get a net profit of $241,496 in our first year.