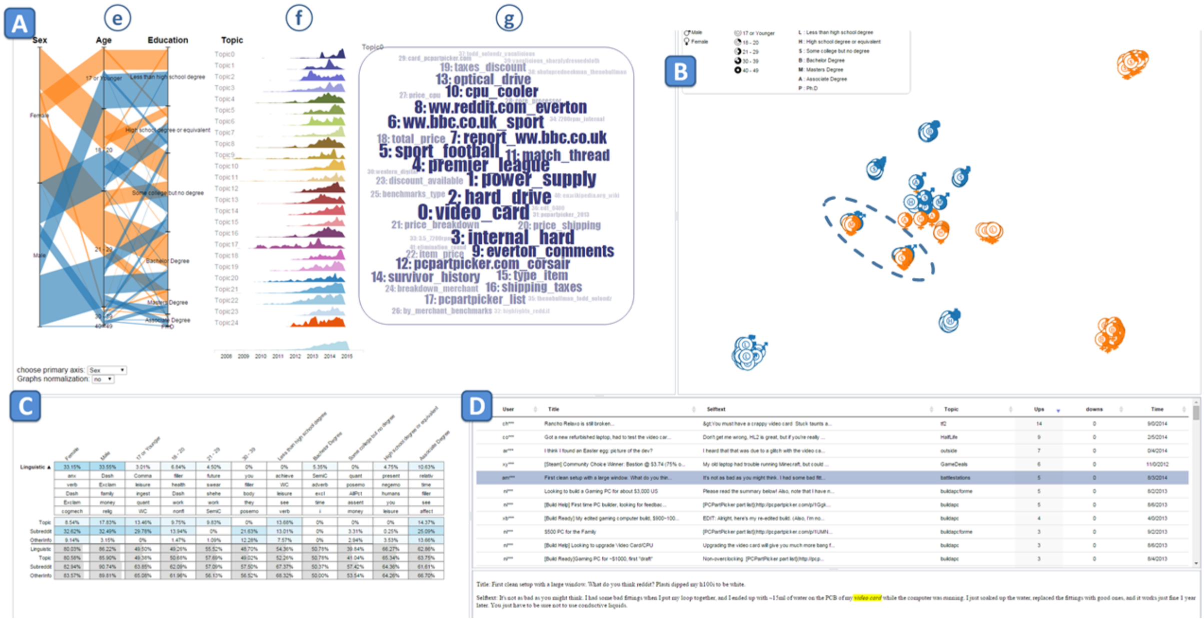
DemographicVis: Analyzing Demographic Information based on User Generated Content

Wenwen Dou[*] UNC Charlotte
Isaac Cho[†] UNC Charlotte
Omar ElTayeby UNC Charlotte
Jaegul Choo Korea University
Xiaoyu Wang Taste Analytics
William Ribarsky UNC Charlotte



choose primary axis: Sex ▾
Graphs normalization: no ▾

	Female	Male	17 or Younger	18 - 20	21 - 29	30 - 39	Less than high school degree	Bachelor Degree	Some college but no degree	High school degree or equivalent	Associate Degree
Linguistic ▲	33.15%	33.55%	3.01%	6.84%	4.50%	0%	0%	5.35%	0%	4.75%	10.63%
	anx	Dash	Comma	filler	future	you	achieve	SemiC	quant	present	relativ
	verb	Exclam	leisure	health	swear	filler	WC	adverb	posemo	negemo	time
	Dash	family	ingest	Dash	shehe	body	leisure	excl	AllPct	humans	filler
	Exclam	money	quant	work	work	they	see	time	assent	you	see
	cogmech	relig	WC	nonfl	SemiC	posemo	verb	i	money	leisure	affect
Topic	8.54%	17.83%	13.46%	9.75%	9.83%	0%	13.68%	0%	0%	0%	14.37%
Subreddit	32.62%	32.49%	29.78%	13.94%	0%	21.63%	13.01%	0%	3.31%	0.25%	25.09%
OtherInfo	9.14%	3.15%	0%	1.47%	1.09%	12.28%	7.57%	0%	2.94%	3.53%	13.66%
Linguistic	80.03%	86.22%	49.50%	49.26%	55.52%	48.70%	54.36%	50.78%	39.84%	66.27%	62.86%
Topic	80.58%	85.90%	49.36%	50.68%	57.69%	49.02%	52.26%	50.78%	41.04%	65.34%	63.75%
Subreddit	82.94%	90.74%	63.85%	62.09%	57.09%	57.50%	67.37%	50.37%	57.42%	64.36%	61.61%
OtherInfo	83.57%	89.81%	65.08%	61.96%	56.13%	56.52%	68.32%	50.00%	53.54%	64.26%	66.70%

User	Title	Selftext	Topic	Ups	downs	Time
ch***	Rancho Relaxo is still broken...	&gt;You must have a crappy video card Stuck taunts a...	tf2	14	0	9/0/2014
co***	Got a new refurbished laptop, had to test the video car...	Don't get me wrong, HL2 is great, but if you're really ...	HalfLife	9	0	2/5/2014
ar***	I think I found an Easter egg: picture of the dev?	I heard that that was due to a glitch with the video ca...	outside	7	0	0/4/2014
xy***	[Steam] Community Choice Winner: Bastion @ $3.74 (75% o...	My old laptop had trouble running Minecraft, but could ...	GameDeals	6	0	11/0/2012
am***	First clean setup with a large window. What do you thin...	It's not as bad as you might think. I had some bad fit...	battlestations	5	0	8/3/2014
n/***	Looking to build a Gaming PC for about $3,000 US	Please read the summary below! Also, note that I have n...	buildapcforme	5	0	8/2/2013
n/***	[Build Help] First time PC builder, looking for feedbac...	[PCPartPicker part list](http://pcpartpicker.com/p/1Ggk...	buildapc	5	0	8/6/2013
xb***	[Build Ready] My edited gaming computer build, $900~100...	EDIT: Alright, here's my re-edited build. (Also, I'm no...	buildapc	4	0	4/0/2013
n/***	$500 PC for the Family	[PCPartPicker part list](http://pcpartpicker.com/p/1UMN...	buildapcforme	3	0	9/2/2013
n/***	[Build Help] Looking to upgrade Video Card/CPU	Upgrading the video card will give you much more bang f...	buildapc	3	0	8/6/2013
n/***	[Build Ready]Gaming PC for ~$1000, first "draft"	Non-overclocking. [PCPartPicker part list](http://pcp...	buildapc	3	0	8/4/2013

Title: First clean setup with a large window. What do you think reddit? Plasti dipped my h100i to be white.

Selftext: It's not as bad as you might think. I had some bad fittings when I put my loop together, and I ended up with ~15ml of water on the PCB of my video card while the computer was running. I just soaked up the water, replaced the fittings with good ones, and it works just fine 1 year later. You just have to be sure not to use conductive liquids.

Figure 1: 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.

Abstract

The wide-spread of social media provides unprecedented sources of written language that can be used to model and infer online demographics. In this paper, we introduce a novel visual text analytics system, DemographicVis, to aid interactive analysis of such demographic information based on user-generated content. Our 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 our modeling process that connects the demographic groups with features that best describe the distinguishing characteristics of each group. Features including topical and linguistic 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, we 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. We present multiple case studies to showcase the utility of our visual analytics approach in exploring and understanding the interests of different demographic groups. We also 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.

Keywords: Visual Text Analysis, User Interface, Social Media, Demographic Analysis

Index Terms: H.5.2 [Information interface and presentation (e.g., HCI)]: User Interfaces—Graphical user interfaces (GUI)

1 Introduction

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 [9]. 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

[*] e-mail: wdou1@uncc.edu
[†] e-mail: icho1@uncc.edu

IEEE Conference on Visual Analytics Science and Technology 2015
October 25–30, Chicago, Il, USA
978-1-4673-9783-4/15/$31.00 ©2015 IEEE

phishing susceptibility among different demographics groups has identified several factors that influence users' online behaviors [21].

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 [20, 23, 3] and Twitter [17, 4]. 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 [2, 6, 15]. 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 [20, 4], 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 paper, we offer DemographicVis[1], a visual analytics approach to demographic analysis that combines the merits of the above two categories of research, in that we take a data-driven perspective and we establish a direct link between demographic groups and meaningful, easy-to-interpret features. More importantly, we 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 [8, 24, 7, 1], 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 papers 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.

2 Related Work

Demographic analysis has been an important research domain. 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.

2.1 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. [6] and Argomon et al. [2] 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 [15], 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. [16] 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. [19] 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.

We see our visual analytics approach as complimentary to the linguistic studies. Through coupling the semantically meaningful topics and relationships between demographic groups identified in our 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 Classification models for predicting age and gender

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. [4] experimented with Support Vector Machines, Naive Bayes, and Balanced Winnow2 [13] 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. [17] introduced stacked-SVM-based classification algorithms over a set of features to classify gender, age, regional origin and political orientation,

[1] http://demographicvis.uncc.edu

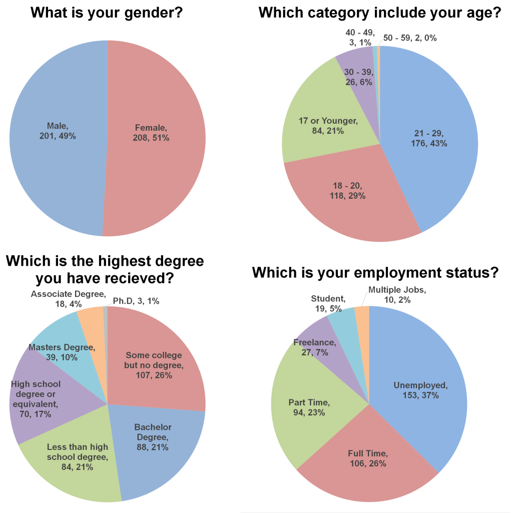


Figure 2: Demographic distribution on gender, age, education, and employment status based on the collected data.

while Schler et al. [19] 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.

3 Data Collection

3.1 Demographic information collection

To collect demographic information from online users, we 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, we first compiled a set of multiple choice questions. The subreddit that allowed us to post the survey is r/SampleSize. This community is dedicated to generating and answering surveys. In our survey, we 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, etc. 482 users participated in our survey, 409 users were included in the final data collection after filtering based on the control questions. Figure 2 presents the summary of information all 409 responses. The summary suggests that our 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 more readily available. However, in our study, we collected more detailed demographic information well beyond age and gender. Although a sample of 409 redditors may not provide sufficient statistical power for generalizing our findings to broader contexts, our visual analytics approach to demographic analysis can be applied to information collected from a much larger population.

3.2 Collection of user-generated content

The objective of our research is to connect demographic information with the content the users posted on social media through visual analytics means. To this aim, we collected the posts from the 409 valid users of our survey. We developed a python crawler to gather the posts of this group of users through Reddit's public API. 169,707 posts were included in the final dataset, 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.

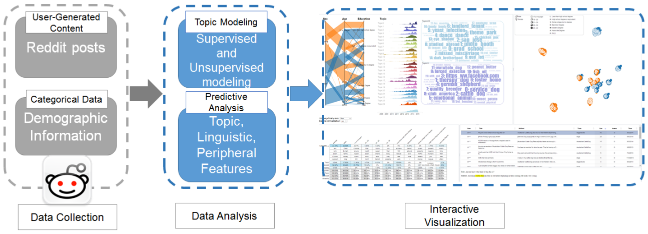


Figure 3: System architecture of DemographicVis. Section 3, 4, and 5 introduce the Data Collection, Data Analysis, and Interactive Visualization component respectively.

In our 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, we 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 our 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}.

4 DemographicVis: A Visual Analytics Approach for Demographic Analysis

Our 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 3. In this section, we focus on introducing the "Data Analysis" component; the "Interactive Visualization" will be described in the next section.

4.1 Data modeling and feature analysis

To describe different demographic groups based on user generated content, we extract features from the reddit posts, including linguistic and topic features. We 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.

4.1.1 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, we perform supervised topic modeling to extract topics for each demographic group. Since the direct relationship between topics and the demographic group is essential to our objective, we want to establish direct links during the modeling process. To incorporate demographic information directly into the topic extraction process, we adopt the Tag-LDA model [14] that was designed to include tags or labels of each document during the topic modeling process. In our approach, each multi-attribute demographic group serves as a tag for each comment a user posted on Reddit. 36 unique multi-attribute demographic groups are found in our data. As a result, we can now use the topic results to describe the interests of each demographic group. For instance, as seen in figure 4, the two topics

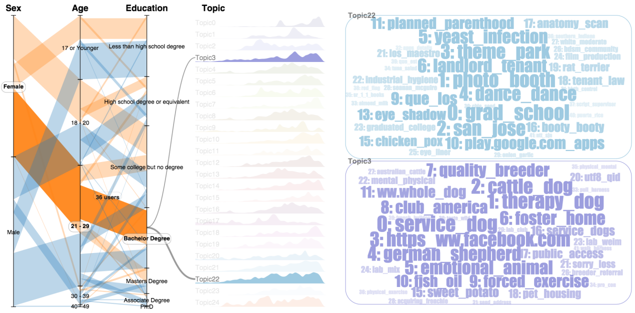


Figure 4: Hovering mouse over demographic group "Female, 21-29, Bachelor degree" leads to the highlighting of the two topics that summarize the interests of the group.

that best describe multi-attribute group "Female 21-29 with Bachelor degree" are Topic 22, which includes keywords "grad_school, study_abroad, theme_park", and Topic3 that focuses on discussing dogs and puppies.

When performing the topic extraction, we employ a bigram approach that treats two consecutive words in a document as a unit of analysis. According to [11], bigrams serve as better feature representations compared to unigrams. Employing bigrams during topic modeling enables us to discover more phrases with richer meaning such as "birth control" and "medical marijuana".

4.1.2 Topic, linguistic, and peripheral 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, we extract a set features and further analyze which feature or combination of features can be used to best predict 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, we describe features we 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.

Our 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, we 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, we will introduce how we derive the features for predictive analysis.

Topic proportions. To obtain the topic proportions, we first construct a term-document matrix using a bag-of-words model based on all available reddit posts. Next, we perform topic modeling using a method called nonnegative matrix factorization [5] with the total number of topics set at 100. We then obtain a 100-dimensional topic-wise vector representing each reddit post. Next, for each user, we sum up these 100-dimension topic-wise vectors of all the reddit posts a user has written, and this aggregated vector works as our topic proportion feature for a particular reddit user.

Linguistic feature. To extract linguistic features from user generated content, we 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[2]. 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. We 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.

4.1.3 Predictive analysis

We obtain a 4,483-dimensional vector for each Reddit user. Then we build a binary classifier by using two groups at a time: one containing users in a particular demographic 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 our data. To this aim, we use a gradient boosting tree (GBtree) [10][3], 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 1C.

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 our evaluation measure rather than a simpler measure such as the prediction accuracy is because our 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 we only incorporate particular features. That is, we 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 we gradually incorporate more features. For example, the first row shows the AUC values only 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. Our 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.

5 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, we 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

[2] http://www.liwc.net/descriptiontable1.php

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

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, we introduce an interactive visual interface that consists of the following three views.

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

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

5.1.1 Parallel Sets for demographic data

Parallel Sets is designed for visualizing relationships between dimensions in categorical data. We applied it to three demographic dimensions: gender, age and educational level. In contrast to the original ParSets layout, we 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 1A.

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 1A, 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.

5.1.2 Topic representation: Word Cloud + Streamlines

To present the topic interests of various demographic groups, we 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 1g, 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, we 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, we 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 1D. 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.

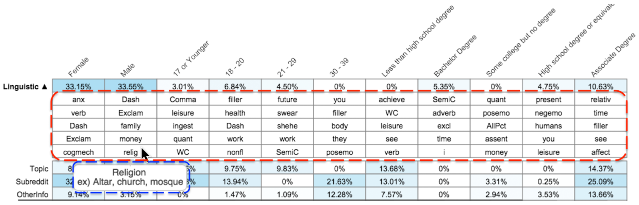
5.2 User Cluster View

To address T2, namely to find out whether the demographic groups have distinct or similar interest, we grouped the 409 redditors that participated in our 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, we leverage the topic results (Section 4.1.1). 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, we leverage a dimensionality reduction method called t-Distributed Stochastic Neighbor Embedding (t-SNE) [22]. t-SNE is particularly well suited for the visualization of high-dimensional datasets since it generates compact yet separable clusters [22].

To further assist users in linking the above dimensionality reduction results to demographic information, we 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 1B, each glyph represents one redditor, with their demographic variables (gender, age, education) captured by the glyph. We 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 1B, 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 1B shows two clusters with each including two different demographic



	Female	Male	17 or Younger	18 - 20	21 - 29	30 - 39	Less than high school degree	Bachelor Degree	Some college but no degree	High school degree or equivalent	Associate Degree
Linguistic ▲	33.15%	33.55%	3.01%	6.84%	4.50%	0%	0%	5.35%	0%	4.75%	10.63%
	anx	Dash	Comma	filler	future	you	achieve	SemiC	quant	present	relativ
	verb	Exclam	leisure	health	swear	filler	WC	adverb	posemo	negemo	time
	Dash	family	ingest	Dash	shehe	body	leisure	excl	AllPct	humans	filler
	Exclam	money	quant	work	work	they	see	time	assent	you	see
	cogmech	relig	WC	nonfl	SemiC	posemo	verb	i	money	leisure	affect
Topic	8	Religion	6%	9.75%	9.83%	0%	13.68%	0%	0%	0%	14.37%
Subreddit	32	ex) Altar, church, mosque	%	13.94%	0%	21.63%	13.01%	0%	3.31%	0.25%	25.09%
OtherInfo	9.14%	3.15%	0%	1.47%	1.09%	12.28%	7.57%	0%	2.94%	3.53%	13.66%

Figure 5: Feature table with linguistic features expanded. Five highest ranked linguistic features are displayed for analysis.

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.

5.3 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, we provide a feature ranking view (figure 5). In the tabular view, the features are aligned by row and the demographic variables are aligned by the column.

As introduced in section 4.1.3, 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 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.[4]

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, we 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 5 red rectangle) are presented. Hovering over a cell brings up the definition of the linguistic feature (figure 5 blue rectangle). Investigating Female and Male groups in our 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.

6 Case Studies

In this section, we 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, we take this opportunity to examine the interests of young crowds.

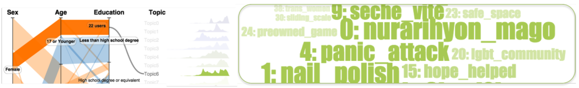


Figure 6: Female, 17 or younger, less than high school degree and the corresponding topic of interest.

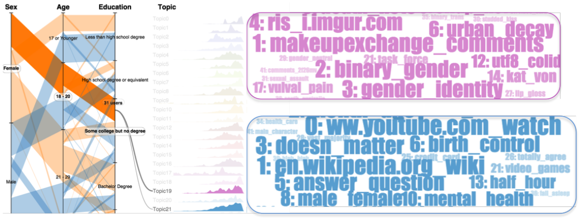


Figure 7: Female, 18-20, Some college but no degree and related topics.

6.1 What are young ladies interested in?

We start by examining the group of "female, 17 or younger, less than high school degree". The individuals in this 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 6. The topic summarizes the young crowds' interests on reading Japanese graphic novel ("nurarihyon_mago") [5], discussing makeup related terms ("nail_polish")), and mentions of "panic_attack", which could be concerning since clicking on the term leads to posts that discuss mental illness.

Next we move on to selecting group "female, 18-20, some college but no degree". Individuals in this group are likely college students. As shown in figure 7, 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"[6], young ladies within this demographic group also discussed gender identity and related issues: "male_female", "gender_identity", "birth_control".

6.2 Exploring diversity and overlap in interests among demographic groups

To identify whether there are overlaps between various demographic groups, we leverage the cluster view in DemographicVis. As shown in figure 8, 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 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[7]. Interestingly, the clusters with different demographic groups denote groups that share similar interests. Here we pick a cluster (a zoomed in view of the cluster is shown in figure 8) 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 8. 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".

[4] 0% is likely due to insufficient data

[5] At first we thought this finding is due to the particular sample on Reddit. However, a study [20] based on 75,000 Facebook uses also found dominating interests in Japanese comics among young users.

[6] A cosmetic brand

[7] 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

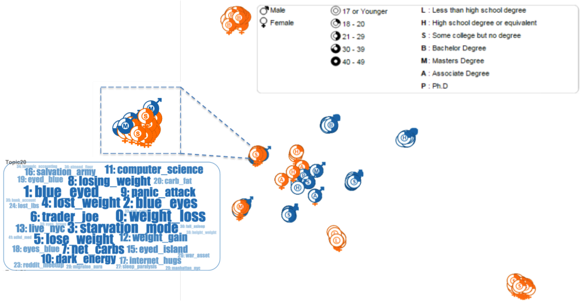


Figure 8: Cluster view that groups Reddit users based on their topic interests. One group with two demographic groups is enlarged and their shared topic is shown.

7 User Study

In this section, we present a user study to evaluate efficacy and usability of DemographicVis by comparing it with the SAS Text Miner [18]. The SAS Text Miner is one of the advanced analytics products developed by SAS that aims to extract insights from unstructured text. We consider SAS Text Miner as the best candidate for comparison because it integrates analysis and visualization capabilities for textual data.

7.1 Experiment tasks and apparatus

We designed 3 tasks for participants to perform with both DemographicVis and the SAS Text Miner. 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, we wanted to test users' understanding of predictive analysis based on various features. However, we could not find functions in the SAS Text Miner that support this task. The 3 tasks we 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. We 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.

We 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.

7.2 Results

In this section, we report the results from the user study. Subjective ratings (7-point Likert scale) on the two systems are reported. We 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. We choose to treat it as ordinal, therefore Friedman's test is used to analyze the results.

7.2.1 Accuracy rate

For Task1, the participants were asked to pick 3 most discussed topics. We 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, we asked participants to find the corresponding demographic groups for the top 3 topics. We 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.

7.2.2 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.

7.2.3 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, we see many participants commented on features provided by DemographicVis that show correlations between topics and demographic groups including "Different views

difficult to model their interests.

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".

8 Discussion and Conclusion

Throughout the design process, we noted that there are aspects we'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, we 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. We 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, we 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 our 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, we can experiment with different topic models to see which one may yield better results. We did experiment on generating different number of topics with our NMF-based topic model, and found that the number of topics does not affect the predictive results and the contribution of the topic features.

9 Conclusion

We 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 our comparative evaluation shows that DemographicVis is quantitatively competitive and subjectively preferred compared to the SAS Text Miner.

Acknowledgements

The work is supported in part by the Army Research Office under contract number W911NF-13-1-0083 and the U.S. Department of Homeland Security's VACCINE Center under award no. 2009-ST-061-CI0002.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *Visualization and Computer Graphics, IEEE Transactions on*, 12(4):558–568, 2006.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.