LEVERAGING VISUAL ANALYTICS FOR MODELING ONLINE USER BEHAVIOR ON SOCIAL MEDIA

by

Omar ElTayeby

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computer Science

Charlotte

2020

Approved by:

_____
Dr. Wenwen Dou

_____
Dr. Jing Yang

_____
Dr. Sara Levens

_____
Dr. Tiffany Gallicano

_____
Dr. Samira Shaikh

ABSTRACT

OMAR ELTAYEBY. Leveraging Visual Analytics for Modeling Online User Behavior on Social Media. (Under the direction of DR. WENWEN DOU)

Analysts and domain experts in various fields rely on collecting data about their subjects to understand and predict their behavior. Characterizing and modeling human behavior requires analyzing extensive amounts of data from heterogeneous sources, which is a challenging task for researchers to achieve when using traditional methods. Social media platforms have been used in social sciences and different industries to understand their subjects in online settings. The advantage of the online setting is the ease of accessing large amounts of data, which solves the problem of data availability that occurs in offline settings. However, the data collected from social media is often messy and noisy.

Therefore, many visual analytics (VA) tools are built for assisting domain experts to overcome those challenges efficiently. In this dissertation, I show how VA systems can leverage data to improve two major types of analysis tasks, which enhance discovering users' behavior on social media. Both analysis tasks are related to the process of inferring the user categories, which are predefined by the domain expert. I illustrate the usability of VA for enhancing these tasks by applying the same research questions on different applications. The first analysis task involves understanding the connection between the social media user's behavior and demographics. The second task involves the labeling of the social media users themselves according to the expert's observations of their behavior. The VA systems characterize the users' behavior through a suite of

multiple coordinated views coupled with predictive models. These models are based on the textual information derived from their posts.

The first application, DemographicVis, supports the understanding of the connection between the user's demographic information and user-generated content. My approach in this application allows domain experts to make sense of the connection between categorical data, which is the users' demographics, and textual data, which is their posts. This connection shows the characteristics of different demographic groups in a transparent and exploratory manner. Users' posts are utilized to model and comprehend the demographic groups with the features that best characterize each group. The interactive interface of DemographicVis also enables the exploration of the predictive power of various features. In the second application, I propose a VA system for domain experts to categorize and label Twitter users. This work was motivated to eliminate bots from social media datasets since they produce noise that impedes the analysis. I address this challenge by providing an interface that enables the communications experts to separate between bots and other types of users in an active learning setting. In this setting, the experts iterate between labeling the users and running predictive models, based on these labels, to enhance their decisions in future labeling rounds.

ACKNOWLEDGMENTS

First, I would like to thank my advisor Professor Wenwen Dou and my committee members Professors Jing Yang, Sarah Levens, Tiffany Gallicano, and Samira Shaikh for their guidance and support during my Ph.D. program.

Also, I would like to express my gratitude to my parents, sisters and extended family for their support throughout my whole academic journey to this day. They have been my source of support, guiding me with their wisdom and life experiences. They appreciated the value of education and always saw it as a life long investment. My parents tutored me and sent me to private tutoring in school to ensure that I had the best education. They never refused a request when it came to benefiting my career path.

My academic journey didn't start with my Ph.D., it started when I discovered my passion for science. There were many teachers who inspired me in subjects like Algebra, Statistics, Physics, Control Systems, Communications Systems, and Machine Learning. I want to express my gratitude to those teachers: Ms. Zahraa, Mr. Mounier Adeeb, Professors Hassan Elkamchouchi, Ahmed Sultan, Sherif Rabea, Peter Molnar, Richard Souvenir. Also, I want to express my gratitude to my internship mentors Dwayne John and Professor Lixia Yao for the great opportunities and their support during and after my internships. Last but not least, I want to thank all of my friends, colleagues, classmates, lab-mates, and coworkers who helped me learn something new everyday.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Social media has been a great platform for studying populations; researchers used many different technologies to characterize users and their activities online. The most popular platforms like Facebook, Twitter, Reddit, and Instagram provide their users with different options to express themselves. Heterogeneous large amounts of data available has created research opportunities to study social media users and characterize their behavior; however, with these opportunities comes challenges. Some of those challenges directly relate to the characterization tasks that the researchers try to achieve, and some are bottlenecks that make the data unreliable to analyze. Quite often these bottlenecks are either overwhelming noisy data from undesired sources or the collected data has missing or incompatible values. Visual analytics (VA) techniques offer great solutions for domain experts to analyze the data in depth and come up with stories about social media users. One of the great advantages of the VA solutions is providing domain experts with the ability to switch between summaries and details in a blink of an eye. There are many techniques that enable this capability, making the analysis faster and more efficient than manual methods.

Thesis statement: Custom VA systems are needed to support domain experts in characterizing and labeling users on social media through organized content analysis on their online behavior. Iterative interaction with predictive models' results on multiple linked visualizations improves the experts' performance over traditional tools.

In this dissertation I study how VA systems can be applied to improve the exploration process of users' behavior on social media for domain experts. This study provides VA methodologies to address important research questions asked by domain experts, which vary according to the context of each application. I present two sets of research questions, either drawn from literature or from interviews, to demonstrate how VA systems enhance analysis tasks discussed in each chapter. Using these research questions, I demonstrate how the proposed VA systems predict and cluster user categories and help contrast the linguistic behavior with their topics of interests.

The process of developing these tools in both chapters follow the same steps; Figure 1 shows the framework of steps that helped answering the research questions. Firstly, from the research questions I was able to derive the task requirements and collect the data needed for the analysis. Secondly, I extracted the important features to build the predictive models. These predictive models are meant to satisfy the analytics capability of the tool, while the interface helps the domain experts explore the data efficiently. I used the task requirements with the models' outcomes to design the interface, which is mostly composed of multiple visualization techniques connected together through interactions. Last but not least, I have conducted use cases and user studies to evaluate the design and the tool's efficiency after implementation.

1.1 Research Questions

The VA systems proposed in both chapters are designed to help domain experts study social media users by showing aggregate summaries of user categories and individual users' information. This combination gives the experts an overview of the

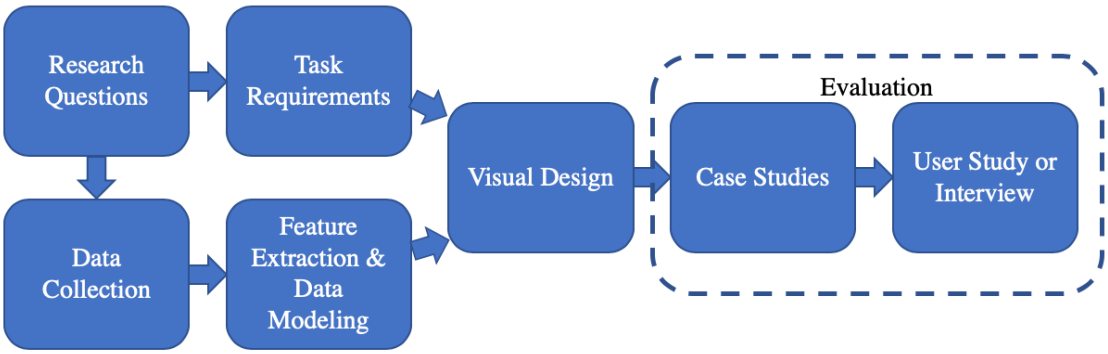

Figure 1: Framework showing the process of developing the VA tools in both chapters.

data while being able to delve into details to answer the research questions below. The research questions are divided into two sets, which correspond to the analytics and visualization components of the tool:

- Analytics component (A1 - A3)

    1. Given the predefined user categories, what is the feasibility to predict them with little information about them on social media? How accurate can the prediction models get?
    2. What are the available features that can achieve this prediction? What is the prediction power of these features? How accurate and important is each feature to the prediction model?
    3. How can awareness be raised to use linguistic, sentiment, topics of interests and other metadata as predictive features to separate different user categories?

- Visualization component (V1 - V3)

1. How can the computational features be presented visually to domain experts without burdening them with technical details?
2. How can user cohorts and their interests be visually represented to the domain experts? How can the similarities and differences between the users on social media be shown?
3. How can the users' posts visually be aggregated to support comparisons between the different user categories?

Using the research questions corresponding to the analytics component, I demonstrate the capability of the VA system to show the prediction power of different features extracted from users' posts and find the linguistic differences (A1 - A3). In addition, the tools' interfaces provide a suite of visualization techniques for domain experts to explore users' behavior, which addresses the last three research questions (V1 - V3). These research questions focus on summarizing the users' interests and representing the information visually in an easy way for domain experts to have better interpretations about the different categories.

The research questions are answered in each chapter according to the context and domain, which the tool was designed for. I address the research questions regarding the analytics component (A1 - A3) for chapters 3 and 4 in sections 3.3 and 4.6.1 (Data Modeling subsection), respectively. The research questions regarding the visualization component (V1 - V3) for chapters 3 and 4 are answered in sections 3.4 and 4.6.1 (Visual Interface subsection), respectively.

1.2 Motivations, Domains & Analysis Tasks

Despite the fact that both chapters are about two different applications, they follow the same structure and development framework shown in figure 1. Mainly, the reason is that both chapters have similar motivations. In this section I present the motivation behind developing both tools, which address challenges for domain experts in the field of business development, marketing, customer relationship management, market research industries and social science studies. I present three major challenges:

1. Traditional data collection methods for social science and business domains are inefficient and time consuming, and social media data is presented as an alternative solution of these methods [36].

2. After the social media data collection stage, the data is usually hard to be used directly for analysis purposes, and it needs to be preprocessed and visualized adequately for domain experts [56].

3. The lack of metadata and expert-defined labels, which help domain experts characterize and understand the users.

Marketing professionals and customer relationship management in many businesses find it very tedious to collect data about their customers for getting feedback about their products and services [41]. This difficulty mainly roots from the unfamiliarity of their customers with the tools and forms used to collect data about them, which creates the bottleneck of low response rates to market study. Along the same type of challenges, psychologists and communications experts consume a lot of time to

collect data about their subjects and sometimes find that user studies prime their subjects towards an artificial environment that skew their results. With the rise of social media platforms, many people created accounts to communicate online about their social lives and experiences [30, 33]. Although, social media has provided the domain experts with more data about their subjects, the data is usually clunky and messy for analysis purposes [56]. In addition, the lack of user categories and types defined by the experts among the sampled users impedes their analysis in making connections between the users' behavior and the cohort that they belong to.

The motivation behind the first application is to provide a tool for business developers, marketers, customer relationship managers, market researchers to understand the connection between social media users' demographics and their online social behavior. The user's behavior is characterized by a suite of textual information derived from their posts on Reddit; for instance their topics of interests, linguistic features and other peripheral information derived from their posts. The demographic information was collected through an online survey as the users' cohort. Then the sampled user's posts were collected to connect their topics with their demographics, which helped answer the research questions in chapter 3 section 3.1.

The second application addresses the need of a labeling tool for domain experts such as psychology, communications studies, political science scholars. The VA system enhances the task of labeling users with expert predefined types. The experts use the system to divide and filter the tweets in order to search for certain user behaviors, and then differentiate the users from each other by labeling them. The system also provides predictions based on the expert's labeling in active learning setting, which

addressed the research questions in chapter 4 section 4.2. In addition, the prevalence of different user types on the topics is shown to the expert in order to help them figure out which topics have been infiltrated with bots.

Another major difference between both tools is the analysis task performed by the domain experts. In the first application the information about the expert-defined cohort is collected from the survey, and the main analysis task performed by the experts is connecting between the user cohort and their extracted textual information. In contrast with the second application, the domain experts label the users according to the extracted textual information rather being collected as part of the metadata. Figure 2 summarizes the difference between analysis tasks in the two applications; however, they fall along the same pattern of tasks that leverage VA to either make connection between the expert-defined cohort and extracted textual information or label the users by making that connection.

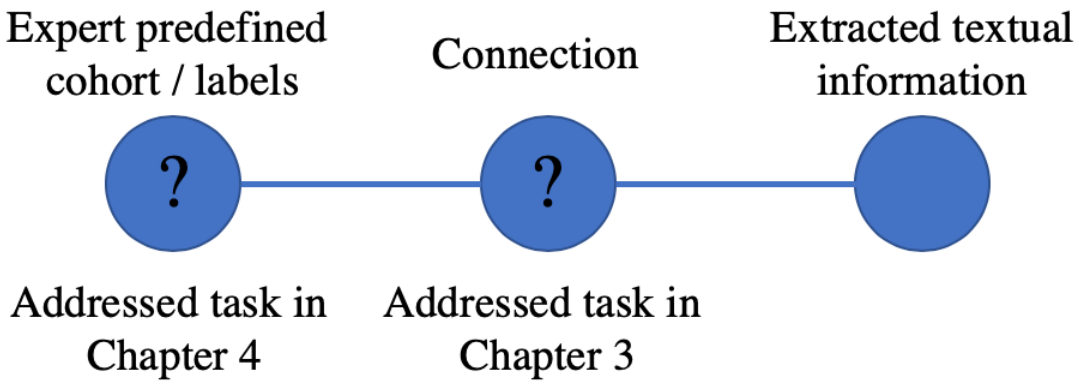

Figure 2: The missing part of the analysis task that the VA system attempts to solve for the domain expert in chapters 3 and 4.

1.3 Thesis contribution

This new kind of understanding of users on social media, which is deepened into the characterization of users' identity and interests, is nuance from simple analysis tasks; it requires much complex modeling and systems that explain the differences between the user categories and types. Because people are complex it's hard to understand their behavior and mostly hard to identify their characteristics as social media users; the interactive VA systems proposed play a vital role in enhancing the exploration process to answer the research questions mentioned, and are more efficient than traditional non-interactive methods. Particularly, these VA systems enhance the domain experts' exploration tasks in:

- Characterizing and linking between user categories and their behavior, which could be derived from their posts like topics of interests and linguistic behavior. This is the main analysis task in chapter 3 when the users' cohort information is available.

- Categorizing and labeling users according to their posting behavior on social media in active learning setting, which is the main analysis task in chapter 4.

- Assessing the predictive power of features extracted from users' posts, which are used to infer and predict the categories that they belong to. This contribution is in the form of answering the research questions A1 - A3 and V1.

- Visually comparing between the user categories from the perspective of their topics of interests and linguistic behavior. This contribution is in the form of

answering the research questions A3, V2 and V3.

- Performing content analysis from both overview and detailed, views to support the previously mentioned tasks with less cognitive effort by the domain expert.

My main argument is that tailored designs that fit required analysis tasks are much more efficient than using off shelf-tools or manual methods to address such complex research questions. I substantiate this argument through user studies and interviews with the domain experts to get insights on how useful are my proposed tools compared to other traditional methods. Also, the aim of these user studies is to provide feedback on how to improve the tool for future work. The methodology that I adopt in both chapters is divided into two steps: the first converts the research questions into analysis tasks, and the second transforms these analysis tasks into VA tasks, which may involve interactions in coordinated multi-view visualizations. The VA tasks shape the design choices of the visualization techniques in addition to the appropriate coordination needed between multiple views.

The user cohort studied is quite unique but has subtle overlapping behavior, nonetheless, there are always exceptions when applying generalizations. For example, in the chapter 3 I show case studies that compare the interests of particular demographic groups. While in the second chapter, I present the intuitions that the communications experts have reached in order to set criteria for differentiating between the four types of users in the dataset. In addition, it's possible to create several classification categories of users from different perspectives. Despite that the prediction models can reach a satisfying level of accuracy to differentiate the users, there is always

room for improvement. For instance, in the third chapter I show the different levels of accuracies and contributions of the different features used to predict the users' demographics.

CHAPTER 2: LITERATURE REVIEW

The main differences between the two main chapters is the analysis task, and therefore, I review two bodies of related work that correspond to each chapter from the perspective of these tasks. The first section is related to connecting demographics to linguistic features addressing the literature for the chapter's domain. The second section is related to the labeling task for domain experts using VA systems which addresses the chapter's analysis task, and the characterization of bots' behavior on social media for the domain aspect of the chapter. Last but not least, text classification models for both main chapters is addressed in section 2.3

2.1 Characterizing demographics on social media

Demographics analysis has been an important domain, where researchers from social sciences gained psychological insights through studies that link language use with age and gender, while researchers from computer science have focused on introducing and improving algorithms to predict demographic information. In the next subsection I review related work to chapter 3.

2.1.1 Linguistic analysis on age and gender

The typical approach of correlating age and gender with language use involves counting word usage over a priori word-categories. The most commonly used word-category lexicon is the Linguistic Inquiry and Word Count (LIWC) dictionary. Several

studies have leveraged LIWC and focused on function words to study age and gender. Research by Chung et al. [19] and Argomon et al. [8] on gender analysis found that males use more articles, while females use more first-person singular pronouns. Also focusing on examining function words, Newman et al. reported several findings [43], including women use more certainty words while men tend to have greater use of numbers, articles, long words, and swearing.

As of age, through linking language use and aging, Pennebaker et al. [44] found that with increasing age, individuals use more positive and fewer negative affect words, use fewer self-references, more future-tense and fewer past-tense verbs. In the context of blogging, Schler et al. [48] identified a clear pattern of differences in content and style: regardless of gender, writing style grows increasingly "male" with age: pronouns and assent/negation become scarcer, while prepositions and determiners become more frequent.

In chapter 3, the VA approach is complementary to the linguistic studies. Through coupling the semantically meaningful topics and relationships between demographic groups identified in my approach, with the general patterns identified by linguistic studies, higher order thought patterns can be revealed and outcomes can be solidified and become more interpretable.

2.2 Characterizing user types and labeling tasks

Along the line of research questions discussed in section 1.1, I am presenting related work for chapter 4. The VA system in chapter 4 provides domain experts with aggregate and detailed views to differentiate different types of users. Since the benefit

of this tool is not only labeling bots but also differentiating them from other persuasive users, I review tools which facilitate the labeling task.

2.2.1 Labeling task in an active learning setting

Labeling tasks for domain experts can be quite tedious and time consuming when it comes to large datasets. The coupling of machine learning and visualizations optimizes the search of the desired user behaviors. In this context, there have been *many studies* that incorporated semi-supervised machine learning and active learning into VA systems to achieve the desired efficiency [50]. The active learning setting in chapter 4 is unlike other related work, where pretrained models are ready to show predictions to the tool users before they start labeling.

Some of the most advanced tools in the field of active learning are ReGroup [5], Basu et al.'s [10] system, Inter-Active Learning by Höferlin et al. [32] and CHISSL [7, 6]. In terms of the machine learning algorithms at the backend that support the classification task, ReGroup uses Naïve Bayes classifier to help social media users group filter their networks based on a number of distinct account features. Basu et al.'s [10] system uses nearest-neighbor approach for textual data; they applied logistic regression classifier to recommend grouping of documents for the domain experts to label them. Inter-Active Learning provides a VA system for experts to label videos for ad-hoc training, where the examples are displayed in clusters. CHISSL can handle different data types, where different classification algorithms are used for each data type.

There is one common aspect in the methodology applied in my system and CHISSL.

Both systems prioritize the data points by showing the most adversarial ones to be labeled first; however, the approach towards implementing it is different. In my VA system, the domain experts search for the persuasive users using topics and the density of their proportions. In CHISSL, the system focuses on presenting the most borderline classified and unclassified data points for their users to correct them, in contrast with ReGroup, they prioritize the most likely accurate and certain ones. Another difference is that my VA system is focused on social media labeling, while CHISSL can be applied on any type of documents.

2.2.2 Bots & users' online behavior

The field of natural language generation (NLG) is advancing everyday which has social media users to become more susceptible to bots [58, 59]; they are able to mimic real users efficiently. This continuous advancement requires automatic bot detection techniques to catch up with the technology. Providing a VA tool for communications experts to label bots and understand their characteristics can make substantial improvements towards identifying the important features for bot detection. In chapter 4 I aim to address the research questions mentioned in chapter 1 by building a machine learning model to investigate the feasibility of predicting the expert's labeling.

Social-bots are software robots that interact with real users on social media, they could be helpful to the social media users by automating online tasks like information retrieval or spreading important news to the public in emergency response situations. However, not all bots are useful, some are intended to harm the users on social media,

for example the spreading of rumors, misinformation, or even malware to infiltration users' privacy [27]. Bots can have dangerous effects on society due to their efficiency in growing and influencing users politically; Abokhodair et al. [1] studies the growth of specific social botnet on Twitter to understand their behavior. Alvisi et al. [4] surveyed the sybil defense that uses social graph as their detection strategy.

In the direction of studies on bot behavior on social media, other researchers studied the users' behavior towards these bots. There has been two approaches towards this characterization problem: the first attempts to create bots and observe their effects on real human users, and the second is detecting the existing bots, then analyzing their effects. Interestingly, many researchers have adopted the method of creating their own bots to discover how the users will respond to them, instead of detecting the bots that already exist to eliminate the chance of error [2, 14, 13]. Wald et al. [59, 60] predicted the susceptibility of regular users to bots. Wilkie et al. [62] characterized three aspects of the bots on Twitter. Dickerson et al. [22] used sentiment analysis to detect bots and then compared their sentiment with human's sentiment on Twitter.

Because of the harmful influence of these bots on social media analysis, researchers have focused their work on detecting them using machine learning techniques. For example, Cao et al. [16] developed SynchroTrap that was able to uncover millions of malicious accounts on Facebook and Instagram using top-notch parallel processing technology. Botometer [1] is an online tool that gives an indication of how much probability a Twitter account could be bot or human operated. The tool is provided in a friendly format through a website and as API for super users. The developer's

[1] https://botometer.iuni.iu.edu/

work was a continuation of their momentum from their contribution at the DARPA Twitter bot challenge. In the developers' paper "BotOrNot" they specified the features that they use in their detection for high accuracies, such as network, user, friends, temporal and content features.

As mentioned earlier, one of the main tasks of social bots is to spread misinformation across social media platforms. They are not the main actuator of misinformation, but they make the problem worse. Thus, detecting misinformation is one of the indicators for bots' behavior, and it is equally important to detecting bots themselves. Hoaxy is a platform that facilitates the analysis of misinformation on social networks and engages in fact checking efforts [51]. The tool is built upon crawling fact-checking assessment sites such as the hoax Facebook page and Snopes.com. Another VA system built for fact-checking news is Verif [38, 37], where the main research goal is to study how confirmation bias and uncertainty could impact the decision-making process of experts in assessing news outlets. Verifi is also considered as one of the VA systems that experimented with visualizing social media data for experts to the labeling task. Their user study required the subjects to label whether the user account as trustworthy or not based on misinformation.

2.3 Comparison between text classification algorithms

The decisions made in section 3.3 in chapter 3 and 4.6.1 in chapter 4 are mainly based on the literature reviewed in this subsection which focuses on selecting the most suitable text classification algorithm. Every classification model has its pros and cons [29] depending on the available resources from computational power, and

time for training the model. When deciding on the machine learning algorithm for the text classification problem, there are three levels of comparisons that we can make: (1) comparison between the algorithms for the classification task in general from previous empirical research, (2) comparison between the algorithms specifically for text classification from previous empirical research, (3) empirical comparison between the algorithms on the dataset in-hand, which the algorithm would be applied to. For the purpose of the work done, the first two levels of comparisons are enough. There is empirical evidence that Random Forest (RF) and Gradient Boosting tree (GBtree) are outstanding competitors amongst other algorithms for text classification for the following reasons.

I review several works that benchmarked classification algorithms on various datasets. Caruana and Niculescu-Mizil [17] made an empirical comparison between 10 algorithms by measuring their performances on classifying the target variables of 11 datasets. In table 4 of this paper [17] they were able to make a bootstrap analysis of the overall rank by mean performance across problems and metrics. The ten algorithms were ranked in the following order: Boosted Trees (BST-DT), RF, Bagged trees (BAG-DT), Support Vector Machine (SVM), Artificial Neural Networks (ANN), k-Nearest Neighbor (KNN), boosted stumps (BST-STMP), Decision Tree (DT), Logistic Regression (LOGREG), Naive Bayes (NB). And thus, the top three ranked algorithms are based on boosted and/or bagged decision trees. In another empirical evaluation paper by Fernández-Delgado et al. [26] found that the parallel random forest achieved the best results among 179 classifiers, which belong to 17 classification algorithm families. They experimented with these algorithms on the

UCI classification problems database. Also, Statnikov et al. [55] found that RF outperforms SVM on microarray-based cancer classification problems.

There are many aspects upon which the best algorithm can be chosen. These aspects depend on the type of problem the algorithm is being applied to. **Complexity**: SVM, RF and BAG-DT are non-parametric models, which are computationally expensive to train when compared to linear models. However, linear models are not as accurate as non-linear models (i.e. RF, BAG-DT, SVM). On the other hand, when comparing between non-linear algorithms RF is faster to train than training BAG-DT and SVM models, especially when RF is trained in parallel [26]. **Multiclass classification**: While RF and BAG-DT are suited for multiclass classification tasks, SVM is more suited for binary classification tasks. Thus, fitting SVM to multiclass classification tasks requires tweaking of the implementation by focusing the model to classify between one of the labels and the rest (one-versus-all) or between every pair of classes (one-versus-one). **Hyperparameters tuning**: For SVM, many parameters need to be tuned to get high accuracy performance such as kernel, regularization penalties, the slack variable, etc... On the other hand, the number of trees and number of randomly selected variables per tree are the only two parameters that need to be adjusted for high performance. **Feature types**: Since SVM's classification depends on the distance between the points in the vector space, one-hot encoding is must for categorical features. Thus, RF models are much easier to deal with when data has categorical features. **Scaling**: SVM requires scaling and centring to the input before feeding it to the algorithm, however, RF do no these kind of transformations to perform well. **Output format**: RF can either be used as a classifier or as a regressor,

however, if we would use SVM as a regressor distances, to the boundary, need to be converted to probabilities.

In addition, when taking Deep Learning (DL) models we have to take the number of training points needed for a good performing model. DL models need large datasets and computing resources in order to produce a model with high accuracy. For BAG-DT models, the individual greedy CART decision trees are problematic. All of the trees produce highly correlated predictions, and thus, combining the ensemble of these predictions leads to high variance. RF solves this problem by combining the ensemble of predictions from weakly correlated models. The trees in RFs are constructed with a number of randomized variables. However, RF can return unreasonable predictions for inputs out of the range. This issue is addressed in chapter 4, which is related to the research questions A1, V1 and V3, where the active learning setting provides the experts the chance to evaluate the regression results.

Computer scientists have also used linguistic features to build and improve models that predict age or gender. Examining information from social media users, Burger et al. [15] experimented with Support Vector Machines, Naive Bayes, and Balanced Winnow2 [39] to build classifiers to predict gender. Descriptions for Twitter user such as screen name and full name are used in addition to tweets to improve the accuracy of the classifier. Rao et al. [45] introduced stacked-SVM-based classification algorithms over a set of features to classify gender, age, regional origin and political orientation, while Schler et al. [48] leveraged style-based and content-based features to classify age and gender for thousands of bloggers.

Comparing to the linguistic analysis research, the above mentioned classification

approaches focus more on predicting age and gender, and less on gaining psychological insights from analyzing the language use of different demographic groups. As a result, interpretable results that distinguish demographics groups are difficult to obtain from the classification models. This challenge is related to research questions A3 and V1 - V3. In chapter 3 I address this challenge by visually connecting between the classification textual features used for prediction with the distinct demographic groups.

In chapter 3, I use the GBtree since the text classification task is pretty well defined for that chapter and the user cohorts are mainly about predicting the demographics. On the other hand, chapter 4 is meant for developing a tool that accepts experts' predefined user types to predict. RF in this case is a better option than GBtree, since GBtree requires more tuning than RF.

CHAPTER 3: DEMOGRAPHICVIS: ANALYZING DEMOGRAPHIC INFORMATION BASED ON USER GENERATED CONTENT

''Today you are You, that is truer than true. There is no one alive who is Youer than You.''

-- Dr. Seuss

In this chapter I show how a VA system, DemographicVis, that can support the task of inferring and making connections between textual corpora and their categorical labels. The aim is to model and infer demographics of users on social media by connecting demographic information with user-generated content and features that distinguish them, along with topical and linguistic features. The tool allows users to understand the characteristics of different demographic groups in a transparent and exploratory manner. This chapter also discusses the prediction of demographic factors, such as topical, linguistic, and peripheral features, which are extracted from both user-generated content and metadata. This chapter is based on my paper published in the VIS conference [2] in 2015 [23].

I transform this characterization task into three specific visual analytics tasks. The first task is to identify the different demographic groups and find the differences between their distinct topics of interest, and that was achieved using the parallel set with the word cloud by connecting demographic groups to topic features. The

[2] http://ieeevis.org/year/2015/info/vis-welcome/welcome

second task is to find groups which have similar interests using user cluster view. And the third task is studying the feasibility of classifying online users into different demographic groups based on the derived features discussed, where little ground truth is available, through the feature ranking view.

3.1 Introduction

The wide spread nature of social media provides unprecedented sources of written language that can be used to model and infer online demographics. In this chapter, I introduced a novel visual text analytics system, DemographicVis, to aid interactive analysis of such demographic information based on user-generated content. my approach connects categorical data (demographic information) with textual data, allowing users to understand the characteristics of different demographic groups in a transparent and exploratory manner. The modeling and visualization are based on ground truth demographic information collected via a survey conducted on Reddit.com. Detailed user information is taken into my modeling process that connects the demographic groups with features that best describe the distinguishing characteristics of each group. Topical and linguistic features are generated from the user-generated contents. Such features are then analyzed and ranked based on their ability to predict the users' demographic information. To enable interactive demographic analysis, I introduce a web-based visual interface that presents the relationship of the demographic groups, their topic interests, as well as the predictive power of various features. I present multiple case studies to showcase the utility of my visual analytics approach in exploring and understanding the interests of different demographic groups. I also

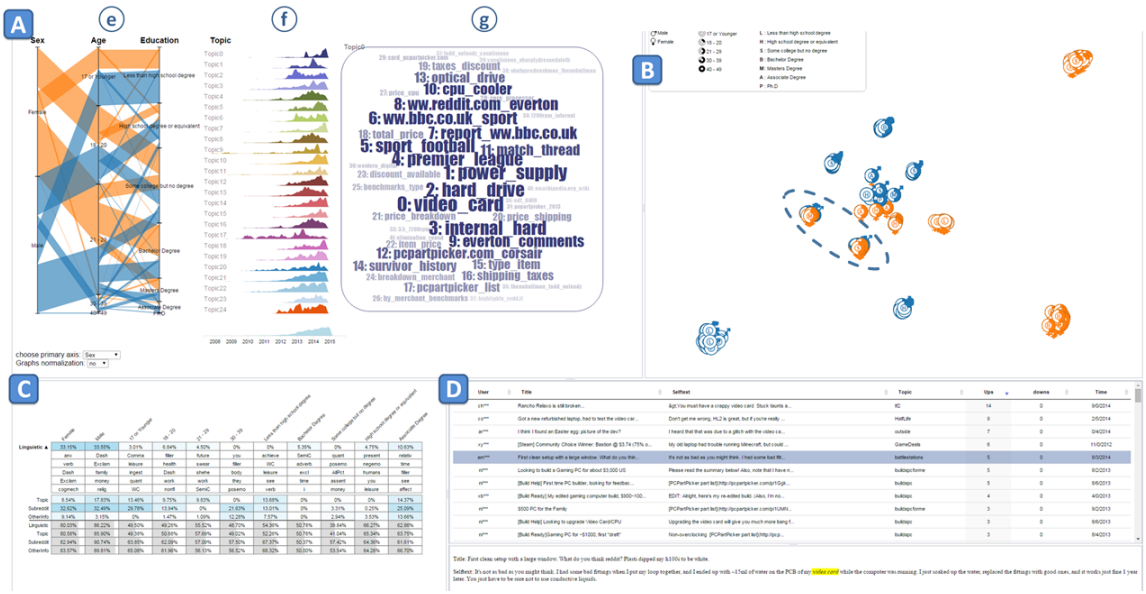

Figure 3: DemographicVis interface: A) Parallel Sets with word cloud view that connects demographic groups to user generated content, B) user cluster view that groups users based on topic interests, C) feature ranking view that presents the predicative power of various features, D) posts view showing details on demand.

report results from a comparative evaluation, showing that the DemographicVis is quantitatively superior or competitive and subjectively preferred when compared to a commercial text analysis tool.

Demographic analysis provides valuable insights on social, economic, and behavioral issues. On a macro level, analyzing demographic information sheds light on a range of future economic factors from gross domestic product growth and inflation to interest rates [25]. On a micro level, demographic analysis yields valuable information on businesses, communities, and other aspects that are closely related to our daily lives. From a business-oriented point of view, understanding demographics is important for business development, marketing, and customer relationship management. Businesses market products or services through targeted approaches to different segments of the population, which are often identified by demographic analysis. Regarding issues

that are more specific to the internet era, analyzing demographic information could help study issues such as online privacy and security. A recent study on phishing susceptibility among different demographics groups has identified several factors that influence users' online behaviors [52].

While demographic information is traditionally collected through census and surveys, the abundance of user-generated content from social media and weblogs provide a unique opportunity for inferring demographic information directly based on users' input. Traditional demographic analysis is usually time-consuming and costly, especially if the survey needs to cover a large population. But with the help of social media, a fast and direct channel can be established with individuals for demographic surveys. In fact, researchers have taken advantage of the channel to collect demographic data through social media including Facebook [49, 61, 9] and Twitter [45, 15]. Various research methods have been developed to analyze the collected data, in order to identify features that can be used to predict demographic information such as age and gender.

Individual users post online discussions regarding their daily lives, international and local events, and other topics of interests. There is an opportunity to establish a direct connection between demographic groups and topics of interests, language style, as well as online social behavior. Such connection provides important insights on users' interests, social and behavioral patterns, and internet cultures, possibly distinguished by different demographic groups.

Much of the previous research on demographics analysis can be organized into two categories. Research in the first category analyzed user-generated content through

counting word usage over pre-determined categories of language in order to distinguish demographic groups [8, 19, 43]. Such approaches yield language usage features that are easy to interpret and can be used to make sense of the commonalities and differences among different groups. The second category of research adopted a more "open vocabulary" approach [49, 15], which does not restrict the analysis to *a priori* word categories. Instead, all words from user generated content can be used as features to classify users into different demographic groups. Such methods employ machine learning algorithms including SVM and Naive Bayes for age and gender classification. The objective of these approaches in the second category is to experiment with different features and optimize the machine learning algorithm to achieve the best accuracy for predicting demographic factors. In contrast to the first category, producing meaningful and interpretable results that distinguish demographic groups is not the focus of these methods. As a result, one drawback is that the features that distinguish the demographics group are not in a form that can be consumed by interested parties.

In this chapter, I offer a visual analytics approach to demographic analysis that combines the merits of the above two categories of research, in that I take a data-driven perspective and I establish a direct link between demographic groups and meaningful, easy-to-interpret features. More importantly, I provide an interactive visual interface for users to make sense of the connection between demographic groups and features that distinguish them, including topical and linguistic features. Compared to previous studies and computational methods on demographic analysis, the novelty of DemographicVis is that it enables interested parties to directly connect demographic information with the computationally extracted yet meaningful features

of the corresponding demographic groups. Previous text visualization systems, such as [24, 63, 20, 3], focus on developing novel approaches to explore and analyze large corpora alone. In contrast, DemographicVis explicitly connects categorical data (demographic factors in this case) with textual contents.

The major contributions of the chapter include:

- A new visual analytics system, DemographicVis, is presented that integrates state-of-the-art analytical methods with a novel visual interface to clearly show the relationship between demographic information and user-generated content. The visual interface includes a rotated parallel set and interactive word clouds, and is tailored to present the connection between demographic information and the features that distinguish various demographic groups. DemographicVis makes explicit connection between categorical data (demographic information) and textual data (user generated content).
- DemographicVis enables a transparent way to conduct demographic analysis, making the features that best describe different demographic groups easy to interpret and ready to consume by the end users. Topical, linguistic, and peripheral features are extracted from both user-generated contents and meta data using multiple machine learning algorithms. The features are also ranked to demonstrate their importance for predicting demographic information.
- A quantitative evaluation is provided to compare DemographicVis to SAS TextMiner, a commercial text mining software for extracting insights from textual data. The evaluation results show that DemographicVis received significantly

higher rating in terms of ease of use and ease of learning with comparable performance on achieving various tasks.

3.2 Data Collection

3.2.1 Demographic information collection

To collect demographic information from online users, I designed and posted a survey on Reddit.com, an online link-sharing community and message board. Reddit has gained popularity in recent years. In order to obtain demographic information directly from this community, I first compiled a set of multiple choice questions. Following IRB approval, the survey was posted on the r/SampleSize subreddit. This community is dedicated to generating and answering surveys. In my survey, I also designed a set of control questions with simple answers to rule out the participants who answer the survey questions randomly.

The information collected through the survey includes each responder's gender, gender expression, age group, education, current location, income level, religious affiliation. 482 users participated in my survey, 409 users were included in the final data collection after filtering based on the control questions. Figure 4 presents the summary of information all 409 responses. The summary suggests that my pool of participants are fairly balanced as to gender, although Reddit is thought to be a male-dominant community. In terms of age group, the results showed a good coverage of individuals ranging from 17 or younger to 39 years old.

Note that previous studies that focus on analyzing and predicting just age and gender tend to have larger sample population, since gender and age information is

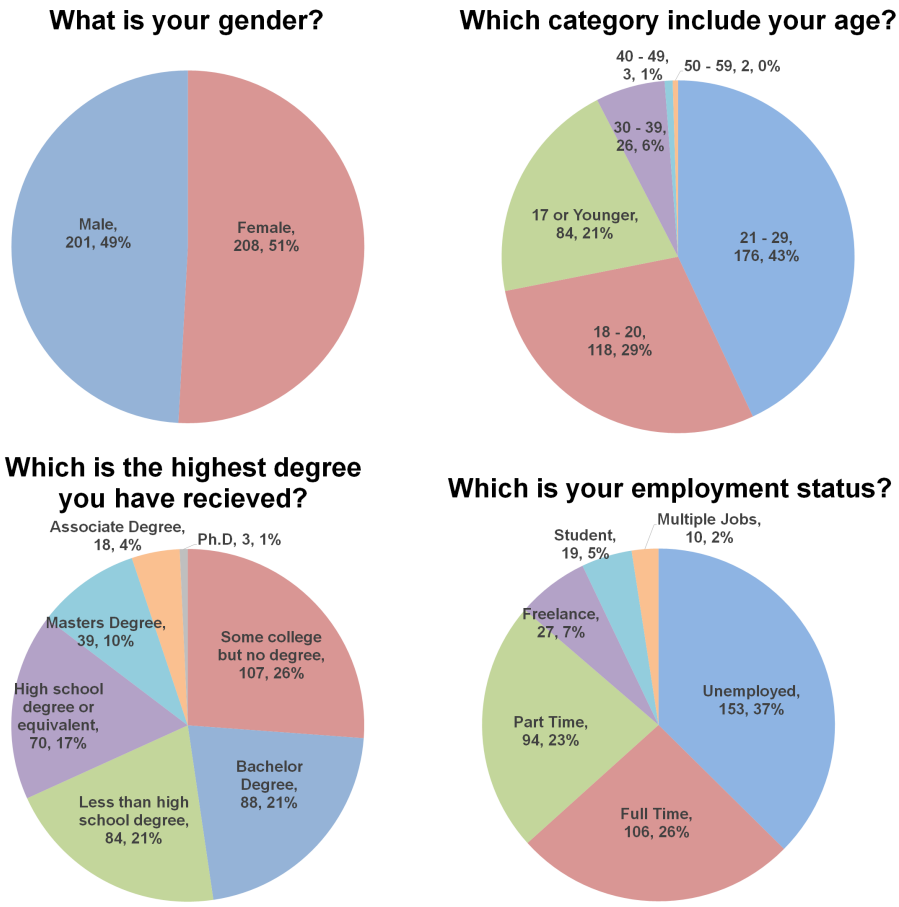

Figure 4: Demographic distribution on gender, age, education, and employment status based on the collected data.

more readily available. However, in my study, I collected more detailed demographic information well beyond age and gender. Although a sample of 409 Redditors may not provide sufficient statistical power for generalizing my findings to broader contexts, my visual analytics approach to demographic analysis can be applied to information collected from a much larger population.

3.2.2 Collection of user-generated content

The objective of my research is to connect demographic information with the content the users posted on social media through visual analytics means. To this aim,

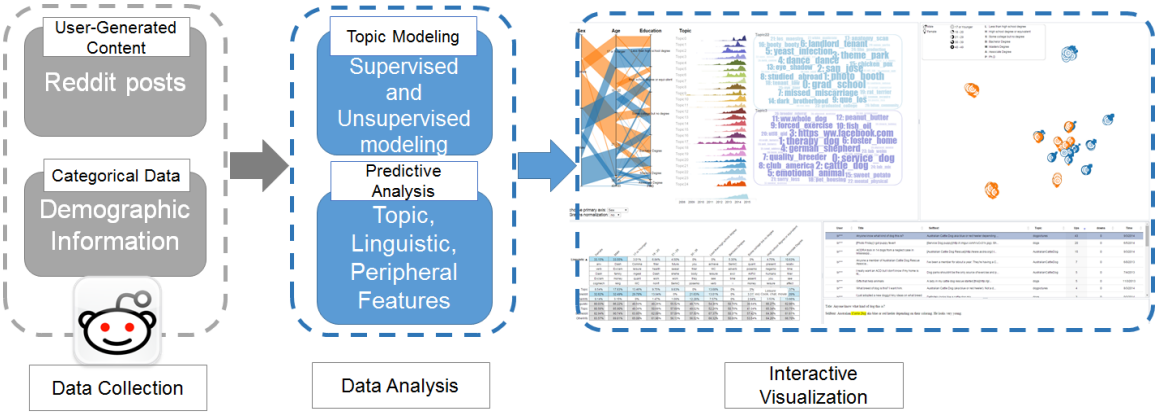

Figure 5: System architecture of DemographicVis. Section 3, 4, and 5 introduce the Data Collection, Data Analysis, and Interactive Visualization component respectively.

I collected the posts from the 409 valid users of my survey. I developed a python crawler to gather the posts of this group of users through Reddit's public API. The final dataset included 169,707 posts, the time stamp of posted comments ranged from 2011 to date. Unlike Twitter, the posts on Reddit do not have length restriction. Therefore a large portion of the comments contained at least a few sentences.

In my final dataset, each record is one post from an individual user. Since each user belongs to a certain demographics group, each record is then associated with the corresponding demographic group. Different from previous studies, I analyze the user-generated content based on multi-attribute demographic groups as opposed to examining attributes such as age, gender, ethnicity individually. Therefore, each user comment in my database is tagged with one multi-attribute demographic group. The attribute combination can be chosen based on the analysis needs. For instance, one common combination is gender, age and education. An example of a particular demographic combination could be {Female, Age 18 - 20, High school degree}.

3.3 DemographicVis: A Visual Analytics Approach for Demographic Analysis

My approach combines analytical methods and an interactive visual interface to enable the analysis of the relationship between demographic information and user-generated content. The system architecture of DemographicVis is shown in Figure 5. In this section, I focus on introducing the "Data Analysis" component; the "Interactive Visualization" will be described in the next section.

3.3.1 Data modeling and feature analysis

To describe different demographic groups based on user generated content, I extract features from the Reddit posts, including linguistic and topic features. I also extract additional features from the meta-data associated with each post; and use them in conjunction with topic and linguistic features for predictive analysis.

3.3.1.1 Topic features for describing demographic groups

To describe the demographic groups based on user generated content, a concise and meaningful summary of interests of each individual demographic group needs to be extracted. To this aim, I perform supervised topic modeling to extract topics for each demographic group. Since the direct relationship between topics and the demographic group is essential to my objective, I want to establish direct links during the modeling process.

To incorporate demographic information directly into the topic extraction process, I adopt the Tag-LDA model [40] that was designed to include tags or labels of each document during the topic modeling process.

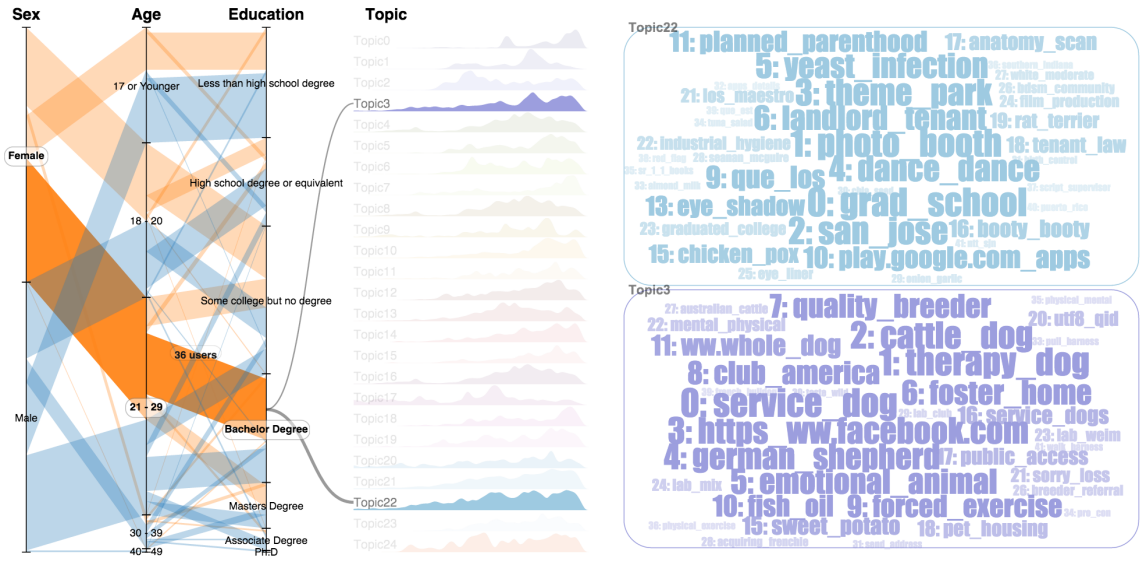

Figure 6: Hovering mouse over demographic group ‘‘Female, 21-29, Bachelor degree’’ leads to the highlighting of the two topics that summarize the interests of the group.

In my approach, each multi-attribute demographic group serves as a tag for each comment a user posted on Reddit. The data contains 36 unique multi-attribute demographic groups. As a result, I can now use the topic results to describe the interests of each demographic group. For instance, as seen in figure 6, the two topics that best describe multi-attribute group “Female 21-29 with Bachelor’s degree” are Topic 22, which includes keywords “grad_school, study_abroad, theme_park,” and Topic 3 that focuses on discussing dogs and puppies.

When performing the topic extraction, I employ a bigram approach that treats two consecutive words in a document as a unit of analysis. According to [35], bigrams serve as better feature representations compared to unigrams. Employing bigrams during topic modeling enables the users to discover more phrases with richer meaning such as “birth control” and “medical marijuana.”

3.3.1.2 Topic, linguistic, and a set features for predicting demographic factors

While topic features are great for presenting a visual summary of the interest of different demographic groups, they can also be used for predicting demographic information based on user generated content. For the task of predictive analysis, I extract a set features and further analyze which feature or combination of features can be used to best predict certain certain demographic information. This is especially useful given that the availability of demographic information on social media is scarce and often unreliable. In this section, I describe features I extracted from both user generated content and meta data for the purpose of predictive analysis. These features are then ranked and presented in the visualization so that the users can interactively make sense of how each feature contributes to the task of predicting different demographic attributes.

The entire feature set contains three subsets of features: topic proportions, linguistic features, and peripheral features extracted from metadata. It would be convenient to reuse the topic proportions from the above supervised topic modeling process, however, I choose not to because linking the labels with the test data that may not have a good coverage on all content will lead to a poor overall prediction performance. In the following descriptions, I will introduce how I derive the features for predictive analysis.

To obtain the topic proportions, I first construct a
term-document matrix using a bag-of-words model based on all available Reddit posts. Next, I perform topic modeling using a method called nonnegative matrix factorization

[18] with the total number of topics set at 100. I then obtain a 100-dimensional topic-wise vector representing each Reddit post. Next, for each user, I sum up these 100-dimension topic-wise vectors of all the Reddit posts a user has written, and this aggregated vector works as the topic proportion feature for a particular Reddit user.

Linguistic feature. To extract linguistic featuress from user generated content, I performed the LIWC analysis. 82 linguistic variables were extracted from the Reddit posts. Such variables include general descriptors, standard linguistic dimensions, etc. A complete list of the variables can be found here[3]. The linguistic feature is then used to perform predictive analysis in conjunction with the topic features.

Peripheral features. Features in this category include the subreddit and other features. I generated the subreddit feature as the 4,296-dimensional vector whose dimension is the total number of unique subreddit categories, where the value represents the count of the Reddit text that a user has written in the corresponding subreddit category. The peripheral feature includes a 5-dimensional feature vector containing the total number of Reddit posts for a particular user, the ratio of original posts (not including comments on other Redditors' posts) to the total number of posts, the total number of thumb-up, the total number of thumb-down received, and the total number of comments that other Reddit users have written in response to the user's Reddit posts.

3.3.1.3 Predictive analysis

I obtain a 4,483-dimensional vector for each Reddit user. Then I build a binary classifier by using two groups at a time: one containing users in a particular demo-

[3]http://liwc.wpengine.com/

graphic group and the other containing those in the rest of the groups. Considering the high dimensionality and the heterogeneity of these feature vectors, it is critical to use the most capable learner that can properly handle the data. To this aim, I use a gradient boosting tree (GBtree) [31] [4], a state-of-the-art ensemble model that adopts a decision tree as an individual learner. The classification performance is shown as a feature table in Figure 3C.

The feature table is divided into two parts: the top table presents the contribution of different features in predicting the demographic variables, while the bottom table presents the accuracy of the predictive analysis for different demographic variables, measured by the Area Under the ROC Curve (AUC). The reason for choosing the AUC value as the evaluation measure rather than a simpler measure such as the prediction accuracy is because the dataset is highly unbalanced between a particular demographic group and the rest. The AUC value is not dependent on the imbalanceness of a dataset.

The top table shows the variable importance scores when I only incorporate particular features. That is, I perform the prediction experiments by using only one feature group corresponding to each row (e.g., 'Linguistic', 'Topic', etc.) at a time and measure how much the binary classification performance **increases** in terms of the AUC value from a random guess classifier of 0.5.

The bottom half of the table shows the **cumulative** AUC values when I gradually incorporate more features. For example, the first row shows the AUC values only

[4]The implementation of GBtree I used is available at
https://sites.google.com/site/carlosbecker/resources/gradient-boosting-boosted-trees

when using the 'Linguistic' features, and the second row shows the AUC values when using the 'Topic' features together with the 'Linguistic' ones. From this table, one can see gender can be well predicted, as shown as the overall performance of 83.57% and 89.81% in the first two columns. This study is one of the very first ones that provides promising results for predicting diverse demographic characteristics in terms of age and education levels, among other aspects, rather than just predicting gender. More importantly, the use of visualization helps users to make sense of the contribution of different features.

3.4 Visual Interface

The interactive visualization permits the sense making and comparison of different demographic groups, as well as identifying features that can be used for demographic information prediction.

To enable interactive demographic analysis based on the features introduced in Section 3, I designed a web-based visual interface that connects categorical data to user generated content. The interface consists of multiple views that were designed based on tasks summarized from interviewing users who are interested in performing demographic analysis. Such users include our industry partners who are interested in performing demographic analysis for marketing purposes, and academic professors who are interested in understanding online behaviors of different demographic groups.

In the context of connecting demographic information with posts from social media, the interviewees are most interested in the following 3 analysis tasks:

T1 What are different demographic groups interested in? Do different demographic

groups have distinct interests that are reflected in what they post?

T2 Which demographic groups share similar interests?

T3 Can we leverage information derived from posts on social media to successfully classify online users into different demographic groups when there is little ground truth available?

To address the three tasks, I introduce an interactive visual interface that consists of the following three views.

3.4.1 Parallel Sets + Word Cloud: connecting demographic groups to topic features

To address T1, I leverage visualizations tailored for categorical data and topic results. Specifically, I combine transformed Parallel Sets with interactive word clouds.

3.4.1.1 Parallel Sets for demographic data

Parallel Sets is designed for visualizing relationships between dimensions in categorical data. I applied it to three demographic dimensions: gender, age and educational level. In contrast to the original ParSets layout, I made a design decision to rotate the ParSets by 90 degrees for two reasons: 1. The dimensions are then drawn from left to right, which conforms to the natural reading direction of most people; 2. Such rotation allows easy connection between the demographic dimensions and the topic word clouds as shown in Figure 3A.

To enable users to explore hypotheses regarding different demographic variables in a flexible manner, the DemographicVis interface permits interactive selection of the starting dimension, since the first dimension in ParSets determines the color assignment. Ribbons connecting adjacent dimensions are sized according to the

number of users falling under the combination of the two demographic variables. As seen in figure 3A, the label for each dimension is on the top while the label for each category within a dimension is placed at the center of the category.

User interaction. When the user hovers over a ribbon, the ribbon is highlighted while the other ribbons are dimmed. At the same time, the corresponding demographic variables are highlighted. To enable examination of a certain demographic variable group, clicking on a ribbon will keep the ribbon highlighted when hovering out. This interaction is important when trying to connect demographic groups to their corresponding topics.

3.4.1.2 Topic representation: Word Cloud + Streamlines

To present the topic interests of various demographic groups, I link interactive word clouds and topic streams to the demographic information. Each word cloud depicts one topic derived through the modeling process (Section 3.1), while each topic stream portrays the temporal trend. The time span of the topic stream ranges from late 2008 to early 2015, with most posts published between 2013 and 2015. Because of the supervised modeling process, each topic describes interests of a specific demographic group. For instance, as shown in figure 3g, topic 0 describes the interests of group "male, 18-20, high school degree or equivalent", which includes keywords related to computer hardware (video_card, hard_drive, etc.) and sports games (premier_league, sport_football). To highlight the ranking of keywords in the word cloud, I use a combination of font size, opacity, and numbering. The size of each bigram is determined by the probability of the bigram in a particular topic. To further

distinguish the most important bigrams, I added a number in front of the leading bigrams to indicate their precise ranking in the topic. The bigrams are animated when popping up, so that the most probable ones appear first. Each topic is drawn inside a rectangular bounding box, with the size of the box dynamically determined based on the total number of topics displayed.

User interaction. Users can explore the relationship between demographic groups and topics via a two-way interaction. On the one hand, hovering the mouse over a certain demographic group would highlight the corresponding topic feature(s) that describe the interests of the demographic group. On the other hand, hovering over a particular topic stream would highlight the demographic group(s) that are interested in this topic.

To help users better understand the topics, clicking on a bigram in a topic brings up a list of posts that contain the bigram. The list of posts will be displayed in the post view shown in Figure 3D. The post view displays information including anonymized user name, the post, the subreddit information, a time stamp, and votes on the posts. Seeing how a bigram is mentioned in the detailed posts helps users to understand certain keywords that might seem obscure at first.

3.4.2 User cluster view

To address T2, namely to find out whether the demographic groups have distinct or similar interest, I grouped the 409 Redditors that participated in the survey based on the content they posted. Such grouping allows one to easily discover whether users belonging to the same demographic group share similar interests.

To generate clusters based on the interests of the users, I leverage the topic results (Section 3.3.1.1). The similarity between two users are computed by calculating the KL divergence of their topic distributions. To map the similarity matrix computed for all 409 Redditors, I leverage a dimensionality reduction method called t-Distributed Stochastic Neighbor Embedding (t-SNE) [57]. t-SNE is particularly well suited for the visualization of high-dimensional datasets since it generates compact yet separable clusters [57].

To further assist users in linking the above dimensionality reduction results to demographic information, I designed glyphs to encode the demographic factors in the clusters. With the glyphs, it is easy to see whether users belonging to the same demographic group are clustered together. As seen in Figure 3B, each glyph represents one Redditor, with their demographic variables (gender, age, education) captured by the glyph. I went through an iterative process to finalize the glyph design so that it is both intuitive and easy to read. The gender variable is represented by the two standard gender symbols denoting male ♂and female ♀. The age variable (5 age groups) is encoded in the outer ring of each glyph, with each tier in age group adding 1/5 of the filling. Lastly, the education variable is encoded in the inner circle of the glyph as the first letter of the education level. The glyph encodings are shown in the legend. As seen in figure 3B, most bigger clusters, especially the ones located on the outer parts of the view, mainly contain one demographic group. Such observation leads to the hypothesis that these demographic groups have fairly distinct interests from other groups. With the two topic groups that seem to involve more than one demographic group, the user can further understand how the demographic groups are

intertwined though interactive analysis.

User Interaction. Hovering mouse over one glyph that represents a particular demographic group highlights all other Redditors belonging to the same demographic group. Such interaction allows easy discovery of whether Redditors in the same group have cohesive or diverse interests. Such interaction also permits rapid analysis of clusters that seem to involve more than one group. Figure 3B shows two clusters with each including two different demographic groups. Such pattern suggests that the two demographic groups share similar topic interests based on their Reddit posts.

The cluster view is coordinated with other views. Hovering over a certain demographic group in the cluster view will lead to highlighting of the corresponding demographic group in the ParSets and the corresponding topics. Conversely, when highlighting a particular demographic group or a topic in the ParSets+word cloud view, the corresponding group will be highlighted in the cluster view.

3.4.3 Feature ranking view

To address T3, namely allowing users to analyze the connection between demographic information and user generated content using features beyond topics, and to represent the predictive power of various features, I provide a feature ranking view (figure 7). In the tabular view, the features are aligned by row and the demographic variables are aligned by the column.

As introduced in section 3.3.1.3, the top table (with blue background) presents the contribution of different features in predicting the demographic variables. For instance, to predict gender as being male, linguistic features make the biggest contribution at

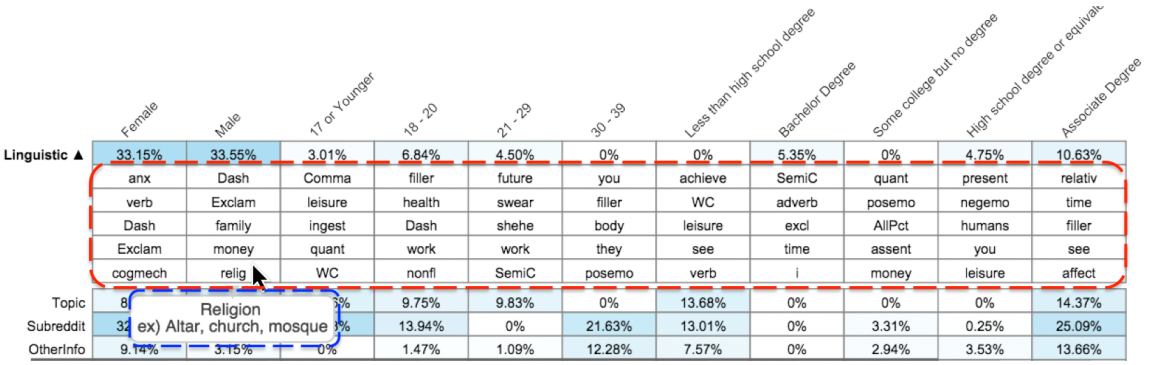

Figure 7: Feature table with linguistic features expanded. Five highest ranked linguistic features are displayed for analysis.

33.15%. Subreddit is the next best feature for such prediction based on user generated content. The background color of the cells is blue and the opacity is determined by the percentage displayed in each cell.[5]

The bottom table presents the cumulative accuracy of the predictive analysis for different demographic variables. Since the accuracy for each demographic variable in a column is analyzed in a cumulative fashion, I want to use the background encoding to reflect the accumulation. The background of the cells in the bottom table is a horizontal bar graph, with the length of each bar determined by the accuracy results. The contribution of the feature analysis view is that it enables the interactive analysis of the contribution of different features to demographic information prediction.

User Interaction. Since the linguistic feature contains 82 dimensions, with each dimension as an interpretable sub-feature, DemographicVis supports the expansion of the linguistic features to show more detailed ranking information. When expanded, the top 5 highest ranked linguistic features (figure 7 red rectangle) are presented. Hovering over a cell brings up the definition of the linguistic feature (figure 7 blue

[5]0% is likely due to insufficient data

Figure 8: Female, 17 or younger, less than high school degree and the corresponding topic of interest.

rectangle). Investigating Female and Male groups in the user pool, one can see the important linguistic features for female are anx (anxiety, e.g. worry, fearful, nervous), verb (common verbs), and cogmech (cognitive processes). Different set of features are highly ranked for male groups, including family, money, and religion. Such interactive analysis can potentially lead to significantly deeper understanding of how the different demographic groups talk and behave online.

3.5 Case Studies

In this section, I present case studies to illustrate how DemographicVis could help users compare and make sense of different demographic groups based on user-generated content from Reddit. Given that the majority of the Redditors are young users, I take this opportunity to examine the interests of young crowds.

3.5.1 What are young ladies interested in?

The individuals in the "female, 17 or younger, less than high school degree" group are teenage girls attending high school or middle school. The most probable topic for describing their discussions is highlighted when hovering over the group, as shown in figure 8. The topic summarizes the young crowds' interests on reading Japanese graphic novel ("nurarihyon_mago") [6], discussing makeup related terms ("nail_polish"))), and

[6] At first I thought this finding is due to the particular sample on Reddit. However, a study [49] based on 75,000 Facebook uses also found dominating interests in Japanese comics among young users.

Figure 9: Female, 18-20, Some college but no degree and related topics.

mentions of "panic_attack", which could be concerning since clicking on the term leads to posts that discuss mental illness.

Individuals in the "female, 18-20, some college but no degree" group are likely college students. As shown in figure 9, two topics are highlighted to describe the interests of this group. One can see both continuation and evolution in topic of interests compared to the younger generation. While continuing the discussion of makeup related terms "makeupexchange_comments" and "urban_decay"[7], young ladies within this demographic group also discussed gender identity and related issues: "male_female", "gender_identity", "birth_control".

3.5.2 Exploring diversity and overlap in interests among demographic groups

To identify whether there are overlaps between various demographic groups, we can leverage the cluster view in DemographicVis. As shown in figure 10, Reddit users are grouped based on their topic interests. Through a quick glance, one can find separable and cohesive clusters around the outer part of the view. For instance, the big orange cluster on the top (female, 21-29, bachelor degree) and the blue cluster

[7] A cosmetic brand

Figure 10: Cluster view that groups Reddit users based on their topic interests. One group with two demographic groups is enlarged and their shared topic is shown.

on the bottom (male, 17 or younger, Less than high school degree) illustrate that the two groups have fairly distinct but cohesive interests. Hovering the mouse over a cluster will highlight the corresponding topics in the ParSets+Word Cloud view[8]. Interestingly, the clusters with different demographic groups denote groups that share similar interests. Here I pick a cluster (a zoomed in view of the cluster is shown in figure 10) with two demographic groups, namely "female, 21-29, some college but no degree" and "male, 30-39, masters degree". Hovering the mouse over the two groups leads to the identification of the common topic shared by the two groups, as shown in figure 10. Through a quick examination of the posts and subreddit where the two groups of users published their posts, we can then observe that the two groups are both interested in "weight_loss, lose_weight".

[8]The mixed glyphs in the center of the figure mainly contain the demographic groups that do not have sufficient numbers of users; as a result, it is difficult to model their interests.

3.6 User Study

In this section, I present a user study to evaluate efficacy and usability of DemographicVis by comparing it with the SAS Text Miner [47]. The SAS Text Miner is one of the advanced analytics products developed by SAS that aims to extract insights from unstructured text. I consider SAS Text Miner as the best candidate for comparison because it integrates analysis and visualization capabilities for textual data.

3.6.1 Experiment tasks and apparatus

I designed 3 tasks for participants to perform with both DemographicVis and the SAS Text Miner after getting an IRB approval for the user study. The main goal is to ask the participants to investigate the demographic groups and identify the connection between the demographic groups and the topic features. When designing the tasks, I wanted to test users' understanding of predictive analysis based on various features. However, I could not find functions in the SAS Text Miner that support this task. The 3 tasks I settled on are:

Task1 : Identify 3 most frequently discussed topics by each demographic groups.

Task2 : For the 3 topics identified in Task1, find the corresponding demographic groups that mainly discuss these topics.

Task3 : Given 3 randomly assigned topics, rate the interpretability of the topics.

Task1 requires users to pick topics that are discussed most often by volume; both DemograhicVis and Text Miner provide functions to complete such a task but with

different visual representations (topic stream vs. pie charts). Task 2 is to identify the demographic groups that discuss the 3 topics picked during the first task, while task 3 focuses on the interpretability of the topics extracted from both systems. In DemographicVis, the users can also access the Reddit posts (by clicking on terms in the topics) to aid the interpretation of the topics.

30 users participated in the user study, 19 males and 11 females (17 Ph.D. students, 5 masters, and 8 undergraduates). The age of the participants ranged from 18 to 40 (M=26). Participants were first asked to fill out a pre-study questionnaire regarding their demographic and experience with text analytics visualizations and Reddit use. I found that not many participants frequent Reddit.com (M=1.9 on a scale of 1 to 7). Many participants rated that they are pretty familiar with visualizations (M=3.5 on a scale of 1 to 7), such as word cloud and tree map. Out of the 30 participants, 5 participants answered that they were familiar with the SAS tool and 7 answered that they are familiar with text summarization methods.

I prepared two training videos on DemographicVis and SAS Text Miner. Prior to starting the study with each interface, the participants watched a 3 minute training video that demonstrates how to use the interface for completing the tasks. The two interface conditions were presented in counter-balanced order across all participants. After a participant finished the user study with both interface conditions, she was asked to fill out a post-study questionnaire regarding subjective preferences on the interfaces as well as each interface's advantages and disadvantages. All participants successfully finished the user study within 60 minutes.

3.6.2 Results

In this section, I report the results from the user study. Subjective ratings (7-point Likert scale) on the two systems are reported. I also summarize the user feedback on the pros and cons of both systems. The subjective rating data were analyzed with Friedman's test with α=.05 level for significance. There is a huge debate ongoing in the socialbehavioral sciences over whether Likert scales should be treated as ordinal or interval. I choose to treat it as ordinal, therefore Friedman's test is used to analyze the results.

3.6.2.1 Accuracy rate

For Task1, the participants were asked to pick 3 most discussed topics. I found no statistically significant difference between DemographicVis and SAS Text Miner (M=0.93 vs. M=0.96). Both systems provide functions that enable such identification.

For Task2, I asked participants to find the corresponding demographic groups for the top 3 topics. I calculate error rate as the number of incorrect answers. The result of an one-way ANOVA shows a significant main effect on accuracy rate of interface condition ($F(1, 29)$=7.760, p=.009, η_p^2=.211). When using DemographicVis, participants exhibit a higher accuracy rate (M=2.83, SD=0.59) than SAS Text Miner (M=2.17, SD=1.15). Based on the user feedback, the reason that DemographicVis outperformed the SAS Text Miner in this task is because of the interactions supported to help users easily connect topics with demographic groups. Although the SAS Text Miner provides similar functions, lack of direct manipulation on the visualization (pie charts) and coordination between multiple windows makes it difficult for users to

identify the demographic groups.

For Task3, although DemographicVis and SAS Text Miner use different methods to extract topics, the participants rate the interpretability ($\chi^2(1)$=.182, p=.670) as comparable.

3.6.2.2 Learnability and usability

Participants rated learnability (easy to learn) and usability (easy to use) for each interface after performing all tasks (on a 7-point Likert scale, 1:*very difficult* to 7:*very easy*). The results of Friedman's test show that there is a significant difference on a learnability rate ($\chi^2(1)$=6.545, p=.011). Median (IQR) learnability rates for DemographicVis and Text Miner are 5.5 (4.75 to 6) and 5 (3 to 6). In addition, there is a significant difference on a usability rate ($\chi^2(1)$=8.048, p=.005). Median (IQR) usability rates for DemographicVis and Text Miner are 5 (4 to 6) and 5 (3.75 to 5.25). Overall, participants rated that DemographicVis is easier to learn and use than the Text Miner to accomplish the designed tasks.

3.6.2.3 Subjective preference

When asked which system one prefers for accomplishing Task1, 20 out of 30 preferred DemographicVis, 7 preferred SAS Text Miner and 3 answered both. For Task2, 24 out of 30 preferred DemographicVis, 5 preferred Text Miner and 1 answered no preference. For Task3, 21 prefer DemographicVis while 6 preferred Text Miner, 2 answered both are same and 1 answered no preference. In terms of overall preference, 25 out of 30 answered that they prefer DemographicVis and 5 answered they prefer SAS Text Miner. From the open-ended comments, I found many participants commented on

features provided by DemographicVis that show correlations between topics and demographic groups including “Different views were synchronized and responsive”, and “It was easier to detect the correlation of topics to groups in DemographicVis”. In contrast, many commented on Text Miner’s lack of view coordination “need to open extra windows”, and lesser visualization quality“the nested pie charts are confusing”.

3.7 Discussion and Conclusion

Throughout the design process, I noted that there are aspects I’d like to continue to improve in both data collection and analysis.

First, to be able to make substantial claims on findings regarding the interests and online behaviors of various demographic groups, I will need to collect a much larger sample. In practice, this is difficult to achieve on Reddit.com alone since the only place one is allowed to post surveys is the r/SampleSize subreddit. I plan to conduct similar surveys on other social media platforms such as Twitter. It will be interesting to conduct comparative analysis on what the demographic group publish on different social media sites.

Second, I would like to improve the feature analysis process to achieve better predictive results. Having a larger sample could help, but more features would also be added to boost the performance of the the predictive results. Some features may be platform specific. For example, in the experiment, subreddit turns out to be a good feature for predictive analysis. Other general features such as how often a user posts (indicating how active she is) or how many different subreddits or topic groups the user posts in may also contribute to the overall predictive analysis. In terms of using

the topic features for predictive analysis, I can experiment with different topic models to see which one may yield better results. I did experiment on generating different number of topics with the NMF-based topic model, and found that the number of topics does not affect the predictive results and the contribution of the topic features.

3.8 Conclusion

I introduce DemographicVis, a visual analytics system that aims to support interactive analysis of demographic information based on user-generated contents. DemographicVis visualizes features that are extracted to either describe or predict demographic factors, and enables the exploration of demographic information in a transparent manner. Results from the comparative evaluation shows that DemographicVis is quantitatively competitive and subjectively preferred compared to the SAS Text Miner.

In this chapter, I address the challenge of characterizing connections between cohorts' categorical metadata and their topics of interests. The correlation that I focus on is between topics of interests for a group Reddit users and their demographics. The users' demographics, which I surveyed, are the metadata while the topics modeled from the users' posts are the unstructured contextual data. Thus, I am assuming that the cohorts' labels are already known and linked to the users' posts. To relate this goal to the main theme of this dissertation, I leverage a visual analytics system to explore the connections between the social media users' posts and their demographics. The VA tool allows its users to understand the characteristics of different demographic groups in a transparent and exploratory manner. The two goals of the VA tool

developed are to model and infer demographics of users on social media by connecting demographic information with user-generated content, and rank features that best describe characteristics of each demographic group. In this chapter I also discuss the prediction of demographic factors, such as topical, linguistic, and peripheral features, which are extracted from both user-generated contents and metadata.

CHAPTER 4: INTERACTIVE SOCIAL MEDIA USER ANNOTATION

''The one who follows the crowd will usually get no further than the crowd. The one who walks alone is likely to find themselves in places no one has ever been before.''

-- Albert Einstein

While in chapter 3, I show how VA systems can be used to support the characterization task for the domain experts by connecting different user categories with their generated content, in this chapter I show how VA systems can support the labeling task to categorize the users, when their categories is the missing information. Therefore the analysis task addressed in this chapter is complementary to the task in the previous chapter as described before in figure 1. The domain experts in this chapter are communication studies and psychology experts. I collaborated with professors from the Department of Communication Studies and Department of Psychological Science at UNC Charlotte. We started a project where its goal was to characterize users in a Twitter dataset. However, it was challenging when the communications expert realized that there are many bots which could skew the results. In addition, we found that there are many "pre-written tweets" from activists who exhibit similar behavior like bots but with subtle differences. The domain experts on our team found the need to separate bots from other types of users in order to contribute to a theory regarding

people's reactions to an activist event. Thus, this research goal has motivated us to facilitate a VA system for labeling users wherein the domain experts can differentiate between different types of users. The aim of the tool is to help the domain experts label social media users efficiently; i.e. with higher accuracy and less time than when using traditional tools.

4.1 Introduction

Researchers attempt to characterize user behavior on social media for different applications and goals. There have been many research studies aiming to characterize the users' behavior on social media, and scientists are usually interested in applying their analysis to serve different applications related to social sciences and many other industries. Despite the unprecedented potential that social media has given the scientific community, interactions on these platforms are plagued with automated deceitful accounts called bots. Recently, Twitter and other social media platforms have been swarmed with bots to interact with regular users for many different purposes, some of them are meant to be harmful and some are for advertisement or educational material [27]. Most of the harmful social media bots target users to steal their personal information or infiltrate their network, which could be used for political purposes. The tweets that are produced by these bots are problematic for communication experts, because they skew the analysis of human behavior. For example, according to [54, 53] Twitter is full of bots that produce a lot of noise and impedes their analyses. While many of these interactions are public, they still contribute to difficulties researchers have when trying to extract meaningful data about conversations people have online.

Unfortunately, these bots impede domain experts' analyses by inducing noise.

As a result of these challenges, computer scientists started prediction models specialized in bot detection aiming to eliminate their noise; however, it is still a dilemma faced by many in the scientific community. Machine learning algorithms have provided us with many new and useful ways of isolating bot activity from real human interactions. While the machine learning algorithms are useful for classifying the users, they are ill-equipped in dealing with state-of-the-art advancements in NLG methods of synthesizing human grade conversations. The NLG field is advancing every day, which has led the social media users more susceptible to bots [58, 59]. This continuous advancement requires bot detection techniques that can catch up with the technology.

Furthermore, the phenomenon of "pre-written tweets" has made it more challenging for these algorithms to solve the problem of isolating the human users from bots, due to similar behaviors between bots and users who post pre-written tweets. The phenomenon was discovered in [28] about human users. Activist organizations use email alerts to invite followers to post the organizations' messages via Twitter and most of these messages contain links which redirects to their organization's website. These listservs are meant for activism as in expressing opinions and spreading the news. Pre-written tweets are identical or nearly identical across different user accounts. The repetition of these pre-written tweets can cause confusion to experts differentiating between bots and activists. This problem opens another door of investigation to identify tweets that exhibit bot behavior, so we can distinguish the influence of bots from pre-written tweets.

To battle these advancements and challenges, new VA systems are needed to provide computer-assisted multi-view interactive explanations coupled with state-of-the-art artificial intelligence techniques. This would allow for trends, relationships and more insightful information to be presented to the domain experts trying to learn about a subject or domain being discussed on social media. Providing a VA tool for domain experts to label bots and understand their characteristics can make substantial improvements towards identifying the important features for bot detection while having this automated detection technology up-to-date. There are bot detection tools that have been published such as Botometer (BotOrNot [21]) which is highly accurate, but these models need to be constantly fed with ground truth data to keep up with the NLG advancements. This could be achieved by making the communication experts' annotation process faster and more accurate. Also, the specificity of the main topic for a corpus influences the training of the predictive models. Thus, when domain experts are able to investigate corpora related to their domain knowledge, the tool will enable the experts to provide ground truth to train models specific to that domain.

In this chapter I present a VA system to support the task of social media user annotation in an active learning setting; the domain experts are able to run a text classification model at the end of each round of annotation to predict the user types and enhance the experts' decisions. The prediction results are displayed as probabilities to help the experts decide on the types of users that exist. The prediction feature is not supported at the initial round of annotation because the experts start the annotation process with no labels at the beginning. The system requirements are based on the interviews that I have conducted with the domain experts, which are mostly tailored

for the task and dataset collected. However, the most important tool features used by the experts to determine the user types can be applied to any Twitter dataset. The contributions in this chapter can be summarized in the following points:

- A new VA system that integrates a unique combination of visualization techniques aided by state-of-the-art prediction models to support the annotation task for domain experts.

- The system enables exploring and annotating specific portions of the corpus by selecting interesting topics. It also enables refining the tweets to focus on persuasive users (including bots and activists) that exhibit the repetition of tweets across the same topic. This refining process is ideal for annotating tweets or documents that would have a high probability of sharing common labels.

- The system supports the annotation task in an active learning setting. The cycle of this setting alternates between annotation sessions and predicting the topic models and classifications of the user types based on the previous expert's annotations.

- Part of the active learning setting, the system also visualizes the topic prevalence for specific predefined labels. For example, in this chapter the domain experts are shown cues of the bot prevalence among the topics. This contribution is one of the most important parts of the chapter's novelty and part of the topic modeling cycle that accommodates active learning setting.

In the rest of this chapter, I will start with the research questions and hypotheses

that I attempt to answer. Afterwards, I will discuss the data collection, and system requirements. Then I will describe the VA annotation system and how it can leverage information for making the annotation task more efficient. The user study of this chapter is coordinated to test the hypotheses in the next section /refchapter4:section2. Lastly, I will discuss the conclusions I have reached after developing and testing the tool.

4.2 Research Questions & Main Hypotheses

The goal of this chapter is to facilitate the VA system for domain experts to label the Twitter users efficiently in terms of speed and accuracy. The system enables the experts to run a text classification algorithm at the backend after each annotation session. The prediction results are shown as how much probability each user could possibly be any of the predefined user types. One important advantage of this system is that the text classification model is topic transductive, meaning that the model performs better on classifying the users in the specific dataset presented to the expert. The models' performances are not dataset agnostic. In other words, if one predictive model was developed for one dataset, it cannot be used for another, even if the labels assigned are the same. This guarantees better performance from the prediction model and the annotation process itself. The domain experts are interested to answer the following research questions which are inherited from the generalized research questions in the introduction chapter (chapter 1):

- Analytics component

    1. What is the feasibility to predict the predefined user types? How accurate

will the predictions be?

2. Will narrowing the topics down help the experts label the users efficiently?
3. Will displaying the predictions to the experts, in the active learning setting, help them label the users efficiently and isolate bots from the rest of the users? Will the predictions be trustworthy for the experts?

- Visualization component

1. How can the prediction results of the topic modeling and text classifier be presented visually to the domain experts without burdening them with technical details?
2. How can the topics of predefined user types be visually summarized to the domain experts without burdening them with technical details? Specifically, how to show the topics that have been infiltrated with bots rather than the other user types?
3. How can the users' posts visually aggregated to support comparisons between the bots and the rest of the users?

The main hypothesis in this chapter is that VA systems would support the domain experts to label the users more efficiently than when using traditional tools. This hypothesis is based on the fact that visualizations make it easier to summarize and fetch data than when using worksheet based tools like MS Excel, Numbers, Google Sheets. These traditional tools are limited in terms of data summarization functions and types of visualizations in which they would convert the data to. For example,

MS Excel doesn't have either cosine similarity or any topic modeling algorithms. On the other hand, tools like SAS have many of these functionalities but not all of them. In addition, SAS needs some programming experience, which is not something that most domain experts would want to spend their time on. Thus, I emphasize in this chapter that domain experts are in need of tailored VA systems in order to improve their efficiency in labeling social media users.

4.3 Data Collection & Extraction

The original goal of this project was to find the causality and effect of certain predefined user categories towards the topics that they discuss on Twitter. We aimed to characterize the users' behavior in a collection of tweets pulled using the keywords related to Unite the Right rally in Charlottesville, in August 2017 through GNIP [9], the Twitter data streaming service. We started with a collection of 706,233 English tweets that were posted within the period of February 7th and October 10th, 2017. The protests where held on August ^{12th}, however all the tweets within this time span was important: starting from legislation decision on the Robert E. Lee Monument removal that occurred in February until the significant drop of the number of tweets pulled from the GNIP query in October. The corpus did not contain any duplicated tweet IDs and excluded retweets. The total number of unique users in this corpus is 335,183. The corpus was collected using a diverse set of rules of keywords as listed below:

> (Charlottesville OR cville OR VA OR Virginia OR McAuliffe OR CvilleCityHall OR VSPPIO) AND (antifa OR Nazis OR Nazi OR neo-Nazi OR "Nazi/KKK" OR "KKK" OR "white supremacy" OR "white supremacists" OR #white-

[9] https://support.gnip.com/sources/twitter/

supremacists OR #whitesupremacist OR "white activists" OR "white activist" OR "James Alex Fields" OR statue OR memorial OR "Robert E. Lee" OR "Robert E Lee" OR "Lee Park" OR "General Lee" OR Confederate OR "Emancipation Park" OR "Stonewall Jackson" OR protest OR march OR marchers) OR cvilleaug12 OR #invisiblecville OR #HeatherHeyer OR #DeAndreHarris OR "DeAndre Harris" OR #unitycville OR #defendcville OR #cvillestrong OR #standwithcharlottesville

During our original planning, the domain experts categorized the users into four theoretical categories according to the users' activity and locality with respect to the Charlottesville incident. The activity of the accounts indicates if the user was actively tweeting about the topic before and after or only after the Charlottesville protest. We refer to "hot-topic users" to users who jumped into this conversation and started tweeting in August when the protests started. While the users who have been tweeting about this topic before the protest are called "enduring users," since they have been discussing and debating about the confederate monuments removal before the protest in August. We decided to separate the hot-issue from the enduring public by the cutoff dates between August 11 and September 3, which is based on the Charlottesville events, which broke on August 11, and the end date was determined by looking at how the conversations on Twitter declined through time across the whole data collection. The locality of the users indicates whether they tweeted from inside or outside Virginia. To determine the location of the tweets, we used the geolocation self-reported by the user in the tweet and created a list of keywords to determine if the location was part of VA. We used these locations to differentiate between the local and non-local publics. These four publics are not the actual labels that the domain experts wanted to annotate the users with, however, they are insightful metadata for the experts during their labeling task.

Our domain experts found many recurring tweets while analyzing these publics, which appeared like bot behavior, so they weren't able to make conclusions. Thus, the motivation for implementing this tool is to help the communication experts remove these bot users from the dataset and focus the analysis on the tweets that truly represent the four publics mentioned. Also, as part of the filtering process we excluded the deleted accounts, since bot accounts are more likely to be deleted. Out of the 335,183 users in our dataset, we found 317,624 (94.8%) active users, and 17,559 (5.2%) deleted accounts.

Another two important metadata information, which we extracted from the tweets, are the links inside the tweets and the tweet source. As part of the domain expert's task to isolate bots from other users, we extract the links inside the tweets and present their subdomain, domain and suffix separately to make it easy for the expert to relate the tweets to activist and media organizations. In most cases, the tweets contain shortened links which don't show the destination websites' names. The destination website's name sometimes indicate the nature and goal of tweet. Some links require multiple hops in order to reach the final destination. We extracted those destination links by recursively redirecting either until we can no longer be redirected to any further link or if the number of hops reached the maximum of 4 hops. The tweet source indicates the device, method, or any automated tweeting platform or social media marketing applications used to post the tweets. This metadata already exists with the Twitter dataset collected through GNIP without need for extraction.

4.4 System Requirements

Throughout the design and development process of the VA system, I interviewed psychology and communication studies professors and graduate students to understand the system requirements that would address their needs. Not all of the requirements were gathered at once, but rather were updated alongside with the rounds of system development. At the beginning the communication experts labeled a sample of the dataset on an MS Excel sheet to test the process of decision making and the challenges, which they are facing when using traditional tools. This test has enabled the communications experts to consolidate the user types into four main categories: bots, activists, media and individual users. They noticed patterns that persist among these four user types which can be summarized as follows; the tweet might have been written by a(n):

- **Bot** user if an identical/nearly identical tweet is posted by the *same* user multiple times. Also, if their tweets contain many mentions of other users via the "@" sign.

- **Activist** user if an identical/nearly identical tweet is posted by *different* users multiple times, and the tweets contain links that leads to an activist organizations' web page, or their screen name indicates activist organizations like: CREDO Action, Color of Change, Force Change or The Petition Site.

- **Media** user if their tweets contain links to news media websites and is identical/nearly identical to other tweets.

- **Individual** user if their tweets are unique and not identical and do not contain activists or media links. Although individuals might share media links, the researchers did not want the headlines to have an undue impact on the subsequent topic modeling.

Since bots and activists have similar behavior, differentiating between them might be a challenge for the domain experts when categorizing them.

In each round of development there were new improvements to meet the experts' needs and solve encountered problems. I summarize these requirements into five points, where every requirement meets a need or a problem to solve. The five system requirements are marked from [**SR1**] to [**SR5**], where **SR** indicates system requirement.

The large corpus has made it hard for the experts to digest this amount of documents as one big chunk to label. They felt the need to summarize the documents and associate their labeling sessions to topics of interest. The experts need to delve into specific portions of the corpus by navigating to topics of their interest. Also, dividing the corpus into topics enables the experts to understand the message that the users want to deliver on a high level [**SR1**].

The second requirement is to be able to focus on persuasive users who post similar tweets within the same topic. They felt the need to start their labeling sessions with tweets which have high similarity than the tweets which equally belong to different topics. They want to select the tweets which are highly similar and condensed within a certain range of topic proportions to increase efficiency in labeling. The experts also

mentioned that they would like to control the degree of similarity that would be used to filter the tweets within a topic [**SR2**].

The user categorization decision making process requires some metadata information besides the tweet body. They want to be able to have the activity, locality, source, generator and links displayed next to the tweets to enhance their decisions. Displaying these features is important for the experts in order to differentiate between bots and other persuasive users, who post pre-written tweet [**SR3**].

One of the most important requirements that the system encompasses is the active learning setting. The experts' efforts after each session can be used to train a classification model and predict the types of users who weren't labeled yet during the previous sessions. The prediction model's outcome would assist the domain experts by displaying the probability of each user, and the expert then makes their own decision. After every round of labeling, the model's accuracy is expected to improve and enhance the expert's confidence. The prediction model is not expected to be very accurate at the first few rounds, which also depends on how many users the expert labels at each round. As exception, the first round would not have any probabilities to assist the experts, since the model depends on the expert's input [**SR4**].

Part of the active learning setting is providing feedback to the experts on the topic level. The experts want to know the bot prevalence for each topic. This requirement comes from their interest in understanding the behavior of the bots. Particularly, they want to know the topics that are mostly infiltrated by bots, and hence, easily decide which topics to focus on first [**SR5**]. This requirement goes hand in hand with the first requirement [**SR1**], which requires the navigation tools to focus on the topics of

their interest.

4.5 Visual Analytics Approach for annotation

In this section I describe my approach to address the system requirements mentioned in the previous section (section 4.6.1). I will start first by showing how the visual interface works, and then get into the details of the modeling in the second part.

4.5.1 Visual Interface

In figure 11, I show the visual interface in the final development phase. The interface consists of three main views besides the control panel on the most left side: the topic overview, topic density plot, and the detailed view. Figure 11 shows the view after filtering the tweets and getting to the stage of labeling. The process of filtering the tweets starts from the topic overview, where the experts select their topic of interest from a network of topic nodes. This filtering addresses [**SR1**]. Each node represents one topic using five words under it. Upon the topic selection, the topic density plot of the selected topic will be shown on the top right view. The topic density plot shows the density of the tweets' volume (on the y-axis) for the different ranges of topic proportions. The topic density plot is meant to guide the experts to find the tweets that are highly similar (topic proportions close to 100% or 1) and condensed within small ranges of topic proportions (large bumps). The expert can select the range of topic proportions they want to label, and the detailed view will automatically show the individual tweets within the selected range. In addition to the topic proportion filtering, the expert can set the tweet similarity threshold through the second scale in the control panel. Both of these filters are needed to meet the system requirement

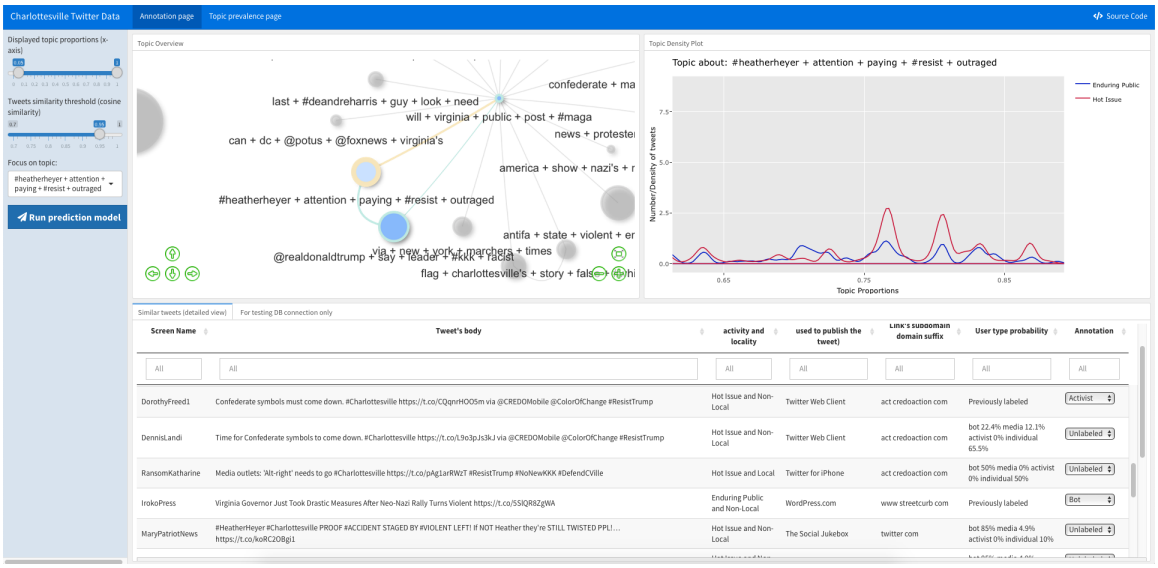

Figure 11: Visual Interface overview: the interface consists of three main views and one control panel. The top left view is the topic overview, the top right is the topic density plot, and the bottom view is the detailed view of the individual tweets. The control panel is the most left column, which contains a few scales for controlling some of the parameters related to the views.

[**SR2**]. Last but not least, the expert can label the users in the detailed view through the drop down menus under the annotation column. In the next few subsections, I will explain each component in detail.

4.5.1.1 Topic Overview

The topics are extracted from the corpus using a topic modeling algorithm, which I will discuss in detail in the next data modeling subsection. The topic overview shows a topic network, where the size of the node represents the relative number of tweets. The color of the node's outline indicates the bot prevalence within a topic after the experts finish the first round of labeling. I used the diverging color palette brown to green (BrBG) spectrum of size 11, from the RColorBrewer library [10], as shown in figure 13. The brown color indicates bots prevalence, and green indicates

[10]https://cran.r-project.org/web/packages/RColorBrewer/index.html

prevalence of other user types. The different effects in-between are represented using the BrBG spectrum. The colors of the node's outlines in figure 12 are the results of incorporating 20% of the users labeled by the experts of the sample they used. The links connecting the topic nodes indicate the topic correlation using a simple thresholding measure, where only edge weights below a certain threshold are truncated. Because the topic network is cluttered with so many topics, I facilitated a topic focus zooming functionality through a dropdown menu in the control panel. If the experts are looking to search for specific keywords among the topics, they can type inside the dropdown menu and the matching topics will appear first and filter out the rest, as shown in figure 12. Also, if the experts doesn't have a particular word to search, the topic zooming function comes in handy when skimming through the list of all topics available. This functionality makes is easier for experts to comprehend all of the topics from a list instead of a cluttered network. The topics are expressed under each node using a list of the highest marginal probability words from the topic model, which I will explain in details in the data modeling section.

4.5.1.2 Topic Proportion & Tweet Similarity Filters

As in [**SR2**], the system is required to enable the experts filter the tweets and find persuasive users who post similar tweets. I provide two filters for the experts in order to navigate through the tweets and make the labeling process more efficient. The topic density plot is the first filter (figure 14), where the expert can select a range of topic proportions by clicking and dragging the cursor on the plot. The second filter is the tweet similarity filter, which is controlled from the control panel on the left as

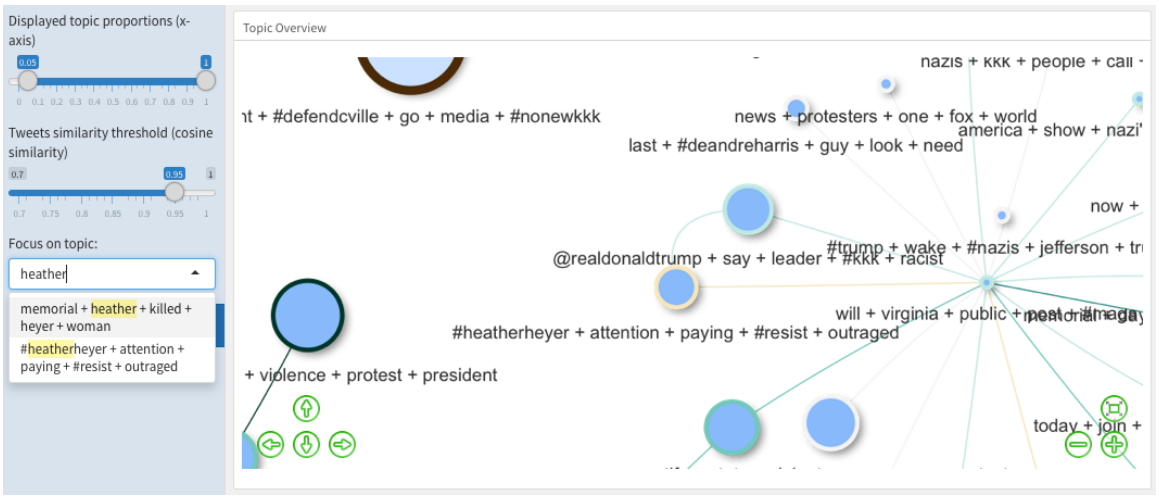

Figure 12: The topic overview provides the option to the expert to select the topics by navigating through a network visualization. The topic focus functionality in action. For example, if the experts search for the word "march" the topics with matching keywords will pop up.

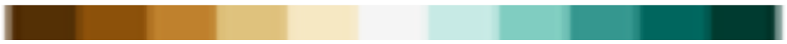

Figure 13: The color palette indicating the bot's effect, where brown indicates bot prevalence and green indicates no bot prevalence.

shown in figure 11. The similarity percentage between the tweets filtered from the topic proportion is calculated using cosine similarity [34]. Similar tweets with cosine similarity more than the threshold set are filtered into the detailed view. In other words, tweets which don't have enough similarity between each other get filtered out of the detailed view.

These two tweet filters provide three advantages for the experts when using the VA system. The first advantage is that the expert will be able to focus on a manageable number of tweets instead of reading large amounts of tweets and users. The second advantage is that they would be able to navigate easily across the spectrum of the topics and their proportions rather than being lost in a big body of corpus. This navigation capability will enable them to easily recall and change the labels of previously visited

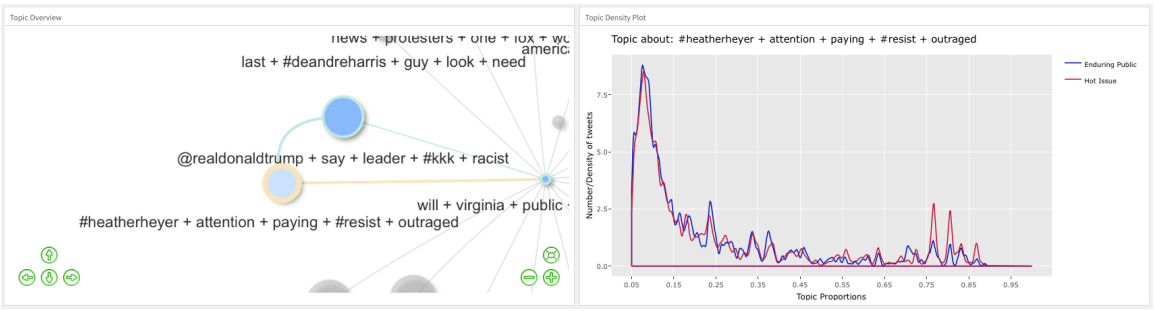

Figure 14: The topic density plot on the right shows the distribution of the tweets among the different topic proportions for hot issue (red) and enduring publics (blue).

tweets. For example, if they want to go back to a certain group of users, they will be able to make a reference to the topic and the range of topic proportions that they visited before to continue labeling or correcting the labels. The third advantage is that it is easier to search for persuasive user behavior. This strategy prioritizes to show the most adversarial users (bots and activists) to be labeled first. This prioritization leads to more accurate labeling performance and thus more accurate prediction model.

4.5.1.3 Detailed View of Individual Tweets

The detailed view displays to the experts the most important metadata in a tabular format to help them decide on the user types. Starting from the most left towards the right, the columns are: the screen name of the user who posted the tweet, tweet's body, account's activity and locality, tweet's source, links inside the tweets are the metadata that the expert use to label the users. These metadata fulfill for the system requirement [**SR3**].

The system meets the requirements [**SR4**] and [**SR5**] by providing the experts a way to incorporate their labeling for active learning. The active learning setting enables the experts to enter the labels for model training in one session, then these

labels are used to train a model to predict the probabilities of the user types for the next session. The user type probability on the second most right column holds the predicted probabilities of the users from the labels of the previous annotation sessions. The most right column contains the dropdown menu of the user types that the domain expert can choose from (i.e. either bot, media, activist, or individual). Once the expert makes a choice from the dropdown menu it gets recorded into a database collection in MongoDB [11]. As shown in figure 15, the users that were labeled from previous annotation sessions are marked as previously labeled under the user type probability instead of showing any predictions and annotation column shows their selected label as "Bot." On the other hand, the unlabeled users have the percentage probabilities displayed per user type, and their annotation column labels as "Unlabeled." When the experts run the prediction model, the new labels are incorporated to train a text classification model and the user type probabilities are updated according to the predictions.

There are some extra features embedded in the detailed view to support the labeling task. The expert doesn't have to worry about searching for the same user within the selected tweets, because the users are grouped together using the "actorId," so the same screen names are grouped together in the table. Another important feature is that the expert can search keywords in each column, which helps in some cases if they want to lookup a specific category within a certain column. The selected range of topic proportions is displayed on top of the table. Also, the expert can download the table in CSV or Excel format in case they want to save this information locally

[11] https://www.mongodb.com/

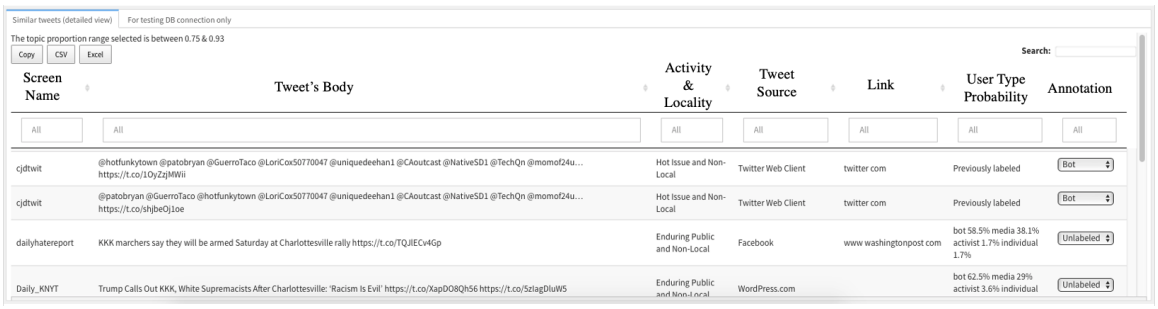

Figure 15: The detailed view shows the important metadata for the experts to decide on the user types, and the annotation column where they can actually input the labels.

on their computer.

4.5.2 Data Modeling

In order to make all of the active learning components available in this VA system, there were two models that had to be tuned first. The two main components of this active learning setting are the topic modeling and user type prediction model. I discuss these two components in this section.

Before getting into the models' details, I will explain the data preprocessing steps and the reasons behind the decisions reached. I conducted a user study to test the system's efficiency as an evaluation for the system's efficiency in terms of design and implementation. In the user study, the subjects are required to label the Twitter users. As an evaluation criteria, the subjects' labels were compared with the expert's labels to measure the speed and accuracy. In addition, for the sake of fairness of the study, the subjects had to be exposed to equal number of users for each user type (i.e. activists, bots, media, individual users). For those two reasons, I filtered out the users who weren't labeled by the experts, and then randomly sampled them down to an equal number of users per user type (1,352 users per user type). The total number of

tweets used, for the user study, were 27,543, tweeted by 5,408 users. While each user can post one or multiple tweets, balancing the number of users is more important than balancing the number of tweets. Thus, the number of tweets are imbalanced which is expected. The dataset contains 3,268; 13,570; 4,849; and 585 tweets posted by activist, bot, individual, and media respectively.

There are few more special preprocessing steps applied to each of the two models separately. For the topic model, I removed the stop words, numbers, symbols, and punctuation. In addition I removed words that appear in less than 5 tweets, and removed tweets with less than 10 words. The words were neither stemmed nor combined to n-grams (only used 1 grams). These steps has lead to the removal of 46 documents in total. Then the corpus was converted to document-frequency matrix format for the convenience of the topic modeling. For the predictive text classification model, I only removed the stop words and numbers, and then removed sparse words with 99% sparsity. The corpus was converted to a document-term matrix format for the convenience of the predictive model. Both models were developed and visualized using R [12].

4.5.2.1 Topic modeling

Topic modeling is one of the best approach in summarizing large collections of tweets based on co-occurrence of words [28]. See more about topic modeling using LDA in [11] and using Structural Topic Model (STM) in [46]. Two of the most important aspects when applying topic modeling on a corpus is the algorithm suitable for the

[12] https://www.r-project.org/about.html

corpus and the number of topics chosen. In this study I used STM because of its ability to estimate the effect of covariates on topic probabilities. This estimation capability provided by the STM enables the estimation of bot prevalence in the next step, which I will discuss in detail. I used the *stm* package in R [13] for its rich functionalities and ease of parameter tuning.

The number of topics is a predetermined parameter that plays a great role in the experts' interpretation of the topics. In practice, the quality of the topics are measured with respect to exclusivity and semantic coherence. Exclusivity measures how much the topics are separable and exclusive within a model. On the other hand, the semantic coherence measures how much the words within a topic are semantically coherent, which is then aggregated per model in order to compare models against each other and pick the best.

My approach towards picking the best number of topics starts with the preliminary selection strategy "Spectral initialization" based on the paper published by Lee and Mimno [42], which is also mentioned in the Roberts et al.'s *stm* paper [46]. Roberts et al. [46] describes this preliminary method as "useful place to start," but warns that it doesn't guarantee to estimate the best number of topics for the randomness introduced in one of the steps. The *stm* function was ran in spectral initialization mode by setting number of topics (k) to zero. The spectral initialization produced a model with 55 topics. When the communications experts interpreted and analyzed the resulting topics from this model, they found that the topics were not coherent enough and felt the need to search for a better model with lower number of topics. They

[13] https://www.rdocumentation.org/packages/stm/versions/1.3.5

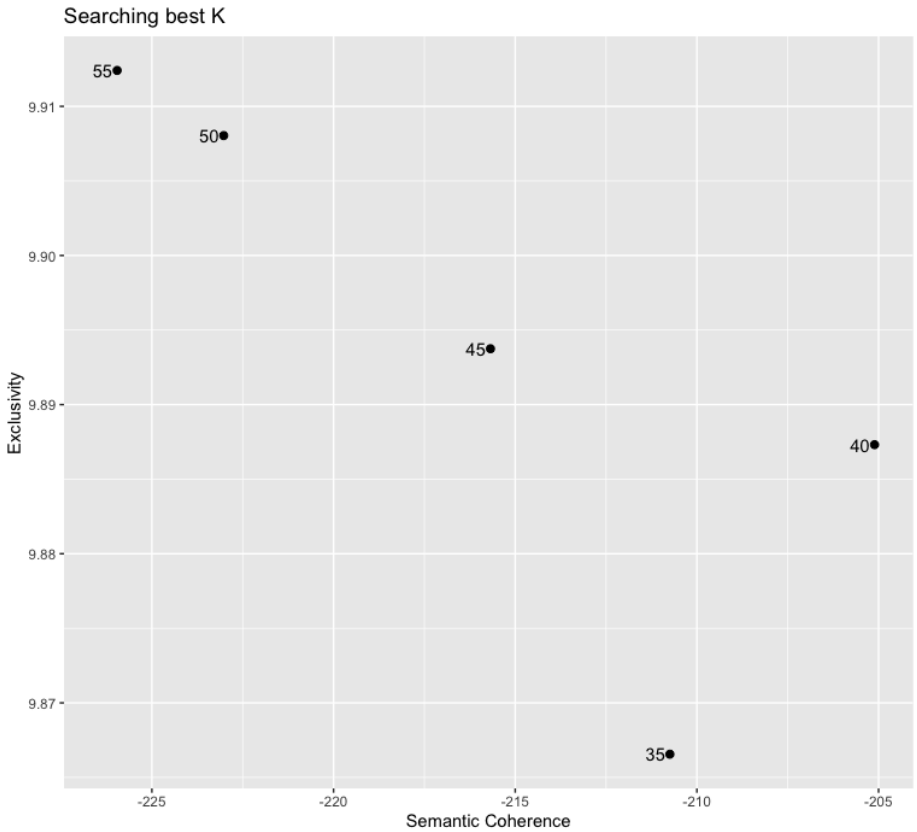

Figure 16: The plot of models' average semantic coherence versus exclusivity.

thought it was a good start but the right number of topics might be a number lower than 55, so I ran STM multiple times while setting k to a different number at each run. I ran the models using the *searchK* function to iterate over k between 35 and 55 with an increment of 5. I plotted the average semantic coherence (x-axis) with respect to the exclusivity (y-axis) of each model in figure 16 in order to compare between the models. Most of the models have relatively close exclusivity, but significantly different in terms of the semantic coherence. Thus, I chose the model with 40 topics, since it is significantly the highest in terms of semantic coherence and has relatively close exclusivity to the other models.

The bot prevalence is part of the feedback that the experts require in [**SR5**]. I used estimateEffect function in the stm package in order to calculate the effect of bots

on the topic probabilities. This means that I have set the user type as a covariate in this equation, and the values are either bot or not-bot (i.e. for other user types). However, that's not the case with the initial model, there are no covariates set. The initial STM model that is used for representing the topics is set without covariates, since there would be no labels provided by the expert initially. This setting means that the *stm* function is set to executing the Correlated Topic Model (CTM), which is the implementation of Blei and Lafferty in 2007 [12], instead of STM. Later on, when the experts start labeling, the STM algorithm is actually ran for the first time by incorporating the bot/not-bot variable as a covariate.

In order to enable this continuous cycle of annotation and modeling, I use the previous model's by-products, the θ (per-document topic distribution) and β (per-topic word distribution) matrices as inputs to the model in the next round and so on. In figure 18 I show the cycle of the topic modeling including the initial round of labeling. This cycle enables one great advantage for domain experts when interpreting topics. The STM does not repeat the initialization process which is non-deterministic. Thus, the topics represented to the experts do not go through significant changes, so that the experts won't get lost or confused between the annotation rounds. This is one of the chapter's novelties which adapts the modeling process for the active learning setting required in the system. Without this adjusted cycle, the domain experts would be confused and lose track of the topics which they previously labeled.

In order to demonstrate how the bot prevalence on topics would be measured by the end of the annotation process, I used all of the communications experts' labels to plot the effect of bots on topic probabilities, as shown in figure 18. I used all

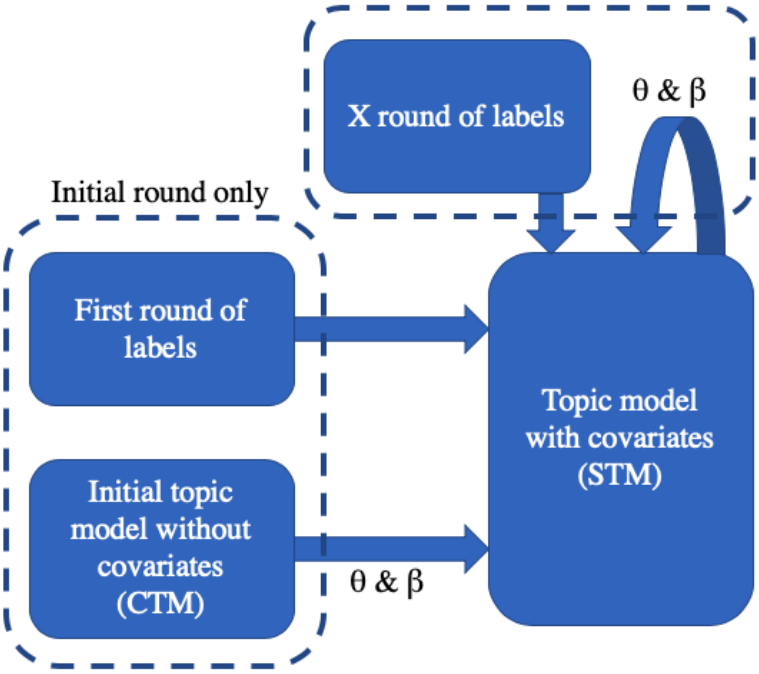

Figure 17: The topic modeling cycle for the first and X rounds of labeling.

of the labels from the experts in the dataset, unlike the user study in section 4.6 where I only considered 20% of the users labeled and the rest are unlabeled. The topics influenced by bots are shifted towards the right side (positive significance), and the topics influenced by other users types are shifted towards the left side (negative significance). The dots and the line segments represent the mean effect of bots on topic probability, and 90% confidence interval of that effect respectively. The x-axis' scale indicates the significance of the bot effect using the positive and negative probabilities.

For instance, in the same figure (18) we can identify 4 topics that are significantly influenced by bots. Their line segments that indicate the 90% confidence intervals are completely on the positive side of the graph. This observation doesn't mean that these topics only contain tweets posted by bots, but it means that they were significantly influenced by them. On the other side of the graph, we can identify 9 topics that

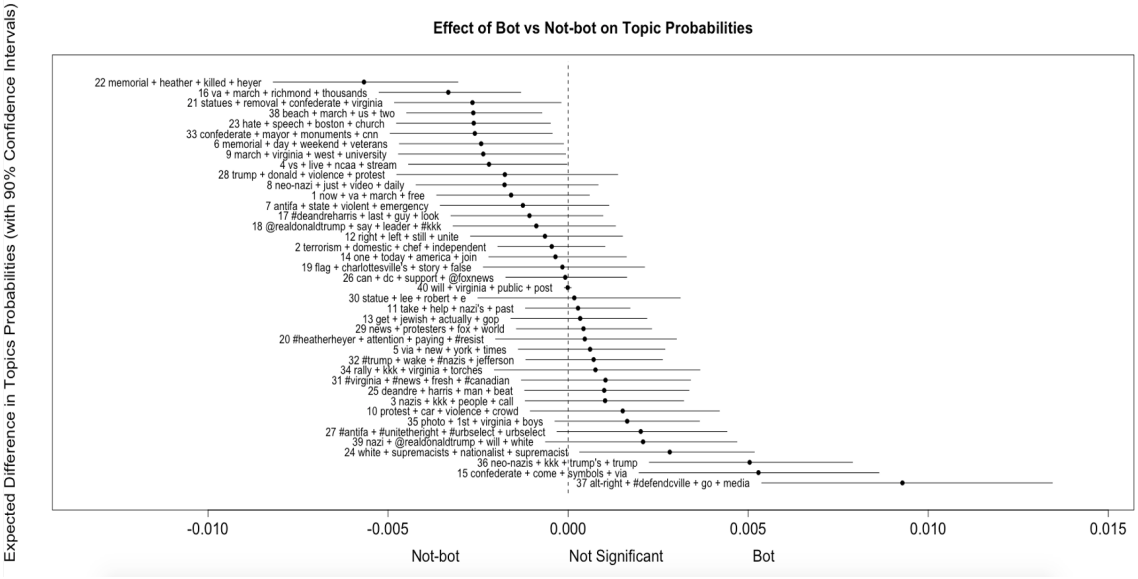

Figure 18: The effect of bots versus other users (non-bots) on topic probabilities.

are significantly influenced by non-bots. Their line segments that indicate the 90% confidence intervals are completely on the negative side of the graph. This observation also doesn't mean that these topics contain no bot-tweets, but it means that they were the least influenced by the presented probabilities on the graph. The rest of the topics have no strong influence from either bots or non-bots.

The experts are able to see this bot prevalence plot by navigating to the "Topic prevalence page" marked at the top of interface as in figure 11. As mentioned before, the colored outlines on the nodes in figure 12 are based on the estimated effects of bots on the topics, which is the alternative direct way to explore estimated bot effects from the topic overview.

4.5.2.2 User type prediction model

The user type prediction model is the second component of the active learning setting. The benefit of this model besides providing the experts with predictions is a

support for their confidence in labeling. The purpose behind these predictions is to give them confidence that would make the labeling process more efficient in terms of accuracy and speed. This hypothesis is confirmed in the user study, which I will discuss in the next section 4.6.

I chose to use the Random Forest (RF) algorithm as a classification algorithm for the reasons mentioned in section 4.2. RF comes in-handy for this particular classification problem addressed here in this chapter. RF is computationally cheaper than BAG-DT and SVM, which is better suited for experts who are going to use in an interface. Multiclass classification, hyperparameters tuning, accommodating feature types, and transforming the output formats are other aspects that proves RF is better than SVM. The domain experts are rarely interested to perform any adjustments to the models. This convenience will allow VA system open to be used for different purposes and datatypes.

I used the *randomForest* [14] and *caret* [15] in R, which facilitates a wide variety of functions for training, prediction and grid search. The *randomForest* function implements Breiman's random forest algorithm. It can be either used as a regressor or classifier with textual data.

The regressor produces the probability of each class for all of the tweets, while the classifier outputs the class that has the highest probability. RF can be used in either modes. I used the *randomForest* function as a classifier first to evaluate the performance using accuracy percentage, confusion matrix, and out of the bag (OOB)

[14]https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest

[15]https://cran.r-project.org/web/packages/caret/vignettes/caret.html

error score. Then I used the regressor to show the predicted probabilities to the experts. RF models need tuning for two parameters only, which are the total number of trees to grow (*ntree*) and the number of variables randomly sampled as candidates at each tree node or split (*mtry*).

First, I performed a grid search to tune the *ntree* and *mtry* parameters using the accuracy as a measure for comparison. I used repeated cross-validation where I split the data into 10-folds and ran 3 repeats per split. Figure 19 shows the accuracy percentage of the models for each combination of the two parameters from 50 to 450 for the *ntree* parameter and 10 to 105 for the *mtry* parameter. I have set the *mtry* to start with square root the number of features (i.e. number of unique words 105) since it is the default value that the *randomForest* package recommends. The grid search graph in figure 19 show that the best performing model of 72.9% accuracy is the one with 250 trees (*ntree*) and 41 variables (*mtry*).

Second, using these parameter values, I built an RF regression model and found that the OOB error rate to be 27.01%. The OOB error is a score that estimates the model's performance on previously unseen data, which is known as the generalization error. The confusion matrix for that model is shown in table 1. The confusion matrix tells us that 11.69% error rate towards predicting activists were mostly false negatives for predicting them as bots. The 14.19% error rate when predicting bots were mostly from predicting them as individuals or media. On the other hand, individual and media users had much higher error rates, 45.08% and 50.42% respectively, and were mostly confused with bots. There are two takeaways from this confusion matrix: (1) the reason behind the high error rates of the media and individual users predictions is

Table 1: Confusion matrix of the RF model using 250 for *ntree* and 41 *mtry*. The rows represent the number of actual values and the columns represent the corresponding number of predicted values.

	activist	bot	individual	media	class error
activist	559	65	3	6	11.69%
bot	54	2353	173	162	14.19%
individual	17	364	525	50	45.08%
media	25	377	192	584	50.42%

because the dataset is imbalanced with respect to number of tweets (not the imbalance in the number of users). The number of tweets posted by bots in the dataset is much higher than the other users. Although the number of users are equal, normally bots tend to post more than other users. (2) the effect of this error rate is not crucial to the main goal which the experts aiming for. The experts are more interested in finding bots and differentiating them from activists. The error in predicting the media and individual users are considered as false negatives from the perspective of predicting bots and activists, which is not as harmful as if it were to be the opposite. Last but not least, since the regression is calculated for each tweet individually, the probabilities are averaged across the same user later when displaying it on the interface.

4.6 User study

In order to evaluate the built VA system I conducted a user study, after getting an IRB approval, to compare it with a traditional tool. I chose MS Excel as the traditional tool to compare with since it is the frontier of traditional tools used by many business analysts for many different purposes. In this user study, I recruited 70 subjects and removed the data points that had errors and pilot experiments, which resulted in 60 subjects. Among these filtered subjects, their age ranged between 18

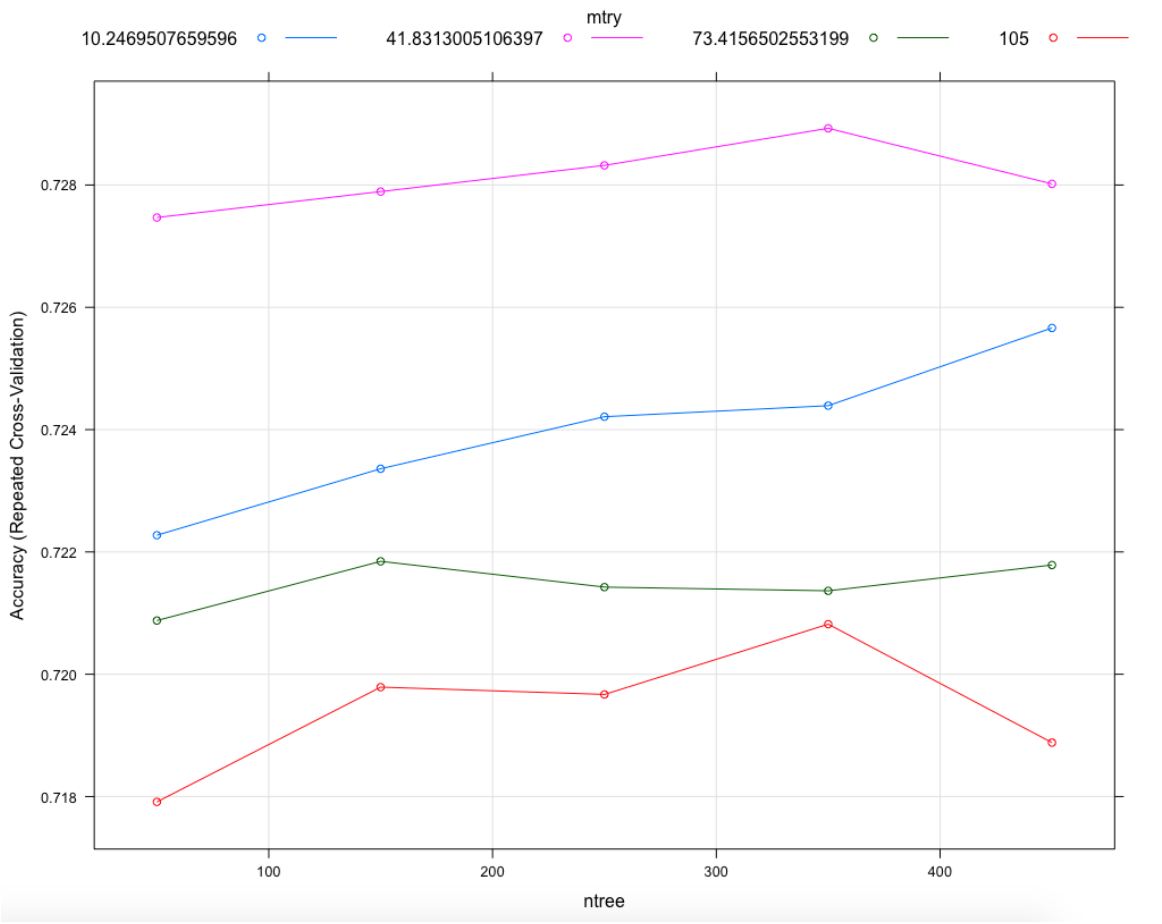

Figure 19: Grid search graph for selecting the best combination of *ntree* and *mtry*.

and 49 (M=24.5, STD=6.8) years old. The participants were 71.6% male and 28.3% female (58.3% undergraduates, 31.7% Ph.D. students, 5% M.S., and other students). Their major categories were distributed as 46.7% STEM, 20% business, 11.7% health and medicine, and 21.6% other categories.

There are two styles of tasks that can be used for the the user study design: either predetermined tasks or exploratory oriented tasks. Since I opened the recruitment for both experts and non-experts, I predetermined the task and restricted the subjects to certain group of tweets to annotate. I didn't limit the recruitment criteria to experts to have a high participation rate, which is needed to obtain results with statistical

Table 2: The chosen topics, their topic proportions selected tweet similarity and number of tweets.

Topic	Topic similarity range	Tweet similarity	Number of tweets
white, supremacists, nationalist, supremacist, protest	0.7 - 0.89	95%	6
trump, donald, violence, protest, business	0.84 - 0.9	95%	96
#heatherheyer, attention, paying, #resist, outraged	0.74 - 0.79	95%	22

significance. In this section, I describe the user study conducted for aspects such as the task, setup, measures, hypotheses, and results.

4.6.1 Experiment Tasks & Apparatus

Task: In this user study, the subjects are asked to label as many users as they can accurately in ten minutes (i.e. four user types: activist, bot, media, individual). Before the labeling task, the subjects filled out a pre-questionnaire, then were trained on either tools (MS Excel or the VA system). Afterwards they start the timed labeling task, and finally answer a post-questionnaire. I prepared three particular topics within certain ranges of topic proportions for each topic, and with a fixed tweet similarity (cosine similarity) threshold. The purpose of this restriction is to make all subjects exposed to the same tweets to prevent bias, as some tweets might be easier to label rather than others. I decided to make the subjects begin with the smallest to the largest topic in terms of number of tweets, so that the subject can switch between topics and use the two filters more before the ten minutes are up. I show the topics chosen and their details in table 2.

Setup: I designed the user study to be a between-subject experiments of three

groups, and labeled them *c0*, *e1*, and *e2*. The first one is control, and the second and third groups are the experimental ones. I designed experiments *e1* and *e2* to emulate the rounds of annotation; the experts would initially start the first round without any previous annotations (*e1*) and rounds further down the road when they have previous annotations incorporated in the models (*e2*). Subjects in *e1* are meant to emulate the experts who would begin with their first round of annotation. While the subjects in *e2* are meant to emulate the experts who already had previous annotation sessions and labeled 20% of the users. These subjects, in *e2*, are not assumed to have experience with the system, but just have the 20% of the labeled users incorporated and the predictions displayed for the unlabeled ones. The accuracy of this model (using 20% instead of 80%) is also around 72% using 350 trees and 40 randomly selected variables used per tree. The difference in the experiment setup between *e1* and *e2* is purposed to find the effect of showing the predictions of the unlabeled users and the actual labels of the previously labeled users. In both experimental groups the subjects followed the normal sequence of steps to filter the tweets using the topic proportions and tweet similarity.

However, MS Excel doesn't have the option of calculating the tweets' topic proportions and cosine similarity. I precalculated these proportions and presented them in different columns for each topic, while the tweet cosine similarities in a separate column. Although the cosine similarity needs to be calculated on the fly upon the selection of topic proportions, as mentioned before (in section), I only presented the cosine similarities for the group of tweets in table 2. The control group subjects had to first search for the topic in the header row and filter by the topic between

the ranges in the table. Then sort the by the tweet similarity to find the ones with 95% similarities. Lastly, the subjects annotated the tweets according to the rules mentioned in the system requirements (section 4.7). I summarize the groups' setups as follows:

- Control (*c0*): using MS Excel to label the users
- Experiment 1 (*e1*): using the VA system to label the users but without the predictions
- Experiment 2 (*e2*): using the VA system to label the users with predictions

Measures: During the timed labeling task, the subjects' labels are recorded and later compared to the communications experts labels (as the ground truth). From there I am able to report three measures of performance to compare between the three subject-groups. The three measures are:

- Speed as a count (#); the number of annotated users
- Accuracy as a count (#); the number of correct labels
- Accuracy as a percentage (%); the percentage of correct labels

4.6.2 Hypotheses

Each of these measures are meant to either reject or fail to reject (accept) a certain hypothesis. The hypotheses' role in this study is to prove the tool's efficiency, which attributes to its usefulness. The tool's usefulness is mainly mentioned in the contributions of this chapter (section). In this user study I measure three main

hypotheses, which account to two main themes. The first theme is that the VA system helps the subjects label the users more efficiently in terms of speed and accuracy. The second theme is that adding the predictions will improve the subject's performance, which is part of the active learning setting. I summarize the three main hypotheses and their sub-hypotheses as follows:

1. Subjects in *e1* and *e2* will be faster than c0 (H1)
   (a) Subjects in *e1* will be faster than *c0* (H1.1)
   (b) Subjects in *e2* will be faster than *c0* (H1.2)
2. Subjects in *e1* and *e2* will be more accurate than *c0* (H2)
   (a) Subjects in *e1* will be more accurate than *c0* (H2.1)
   (b) Subjects in *e2* will be more accurate than *c0* (H2.2)
3. Subjects in *e2* will be more accurate than *e1* (H3)

4.6.3 Results

In order to apply the appropriate statistical significance test when comparing between populations, first I needed to know whether the measured speed and accuracies were following the normal (Gaussian) distribution or not. In order to answer this quesion, I applied three normality tests: Shapiro-Wilk test, D'Agostino's test, and Anderson test. I applied the Mann-Whitney U test whenever there is any measure with non-normal distribution in the comparison equation, and Welche's t-test whenever the measures on both sides of the comparison are normally distributed. I chose Welche's

t-test over the vanilla t-test, since all of the variances are different. I found the following:

- Speed as counts: *c0*: does not look normal, *e1*: looks normal, *e2*: looks normal
- Accuracy as counts: *c0*: looks normal, *e1*: looks normal, *e2*: looks normal
- Accuracy as percentages: *c0*: does not look normal, *e1*: looks normal, *e2*: looks normal

In table 3 shows a summary of the means (M) and standard deviations (STD) of each measure per subject-group. Table 4 shows the comparisons between the subject-groups and the statistical significance of these comparisons (i.e. reject or fail to reject hypothesis). Also, each cell mentions which hypothesis can be supported or not with color coding. In addition to these two tables, the speed and accuracy measures are visualized in figures 20, 21, and 22 using box-and-swarm plots for the speed count (figure 20), accuracy as a count (figure 21), and accuracy as a percentage (figure 22). In a nutshell, hypotheses H1.1 and H1.2 were confirmed using the speed measure when comparing *c0*, *e1* and *e2*. The accuracy measure as counts confirmed H2.1 and H2.2, but not H3. On the other hand, the accuracy measure as a percentage confirmed H2.2 and H3, but not H2.1. Therefore, all of the hypotheses mentioned were satisfied.

4.7 Discussion & Future Work

In summary to this chapter, I aim to help communications experts annotate social media users efficiently. The experts' annotations enabled them to remove the bots and

Table 3: The means (M) and standard deviations (STD) for each of the measures (speed as a count, accuracy as a count and accuracy as a percentage) for each of the subject-groups (*c0*, *e1*, *e2*).

Group code	Speed as a count (#) comparison for H1	Accuracy as a count (#) comparison for H2 & H3	Accuracy as a percentage (%) comparison for H2 & H3
c0	M = 28.9, STD = 16.00	M = 11.10, STD = 6.81	M =38.97%, STD = 19.93%
e1	M = 53.95, STD = 23.48	M = 20.2, STD = 10.22	M = 38.78%, STD = 16.38%
e2	M = 47.45, STD = 21.87	M = 25.9, STD = 12.45	M = 54.65%, STD = 9.25%

Table 4: The comparison between the subject-groups and their statistical significance in terms of the three measures. The cells are color-coded according to the status of significance towards the hypotheses as follows: green as fail to reject, grey as no significance (neither rejects nor accepts the hypothesis), and red as reject. Double asterisks (**) means the highly significant with p-value <0.05, and single asterisk (*) means the moderately significant with p-value <0.1.

Group code	Speed as a count (#) comparison for H1	Accuracy as a count (#) comparison for H2 & H3	Accuracy as a percentage (%) comparison for H2 & H3
c0 & *e1*	*c0* <*e1*** fails to reject (accepts) H1.1	*c0* <*e1*** fails to reject (accepts) H2.1	*c0* <*e1* no significance to H2.1
c0 & *e2*	*c0* <*e2*** fails to reject (accepts) H1.2	*c0* <*e2*** fails to reject (accepts) H2.2	*c0* <*e2*** fails to reject (accepts) H2.2
e1 & *e2*	*e1* >*e2* no significance and no hypothesis originally	*e1* <*e2* no significance to H3	*e1* <*e2*** fails to reject (accepts) H3

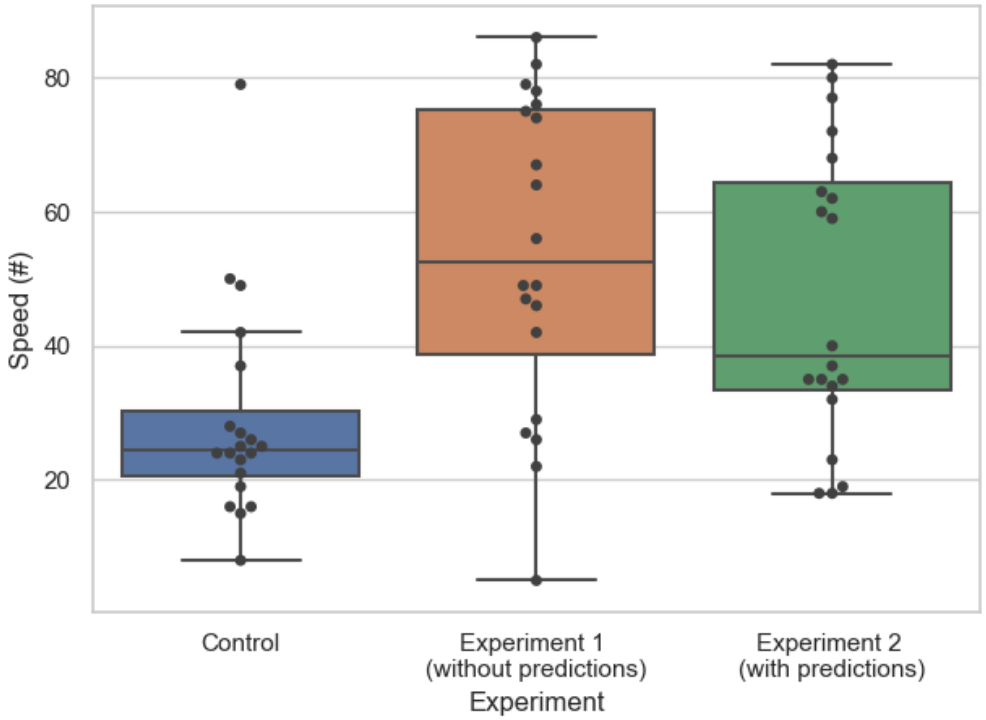

Figure 20: The box and swarm plots of speed measured as a count of correct labels.

rerun their models without the noise that used to clutter their analysis. This goal aligns well with my dissertation statement mentioned in the introduction; where I attempt to emphasize the importance combining interactive visualizations with predictive models to study user's behavior on social media. My attempt to support the dissertation statement was the motivation behind documenting the system requirements, which gave me directions for developing the VA system.

At the end of the chapter I conducted a user study to evaluate the usefulness of the tool, which is linked to confirming the hypotheses tested. I compared between the VA system and MS Excel by measuring the speed and accuracy of the user study's subjects when annotating the social media users as the evaluation measures. In addition, I tested the effect of providing predictions to the subjects as part of evaluating the

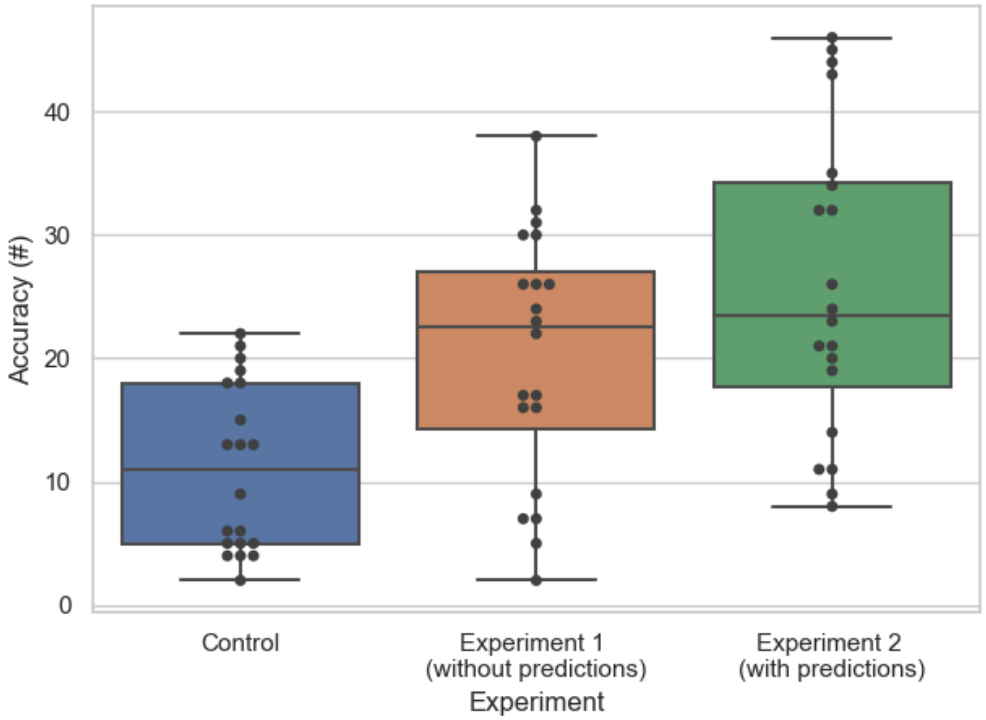

Figure 21: The box and swarm plots of accuracy measured as a count of correct labels.

active learning settings. The hypotheses were shown to be accepted with statistical significance, where the subjects using the VA system were faster and more accurate than subjects using MS Excel. Also, the VA system supported by the prediction model was proven to improve the accuracy of the subjects.

The VA system can also be applied to different datasets and applications. It is not only restricted to this Twitter dataset, and also applicable to any textual data with the same characteristics. Applying other datasets or applications to the system is possible by adjusting the topic modeling and text classification parameters, and identifying the useful metadata. The topic modeling and text classification algorithms applied is designed for corpora that has documents with only one tag or label per document. This system can also be used in applications or purposes other than social

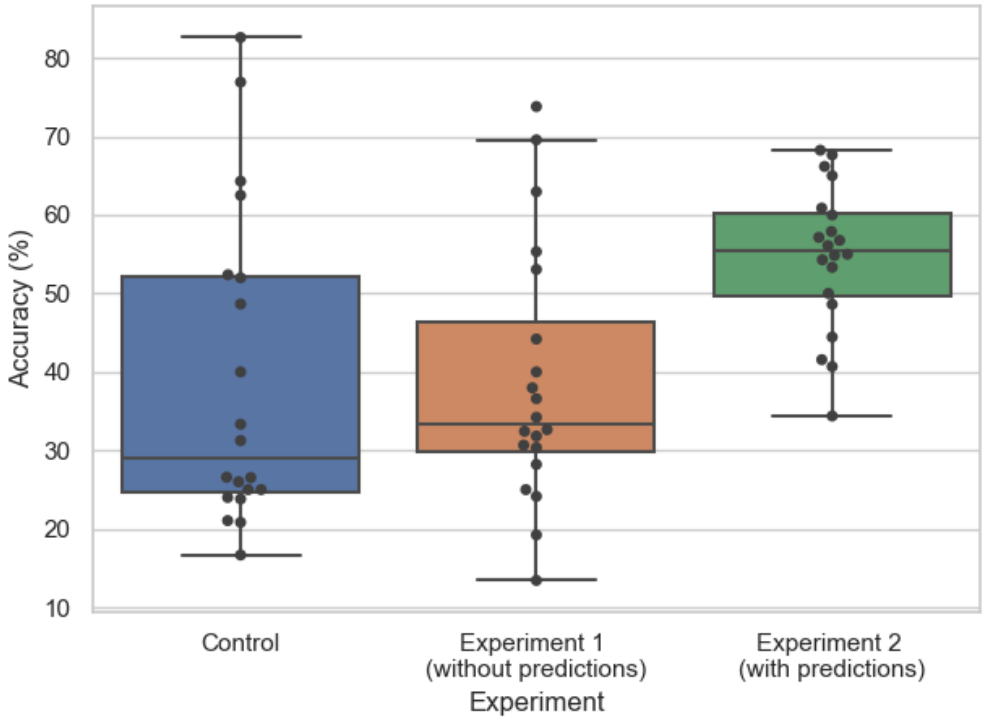

Figure 22: The box and swarm plots of accuracy measured as a percentage of correct labels.

media labeling, the application just needs to follow the same pattern of requirements. For example, the VA system can be modified to be used in the legal domain, business development, market research, interventions, and advertising industries. One possible feature that can enable its use in different applications is having a home page for domain experts to enter their predefined labels, which would appear later in the dropdown menus of the annotation column.

One of the prospected future work I would like to implement is enabling the domain experts to collaborate online through a crowdsourcing platform. The platform could enable multiple experts to work together on the same corpus by dividing the labeling efforts amongst them. Another prospects of the future work are the improvements that can be done on the system in order to attract more domain experts to use

it. The first aspect is the improvement of the backend models such as the text classification model. In order for the experts to trust this system, they need to know that the model is reliable. Thus, I will use area under the curve (AUC) instead of absolute accuracy percentage to report the predictions, and use sampling techniques to overcome imbalanced datasets.

Another future work aspect I would like to add is giving the domain experts control to change the number of topics and the topic names as part of the annotation process. This feature would give them the ability to adjust the number of topics according to their point of view. Also, part of my future work is to use hierarchical topic modeling, so that the experts can summarize a large number of topics when needed. One of the challenges that the domain experts faced with this system is the enormous number of topics that need to be analyzed. To apply this idea on the system, the topic overview would show the topic hierarchy using collapsible tree layout instead of topic network. This would enable the experts to view the topics from high level first, and then delve into more detailed ones. In addition I would like to evaluate the difference between the topic words generated from different iterations of the topic model cycle. When the difference between the words of the same topic from different iteration is big, the domain experts could get confused. The results of this evaluation depends on many factors such as the size of corpus (number of documents), size vocab (number of unique words), word frequency distribution, and the semantic coherence of the topics themselves.

The other type of future work which I would like to work on are visual features that would improve the experts' perception of the data for easier and quicker understanding.

For example, representing the prediction probabilities as barcharts instead of numbers only. Another one is about swapping the columns in the detailed view to make the annotation column next to the screen name.

CHAPTER 5: CONCLUSIONS

In conclusion, the advancements in Artificial Intelligence (AI) require both, transferring the domain expert's knowledge to machines and simplifying the analysis for experts. Efforts in both of these directions address the research gaps in the fields of AI. VA systems play a major role in progressing these efforts. They enhance various analytics tasks for domain experts to achieve the technological progress needed. This dissertation demonstrates how customized VA systems can support domain experts to perform their tasks more efficiently than when using manual traditional tools or other platforms. The dissertation is composed of two main projects, each aiming to enhance two different analysis tasks. The first enhances making sense of the connection between user cohorts and their generated online content. The project's application is meant for business developers and market researchers who want to understand the topics of interests for different demographics groups on Reddit. The second project supports the labeling task of users, and the application is to differentiate bots from other users and remove them from the dataset. One of the most important aspects in this dissertation is not only how to design the interface to support these tasks, but also how to incorporate machine learning models to enhance this support.

The design aspects of incorporating prediction models into the VA system has brought up six main research questions that were applied to both projects. The first three questions (A1 - A3) are related to the analytics side of the system, and the

second three questions are related to the interface part (V1 - V3). These research questions were addressed through the development of visual and analytics features, and then the systems were evaluated to prove the efficiency of those features. The first set of analytics research questions wonder about the feasibility to predict predefined user types and cohorts, the power of different predictive features, and the ability to raise awareness for using those powerful predictive features. The second set of questions regarding the visualization component set an example of how to present computational features and aggregate model results and details of the users' posts.

With models that reach 89.81% accuracy for predicting gender and 85.81% for predicting bots, we can see that it is feasible to create models to detect predefined user types on social media. In addition, I proved the ability to extract the powerful features and incorporate the domain experts into an active learning setting by devising the appropriate customized visualizations. Also, given the VA systems evaluated in both user studies, we can see how the visual interfaces were more efficient when compared to SAS and MS Excel in both, presenting the models' results and aggregating details of the users' posts. The user studies show significant improvements when using the VA systems.

In the future, my main efforts will be in the direction of building more VA systems that address problems for domain experts in other fields besides market research and social sciences such as advertisements, interventions, and law firms in legal systems. In these fields, experts also use textual data and are in need of natural language processing and topic modeling to summarize and interpret their data efficiently. One of my important recommendations for other researchers in the VA field is to focus on

proving the concept that addresses the system requirements first, and then afterwards transition to scaling performance and improving the aesthetics. Lastly, there is one limitation that I have not mentioned in both user studies. Since human-computer interaction is an important part of VA systems, psychologists' perspective in decision making is an important angle in user studies. The influence of decision making strategies makes a difference in measuring the usefulness of the VA systems.

REFERENCES

[1] N. Abokhodair, D. Yoo, and D. W. McDonald. Dissecting a social botnet: Growth, content and influence in twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 839--851. ACM, 2015.

[2] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo. People are strange when you're a stranger: Impact and influence of bots on social networks. *arXiv preprint arXiv:1407.8134*, 2014.

[3] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 173--182, Oct 2014.

[4] L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi. Sok: The evolution of sybil defense via social networks. In *Security and Privacy (SP), 2013 IEEE Symposium on*, pages 382--396. IEEE, 2013.

[5] S. Amershi, J. Fogarty, and D. Weld. Interactive machine learning for on-demand group creation in social networks'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 21--30, 2012.

[6] D. Arendt, E. Grace, and S. Volkova. Interactive machine learning at scale with chissl. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2018.

[7] D. Arendt, C. Komurlu, and L. M. Blaha. Chissl: A human-machine collaboration space for unsupervised learning. In *International Conference on Augmented Cognition*, pages 429--448. Springer, 2017.

[8] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9), 2007.

[9] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22nd international conference on World Wide Web*, pages 131--140. International World Wide Web Conferences Steering Committee, 2013.

[10] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 11. ACM, 2004.

[11] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77--84, 2012.

[12] D. M. Blei, J. D. Lafferty, et al. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17--35, 2007.

[13] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th annual computer security applications conference*, pages 93--102. ACM, 2011.

[14] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. Design and analysis of a social botnet. *Computer Networks*, 57(2):556--578, 2013.

[15] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301--1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] Q. Cao, X. Yang, J. Yu, and C. Palow. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 477--488. ACM, 2014.

[17] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161--168. ACM, 2006.

[18] J. Choo, C. Lee, C. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):1992--2001, Dec 2013.

[19] C. Chung and J. W. Pennebaker. The psychological functions of function words. *Social communication*, pages 343--359, 2007.

[20] W. Cui, S. Liu, Z. Wu, and H. Wei. How hierarchical topics evolve in large text corpora. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2281--2290, Dec 2014.

[21] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273--274. International World Wide Web Conferences Steering Committee, 2016.

[22] J. P. Dickerson, V. Kagan, and V. Subrahmanian. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 620--627. IEEE Press, 2014.

[23] W. Dou, I. Cho, O. ElTayeby, J. Choo, X. Wang, and W. Ribarsky. Demographicvis: Analyzing demographic information based on user generated content. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 57--64. IEEE, 2015.

[24] W. Dou, X. Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 231--240, Oct 2011.

[25] D. Elliott. Crowd gazing - understanding demographic forces can help us better prepare for the problems caused by the world's rapidly expanding population. Accessed: 2015-03-30.

[26] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133--3181, 2014.

[27] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96--104, 2016.

[28] T. D. Gallicano, R. Wesslen, and J.-C. Thill. From cluster tweets to retweets: A big data, rhetorical exploration of digital social advocacy in the context of the charlotte protests on twitter. In *20TH INTERNATIONAL PUBLIC RELATIONS RESEARCH CONFERENCE*, page 75, 2017.

[29] A. Gupte, S. Joshi, P. Gadgul, A. Kadam, and A. Gupte. Comparative study of classification algorithms used in sentiment analysis. *International Journal of Computer Science and Information Technologies*, 5(5):6261--6264, 2014.

[30] M. N. Hajli. A study of the impact of social media on consumers. *International Journal of Market Research*, 56(3):387--404, 2014.

[31] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83--85, 2005.

[32] G. Heidemann, D. Weiskopf, M. Hoferlin, R. Netzel, and B. Hoferlin. Inter-active learning of ad-hoc classifiers for video visual analytics. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 23--32.

[33] A. Hermida, F. Fletcher, D. Korell, and D. Logan. Share, like, recommend. *Journalism Studies*, 13(5-6):815--824, 2012.

[34] A. Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9--56, 2008.

[35] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander. Large scale personality classification of bloggers. In *Affective Computing and Intelligent Interaction*, pages 568--577. Springer, 2011.

[36] S. M. Kabir. *METHODS OF DATA COLLECTION*, pages 201--275. 07 2016.

[37] A. Karduni, I. Cho, R. Wesslen, S. Santhanam, S. Volkova, D. Arendt, S. Shaikh, and W. Dou. Vulnerable to misinformation? verifi! *arXiv preprint arXiv:1807.09739*, 2018.

[38] A. Karduni, R. Wesslen, S. Santhanam, I. Cho, S. Volkova, D. Arendt, S. Shaikh, and W. Dou. Can you verifi this? studying uncertainty and decision-making about misinformation using visual analytics. 2018.

[39] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285--318, 1988.

[40] Z. Ma, W. Dou, X. Wang, and S. Akella. Tag-latent dirichlet allocation: Understanding hashtags and their relationships. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 260--267. IEEE, 2013.

[41] H. McDonald and S. Adam. A comparison of online and postal data collection methods in marketing research. *Marketing intelligence & planning*, 2003.

[42] D. Mimno and M. Lee. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1319--1328, 2014.

[43] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211--236, 2008.

[44] J. W. Pennebaker and L. D. Stone. Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85(2):291, 2003.

[45] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37--44. ACM, 2010.

[46] M. E. Roberts, B. M. Stewart, D. Tingley, et al. stm: R package for structural topic models. *Journal of Statistical Software*, 10(2):1--40, 2014.

[47] SAS. Sas text miner. http://www.sas.com/textminer. Accessed: 2015-03-30.

[48] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199--205, 2006.

[49] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[50] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[51] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745--750. International World Wide Web Conferences Steering Committee, 2016.

[52] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 373--382. ACM, 2010.

[53] S. K. Sikdar, B. Kang, J. O'Donovan, T. Hollerer, and S. Adal. Cutting through the noise: Defining ground truth in information credibility on twitter. *Human*, 2(3):151--167, 2013.

[54] T. O. Sprenger, P. G. Sandner, A. Tumasjan, and I. M. Welpe. News or noise? using twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting*, 41(7-8):791--830, 2014.

[55] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):319, 2008.

[56] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger. Social media analytics – challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39:156 -- 168, 2018.

[57] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[58] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts (# MSM2012)*, 2(4):1951--1959, 2012.

[59] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner. Predicting susceptibility to social bots on twitter. In *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)*, pages 6--13. IEEE, 2013.

[60] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner. Which users reply to and interact with twitter social bots? In *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*, pages 135--144. IEEE, 2013.

[61] Y.-C. Wang, M. Burke, and R. E. Kraut. Gender, topic, and audience response: an analysis of user-generated content on facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 31--34. ACM, 2013.

[62] A. Wilkie, M. Michael, and M. Plummer-Fernandez. Speculative method and twitter: Bots, energy and three conceptual characters. *The Sociological Review*, 63(1):79--101, 2015.

[63] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1763--1772, Dec 2014.