Talkographics: Learning Audience Demographics and Interests from Text to Make TV Show Recommendations

Shawndra Hill
Wharton School of Business, University of Pennsylvania, Philadelphia, PA 19103, shawndra@wharton.upenn.edu
Adrian Benton
Wharton School of Business, University of Pennsylvania, Philadelphia, PA 19103, adrianb@wharton.upenn.edu

This paper presents a novel recommendation system (RS) based on the user-generated content (UGC) contributed by TV viewers via Twitter, in order to demonstrate the value UGC presents for firms. In aggregate these TV viewers' tweets enable us to calculate the affinity between TV shows and explain the similarity between TV show audiences. We present 1) a new methodology for collecting data from social media with which to generate and test affinity networks; and 2) a new privacy-friendly UGC-based RS relying on all publicly-available text from viewers, rather than on preselected keywords. This data collection method is more flexible and generalizable than previous approaches and allows for real-world validation. We coin the term *talkographics* to refer to descriptions of any product's audience revealed by the words used in their Twitter messages, and show that Twitter text can represent complex, nuanced combinations of the audiences features. To demonstrate that our RS is generalizable, we apply this approach to other product domains.

Key words: social tv, recommendation system, user generated content, brand affinity networks

1. Introduction

A wide variety of data is available on consumers, including data that they themselves have freely made available to the public eye online. More and more, firms and researchers are deriving value from online user generated content (UGC). These data are being used for target marketing and advertising, and to help improve the precision of product recommendations for consumers. Both firms and consumers stand to gain when firms can make more reliable inferences about the characteristics of consumers purchasing their brands, as much in terms of demographics and interests as of product preferences, because these are the data that form the majority of bases for most recommendation system predictions.

Increasingly, businesses have used recommendation systems (RSs) to offer consumers suggestions about products, services, and other items that they might be interested in. RSs increase sales by

directing consumers to items that will likely suit their wants and needs and encouraging them to purchase them (Adomavicius et al. 2005). Since the worth of a RS is directly linked to its accuracy in predicting consumer preferences, particularly when it comes to new products and services, improving a system's accuracy in recommending a diverse set of items (Fleder and Hosanagar 2009) allows a business to offer its customers added value while gaining and retaining their trust and loyalty.

Existing RSs differ both in the types of data collected and in data-gathering techniques. Traditionally, RSs have relied on either content-based (CB) or collaborative-filtering (CF) methods to collect and categorize data about products, services, or people. CB methods calculate similarities among *products* and then recommend products similar to those a user has previously indicated. CF methods, in contrast, assume that similar people tend to have similar preferences, and therefore look for similar *users* who share their product preferences; based on these patterns of affinity, they recommend items that comparable users have purchased or shown interest in.

Data used in these older RSs are therefore straightforward indicators of users' preferences, such as information about product ratings or lists of products purchased by individual users. But it has become apparent that the Internet in particular offers much greater possibilities. In recent years, researchers and firms have experimented with extracting information about consumers from contexts (e.g., geographics, location, time and mood), social networks (e.g., Twitter and Facebook, what friends are doing and buying), and text (e.g., online consumer reviews and Twitter posts) to make more effective product recommendations. In particular, social media provides a rapidly growing body of user-generated content (UGC), such as text, images, and videos, that have the potential be used to improve RSs, and therefore deliver greater value to firms.

In this paper we ask whether we can derive value from Twitter-sourced UGC to make TV show and brand recommendations. In our proposed approach, TV shows and brands are represented by what people who follow them say, not only on the subject of the shows and brands, but in general. Text features from the UGC are derived to represent the shows. Based on these representations, we then use a content-based framework to calculate the similarity between the TV shows and brands followed and others that could be recommended to given users. What is unique in our approach is that we are able to use an aggregate level collection of general publicly available tweets to predict viewers' aggregate level features, for example their demographics and interests, remarkably well; furthermore, our approach is both privacy-friendly and generalizable to all product domains.

The micro-blogging platform Twitter is a promising source of data that provides a rich collection of real-time commentaries on almost every aspect of life, including consumers' responses to advertisements and television shows Hill and Benton (2012). Researchers have focused on Twitter extensively as a research testbed, improving recommendations by using the usage frequency

of certain words mentioned in tweets to determine the characteristics of users, including their demographics and geo-location. Twitter feeds have also been analyzed to build RSs that suggest particular websites or news stories that might be of interest to given users (Chen et al. 2010, Phelan et al. 2009), and used text mining to analyze UGC (such as online product reviews) and use it as a base for better recommendations. Other researchers aiming to improve recommender systems have worked with "folksonomies", arbitrary words or "tags" used to label uploaded content. But all of these approaches require some type of quite costly ontology, as well as rarely being privacy friendly, as they use individual level data to make recommendations. In our research, by contrast, we use all text features that users contribute, without using an ontology. We examine ways to collect the entirety of the public online text that followers of TV shows have shared on Twitter, and then use that information in aggregate to calculate affinity networks between shows, thereby finding related TV shows to recommend to viewers.

To ensure privacy while building our affinity model, we take all tweets posted by followers of the shows and brands and erase the tweeters' identities. To provide baselines for comparing our new approach to previous ones, we use several different types of data to calculate the similarity among shows. The primary baseline is a product network-based approach, combined with a very basic association rule strategy. In this approach, we calculate the similarity between shows based on the number of Twitter followers the shows have in common. It is important to note that we also compare our approach to a related baseline of the incidence of co-mentions of brands in tweets. However, this baseline performed very poorly, in part because of the sparseness of co-occurrence of TV show and brand mentions in tweets. We compare the product-network approach with our proposed new text-based method, which uses the Twitter texts of show followers to create a **talkographic** profile of a TV show, then uses the talkographic profile to calculate the similarity between one TV show and another. We demonstrate that these talkographic profiles reflect the interests, demographics, location, and other characteristics of users. The text that reveals demographics or specific interests helps us to explain why certain shows attract similar audiences. This ability to determine the nexus between users and product opens up multiple possibilities for businesses, far beyond the mere construction of new RSs.

The results published here build on prior work demonstrating that tweets and their content are reflective of both demographics and psychographics at the individual level (Michelson and Macskassy 2010, Schwartz et al. 2013). In this prior work, researchers linked answers given on a personality test taken by individuals to the text these individuals typed in their Facebook status updates. We extend this earlier work by showing that aggregate-level profiles, rather than individual-level profiles, are able to predict the aggregate-level demographics of viewers remarkably well. Thus, individual level demographics that might be hard to come by or infringe on privacy

rights are not required in our approach. By constructing aggregate-level profiles of TV viewers, our approach remains privacy-friendly, unreliant on any individual-level demographic data to make predictions or gain insights into the viewers and/or followers of products, services, and TV shows. Our RS capitalizes on what users are contributing in public, for free, about all aspects of their daily lives. In aggregate, these data allow us to estimate the demographics of populations of tweeters – in our case the populations that follow TV shows and brands. Our work also differs from prior work incorporating UGC into business decision-making in that we don't restrict our data to only those comments and tweets regarding the products. Instead we consider all tweets contributed by the TV show viewers, including those narrating the details of their daily lives. We can isolate the words and terms that best reflect the similarity between shows as well as establish which demographics, interests and geographics the words are most associated with. Our goal here is twofold. First, we show that UGC is valuable in that it can be used to generate TV show profiles, what we call the talkographic profile, that do not require an ontology and can therefore apply to a wide variety of products and services discussed and followed on online social networks; specifically, we show that these profiles can be used to make TV show-viewing predictions. Second, we determine those product types (popular versus niche, specific demographic audience versus niche interest audience) for which the UGC text is more effective for making predictions. We validate our approach using a novel data set we constructed, sourced from publicly available data. The data we collected combine the user-generated text with the users' TV show-viewing preferences, as indicated by the TV shows these users follow on Twitter, for a large subset of Twitter users.

As *talkographic* profiles may be generated for any product of interest for which a subset of consumers of the product that talk online can be identified and observed (something that Twitter makes possible for just about any brand, topic, or individual), their potential application as much for research as for business is virtually unlimited. In particular they offer a new model for firms to take advantage of the wealth of data consumers share online about their daily lives with very little cost.

2. Background

Recent years have seen an explosion of digital content concerning consumers, mostly user-generated content, that has been used by firms to gain business insights into their customers. By gaining the ability to make recommendations about services according to the customers' interests and characteristics, firms gain a marked advantage that is reflected in both revenue and customer satisfaction and loyalty. The data available for gaining these insights has grown even larger as users have begun to freely reveal their preferences in public in the form of posts on social networking sites like Facebook and Twitter. While the possibilities inherent in this rapid expansion of potential

information, both for academic and industry researchers, have not yet received the full attention they deserve, it is becoming ever more apparent that this valuable data will be at the base of future marketing efforts. One particular way in which they can be used is in constructing superior recommendation systems.

Table 1 lists the most relevant papers using user generated content to derive value for firms and users and how they compare to our work. To define the similarities and differences in approaches, we use a set of five main features contained in our approach. While some prior research, as can be seen in Table 1, exhibits one or more of these dimensions, to our knowledge no papers exist combining all five of these important characteristics in their demonstration of value. Our approach is thus substantially new and allows us to confirm its value by using UGC to construct effective affinity networks between brands which we can then test in a recommendation system context.

These five features are: 1) we use publicly available data; 2) we use aggregate level data, making our approach privacy-friendly; 3) we do not require an ontology, taxonomy or preselected set of keywords; 4) we capture demographic, geographic, and interest level features of the products' audience by including all words used by viewers on social media, not just words about the products; and 5) we validate our results using a predictive model using 10 fold cross validation on hold out sample data.

As can be seen in 1, prior research has looked at a variety of available consumer data. These data range from clickstream data and user-generated content on social networks and review sites Dellarocas (2003) to data being generated by mobile health applications on consumers' daily physical activity and consumption patterns. These data are being used in various ways by firms for business intelligence, for example to predict sales on Amazon Hu et al. (2008), movie success rates Eliashberg et al. (2007), and stock price movement Das and Chen (2007). Individual level clickstream data has been used in the past both to identify users based on their behavior and to categorize their demographics Montgomery and Srinivasan (2002) for better personalization. Clickstream data is highly proficient at inferring individual level demographics. However, the data are generally costly to acquire, especially when they must be gathered across multiple websites. This is also true of product review data, which has been used by many researchers to infer consumer preferences from the reviews they write Ying et al. (2006), Ghose et al. (2012), Decker and Trusov (2010) as well as infer important product features with ontologies Archak et al. (2011) and without Lee and BradLow (2011) based on what is said by consumers about the products.

2.1. Twitter UGC as a Research Testbed

Among the potential data sources for UGC, the micro-blogging network Twitter in particular may offer especially useful information for recommender engines. In recent years it has opened

Table 1 Summary of Prior Work

Paper Title	Authors	Domain	Publicly Available Data	Privacy Friendly	No Ontology	Generalizable	Prediction	Citation
Learning About Customers Without Asking	Alan L. Montgomery, Kannan Srinivasan	clickstream data			✓		✓	(Montgomery and Srinivasan 2002)
E-Customization	Asim Ansari, Carl F. Mela	email data, predicting which features of an email lead to more access of the website			✓			(Ansari and Mela 2003)
Leveraging missing ratings to improve online recommendation systems	Yuanping Ying, Fred Feinberg, and Michel Wedel	making movie recommendations, based on customer reviews		✓			✓	(Ying et al. 2006)
From story line to box office: A new approach for green-lighting movie scripts	Jehoshua Eliashberg, Sam K. Hui, and Z. John Zhang	movie spoilers, predict success	✓	✓	✓		✓	(Eliashberg et al. 2007)
Yahoo! for Amazon: Sentiment extraction from small talk on the Web	Sanjiv R. Das, Mike Y. Chen	messages about stocks from Morgan Stanley High-Tech Index message boards, predicting stock movement	✓	✓			✓	(Das and Chen 2007)
Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects	Nan Hu, Ling Liu, Jie Jennifer Zhang	Amazon product reviews (books, DVDs, videos), predicting sales rank based on consumer reviews	✓	✓	✓			(Hu et al. 2008)
Estimating Aggregate Consumer Preferences from Online Product Reviews	Reinhold Decker, Michael Trusov	mobile phone product reviews, predict user preference	✓	✓				(Decker and Trusov 2010)
Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics	Anindya Ghose, Panagiotis G. Ipeirotis	Amazon product reviews (audo and video players, digital cameras, and DVDs), along with sales rank	✓			✓	✓	(Ghose and Ipeirotis 2011)
Deriving the pricing power of product features by mining consumer reviews	Nikolay Archak, Anindya Ghose, Panagiotis G. Ipeirotis	Amazon product reviews (digital cameras and camcorders), inferring economic impact of these reviews	✓	✓		✓	✓	(Archak et al. 2011)
Automatic Construction of Conjoint Attributes and Levels from Online Customer Reviews	Thomas Y. Lee, Eric T. Bradlow	Epinions.com digital camera reviews	✓	✓	✓			(Lee and BradLow 2011)
Mine Your Own Business: Market-Structure Surveillance Through Text Mining	Oded Netzer, Ronen Feldman, Jacob Goldenberg, Moshe Fresko	message board data (diabetes/sedan forums)	✓	✓				(Netzer et al. 2012)
Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content	Anindya Ghose, Panagiotis G. Ipeirotis, Beibei Li	Travelocity.com/TripAdvisor.com/neutral third-party site hotel reviews		✓			✓	(Ghose et al. 2012)
Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach	H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, Lyle H. Ungar	facebook status posts, generating hypotheses about language use from different subpopulations		✓	✓		✓	(Schwartz et al. 2013)
Talkographics: Using What Viewers Say Online to Estimate Audience Demographics to Calculate Affinity Networks for Social TV-based Recommendations	This paper	Twitter status updates and networks, predicting TV show recommendations	✓	✓	✓	✓	✓	2013

up entirely new possibilities for assembling data that can be used for various purposes, including recommender systems. In tweets of 140 characters or less, people offer real-time news and commentaries about various happenings in the world. This includes their responses to television shows and advertisements. This rich trove of data has been put to use in various ways to inform recommender systems.

For example, Twitter users often comment on news stories as they appear; tweets therefore contain information about what interest Twitter users take in various news topics. Researchers have thus been able to use information from Twitter feeds to recommend particular news stories for a user's favorite RSS feeds (Phelan et al. 2009, 2011). Abel et al. (2011) identified topics as

well as entities (i.e., people, events, or products) mentioned in tweets and used semantic analysis of these tweets to improve user profiles and provide better news story recommendations. In research with a somewhat different goal, Sun et al. (2010) analyzed the diffusion patterns of information provided by a large number of Twitter users, who were essentially acting as news providers, to develop a small subset of news stories that were recommended to Twitter users as emergency news feeds.

Twitter users comment on a variety of information other than news stories, of course, including suggesting websites to other users. Chen et al. (2010) analyzed the best ways to develop content recommendations using a model based on three different dimensions: the source of the content, topic interest models for users, and social voting.

Twitter users' networks also provide a great deal of information. Each user will generally follow the tweets of a selected group of other users while at the same time being followed by a different group of users. The choices made about which Twitter feeds to follow hold a great deal of implicit information about a user's interests, and Hannon and colleagues have developed a recommender system that suggests new users for Twitter users to follow (Hannon et al. 2010, 2011).

A number of researchers have combined Twitter data with various sorts of outside information in efforts to improve recommendations. Morales et al. (2012) combined information from users' Twitter feeds with details about the users' social Twitter neighborhoods (or followers and friends) as well as the popularity of various news stories to predict which news stories would prove most interesting to Twitter users. They reported achieving a high degree of accuracy in predicting which news stories Twitter users would choose to read. Pankong and Prakancharoen (Pankong and Prakancharoen 2011) tested 24 different algorithms for making content recommendations on Twitter, with the algorithms taking into account various combinations of topic relevance, the candidate set of users to base predictions on, social voting, and metadata mapping. They studied recommendations in the areas of entertainment, the stock exchange, and smart phones, and found that one of these algorithms created very effective recommendations.

Given the richness of data that Twitter provides, it is somewhat surprising that more has not been done to harness this data in the service of providing useful recommendations and finding other ways to provide value to businesses. This may be due in part to the difficulty of analyzing Twitter messages, which, at a maximum of 140 characters, are prone to abbreviations and other forms of shorthand, sentence fragments, and reliance on context and previous messages to make their meaning clear, all of which increase the difficulty of making sense of them. The potential rewards, however, are great enough that it is worthwhile to attempt to find more and more useful ways to apply the data offered through Twitter to recommendations systems. In our case, we observe which TV shows are followed by Twitter users and combine that data with the followers' tweets

to represent a TV show. What we are using is the concept of Social TV, the fact that people are linking to and discussing TV shows online on a large scale.

2.2. Social TV

With the rapidly increasing popularity of social networks, producers of a growing number of television shows have sought to expand their popularity and viewership by adding an online social component to the experience of watching television. There has, of course, always been a certain social element to TV watching; from the beginning television made its effect by bringing news and actors directly into the living room, seemingly establishing a one-on-one relationship between viewer and viewed, and with family members, friends, acquaintances, and sometimes even total strangers gathering around a television to watch a show, sharing in the experience, observing one another's responses, and discussing what they were watching. Today a similar experience can be made available virtually, with viewers spread across thousands of miles but still able to observe one another's interactions with the show and to share their thoughts and reactions. This shared experience and interaction is referred to as social television or social TV (Mantzari, Lekakos et al. 2008). It has been reported that Twitter is by far the dominant player in the social TV market in terms of viewers commenting about TV shows in real time while watching.

Because Social TV is still in its infancy, it is not yet clear how best to design TV-centric social interactions or even what sorts of social interactions will be most desired by users (Geerts and De Grooff 2009). Thus researchers are examining which factors result in effective social TV interactions (Chorianopoulos and Lekakos 2008), and guidelines have been suggested for the best approaches to designing social TV experiences (Ducheneaut et al. 2008, Gross et al. 2008).

Researchers are also only just beginning to explore ways in which information provided by participants in social TV can be put to work, such as making recommendations on which shows to watch or become interactively involved with. For example, Mitchell et al. (2010) discuss how social networks could be used to identify what portion of the Internet's vast amount of available content is worth watching for Internet television users. But relatively few studies examine social TV and, in particular, the best way to shape its social interactions to achieve various ends, including both improving the user experience and helping advertisers and other businesses improve their own bottom lines. This is an area ripe for exploration.

In our case, we look at one potential application from which both users and firms can derive value: recommendation systems. In this paper, we build on prior work (summarized in Table 1) that shows the value of user-generated content to both consumers and firms.

Our work intersects with the disciplines of information systems, marketing, and computer science on the topics of text mining and RSs. As noted earlier, prior studies vary in terms of data used

and motivating problem. Some research approaches are privacy-friendly, relying on aggregate level data, while others rely on individual level data. Lastly it should be noted that researchers dealing with text mining often develop an ontology that requires extensive work and is rarely generalizable to all domains. With respect to validation, only a few papers were able to use a holdout validation set approach in part to limitations in the data collection.

2.3. Text Mining

Much of the user-generated content on social networks and on the Internet in general is in the form of text. This text is generally informal and unstructured and it can thus be challenging to extract meaning from it. The goal of text mining is to overcome these challenges and find effective ways to pull meaningful and useful information from different types of text (Dörre et al. 1999, Feldman and Sanger 2006). Among the most avid users of text mining tools are businesses, which have applied these tools in a variety of ways, from analyzing various types of information posted by consumers on the Web to looking for patterns in the vast amount of financial report data (Feldman et al. 2010) available to the public.

A great deal of attention has been paid to the use of data mining to analyze user-generated content, such as that found on social networks, with the goal of developing insights into the attitudes and opinions of groups of individuals. Much of this work has appeared in the computer science literature, as reviewed by Pang and Lee (2008) and Liu (2011). Among the various approaches to deriving information through text mining, a common denominator is that most of the approaches require significant analytical effort to obtain reliable and useful information.

For example, Netzer et al. (2012) combined text-mining tools with semantic network analysis software to extract meaning and patterns from online customer feedback on products. Archak et al. (2011) decomposed customer reviews collected through text mining into independent pieces describing features of the product and incorporated these into a customer choice model. Because reviews generally did not touch on all features and some features were mentioned in very few reviews, the team clustered rare opinions through pointwise mutual information and also applied review semantics to the mined text. Ghose et al. (2012) devised a system for recommending hotels to consumers. In addition to data collected with data-mining techniques from social media sources, they used a dataset of hotel reservations made through Travelocity over a three-month period, human annotations, social geotagging, image classification, and geomapping. Inserting the data into a random coefficient hybrid structural model, they estimated the utility of staying at various hotels for different purposes, making it possible to determine the hotel that represents the best value for a particular customer. In all of the aforementioned papers, a preconceived ontology was used. Only recently has work focussed on gleaning important features from text in an automatic fashion. In Lee

and Bradlow (2011), the authors automatically elicited an initial set of attributes and levels from a set of online customer reviews of digital cameras for business intelligence. While existing computer science research aims to learn attributes from reviews, our approach is uniquely motivated by the conjoint study design challenge: how to identify both attributes and their associated levels.

While all of these approaches provide useful information, they require analytical sophistication and that significant effort be put into designing and developing an ontology. A simpler text-mining approach that could extract useful information with much less effort would be valuable. For example, researchers Schwartz et al. (2013) have recently used an ontology-free approach to link the text of Facebook status updates to answers to personality tests, linking individual-level text features to individual-level answers to the questions. Their approach is similar to ours except that we link aggregate level text features to aggregate level demographics, thereby not relying on private information. Our approach uses not only have text data but also product network data, the combination of which allows to make better recommendations than using either data source alone.

2.4. Text-based Recommender Systems

The vast amount of information contained on the Web, including the information available on social networks, makes it difficult for users to find the information that is most relevant to them. RSs address that difficulty by offering personalized recommendations of everything from consumer goods to websites and other users. However, designing an effective recommender that can infer user preferences and recommend relevant items is a challenging task. Researchers from several fields, including computer science, information systems, and marketing, have addressed this issue, devising a variety of approaches to making effective recommendations. We will highlight the most recent work on RSs used in business contexts.

After surveying the RSs literature, Adomavicius and Tuzhilin (Adomavicius and Tuzhilin 2005) found that most RSs could be classified as one of three types: content-based, collaborative filtering, and hybrid. Content-based systems make recommendations by finding items with a high degree of similarity to consumers' preferred items, with those preferences generally being inferred through ratings or purchases (Mooney and Roy 1999, Pazzani and Billsus 2007). One advantage of such content-based designs is that they can handle even small sets of users effectively; their major limitation is that one must be able to codify features of the products in a way that can be used to calculate similarity between products (Balabanovic and Shoham 1997, Shardanand and Maes 1995). Because our approach uses freeform text to quantify the audience of a brand or TV show, our approach naturally represents the brand in nuanced ways. CF systems base item recommendations on historical information drawn from other users with similar preferences (Breese et al. 1998). Using collaborative filtering in a RS makes it possible to overcome some of the limitations of content-based systems because information about the products does not have to be codified at all, but this

approach suffers from the new item problem - that is, the difficulty of generating recommendations for items that have never been rated by users and therefore have no history. The hybrid approach combines collaborative- and content-based methods in various ways (Soboroff and Nicholas 1999).

Researchers have also studied how to improve the accuracy of recommendations by including information other than customers' demographic data, past purchases, and past product ratings. Palmisano et al. (2008) showed, for instance, that including context can improve the ability to predict behavior. Using stepwise componential regression, DeBruyn et al. (2008) devised a simple questionnaire approach that helped website visitors to make decisions about purchases based on answers given about the visitor's context. Adomavicius et al. (2005) described a way to incorporate contextual information into recommender systems by using a multidimensional approach in which the traditional paradigm that considers users and item ratings together was extended to support additional contextual dimensions, such as time and location. Panniello and Gorgoglione (2012) compared different approaches to incorporating contextual dimensions in a recommender system. Similarly, Ansari et al. (2000) studied how to rank users by their expertise in order to better estimate the similarities between users and products. Sahoo et al. (2008) built a multidimensional recommender system to use information from multidimensional rating data. Atahan and Sarkar (2011) describe how to develop user profiles of a website's visitors that could offer targeted recommendations to users. Likewise Ansari et al. (2003) worked at targeting recommendations better by focussing on the email text of marketing messages. They found through text-mining that individually customizing the text of target marketing emails led to substantially greater use of the websites being targeted. While an ontology was not needed in this case, the data were both proprietary and required individual level data.

One notable line of research has examined how to modify recommender designs in order to increase the diversity of recommendations across product types and features. McGinty and Smyth (2003) investigated the importance of diversity as an additional criterion for item selection, and showed that it is possible to achieve significant gains in the effectiveness of recommendations if the way diversity is introduced is carefully tuned. Fleder and Hosanagar (2009) demonstrated that recommender systems that discount item popularity in the selection of recommendable items may increase sales more than recommender systems that do not. Adomavicius and Kwon (2012) showed that ranking recommendations according to the predicted rating values provides good predictive accuracy but poor performance with respect to recommendation diversity. They proposed a number of recommendation-ranking techniques that impose a bias towards diversity. Our UGC text-based method offers greater diversity than the product network with the additional feature that it can be tuned to skew towards popular shows if need be.

3. Testbed

To validate our social media text-based method against a variety of baselines, we compiled a large database of TV-related content with which to train our RS and evaluate social media-based RSs. Our methodology for data collection is itself a contribution to the recommendation systems literature, because it enables RSs researchers to both build and evaluate complex recommendation strategies using publicly available large-scale data. Our data collection process is illustrated in the flowchart in Figure 1a, and the schema for this database in Figure 1b.

3.1. Data collection needed for text-based RS

The data preparation and collection for this project was extensive and sourced data from a variety of online sources, including Amazon Mechanical Turk, Twitter, IMDb, TVDB, and Facebook. The data-collection process consisted of six main steps. We firstly selected a list of TV shows from the Wikipedia article titled "List of American Television Series" [4 add reference]. Since Twitter was established in July 2006, our selection of TV shows was restricted to those that ran from January 2005 through January 2012; we also removed those that were canceled or were no longer aired, yielding a list of 572 TV shows. We then pared down this list to 457 shows, removing very unpopular shows with little available data on from either Twitter or IMDb and TVDB.

Secondly, we pulled metadata for these shows from the popular websites TVDB.com and IMDb.com, which provide a wealth of information on TV shows, and stored them in the Show table. The metadata collected from these sites included genre labels, the year the show aired, and the broadcast network for each show (along with many other features listed in Figure 1b).

In the third step, we collected the official Twitter accounts for TV shows and the Twitter handles used to refer to them. Online volunteers with the Amazon Mechanical Turk service were employed to identify the Twitter handles for these 457 TV shows. For each show, we received data from at least two different volunteers to ensure correctness. We then manually examined the raw data to filter out incorrect or redundant handles.

The fourth step was to use the Twitter API to grab relevant network data and tweet streams for our analysis. For each of the Twitter handles collected by Mechanical Turkers, we queried the Twitter API to retrieve a list of all followers of that handle. This step provided us with over 19 million unique users who followed any one of these TV show handles. These unique users were then filtered to only those who followed at least two or more accounts, yielding approximately 5.5 million such users. The fact that the users in this set follow two or more shows gained its importance for it allowed us to evaluate our RS, by inputting one show to our RSs that we knew the user followed and then evaluating its ability to predict output shows also followed by the user. This ability to build recommenders and evaluate those recommenders on a hold out sample of users who follow

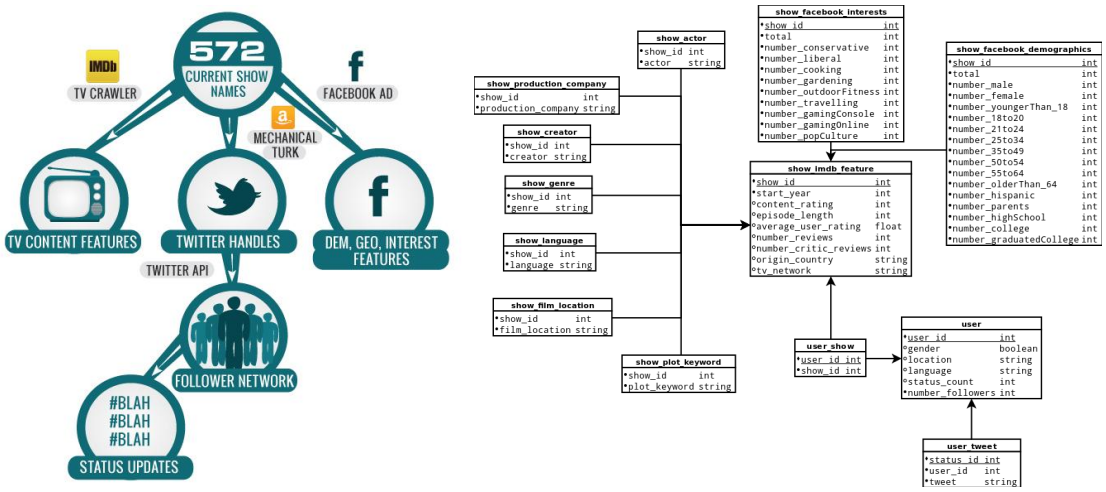


Figure 1 **a) Flowchart of the data collection process. Given a list of 457 seed shows, features and characteristics of shows were scraped from IMDb.com. At the same time, Twitter handles for these TV programs were gathered using Amazon Mechanical Turk. All user IDs following these programs on Twitter were collected, along with up to 400 of the most recent status updates they posted. b) Schema for database collected. The database we generated contains show features such as show content rating and genre, a mapping from shows to user IDs following that show, user profile features for a subset of TV followers (e.g., inferred location and gender), and a collection of tweets corresponding to each of the users in this subset.**

more than one TV show is the main novelty in our data. The scrubbed user account information was stored in a User table and the user-show relationships in a User Show table. Identifying user information was all eliminated, ensuring the users' privacy.

The fifth step was to extract tweets from the target users in Step Four. As s lack of time and computation left us unable to collect this fine-grained information for all users in our network, we opted instead to randomly sample users. Out of the 5.5 million users identified in Step Four, we randomly sampled approximately 99,000 users from the collected list of TV show followers. By sampling in this way, we ensured that our resulting set of users would provide sufficient coverage of all the TV show accounts under consideration. Users and tweets were filtered according to two criteria. First, a target user had to have self-identified as an English speaker in their profile's language field. Secondly, to prevent biased results, we focused on followers who were not public figures. It has been noted that celebrities and businesses usually have a large number of followers, so we restricted ourselves to users with no more than 2,000 followers each. We then extracted each of these 99,000 acceptable users' last 400 tweets.

In summary, our data consist of over 29 million tweets from about one hundred thousand Twitter users, containing hundreds of millions of words, along with their relationship to our set of 457 seed TV show handles. In Figure 2 we provide plots to illustrate the distribution of the number of shows followed by users (left), the log number of users per show (center), and the distribution of users per show pairs (right).

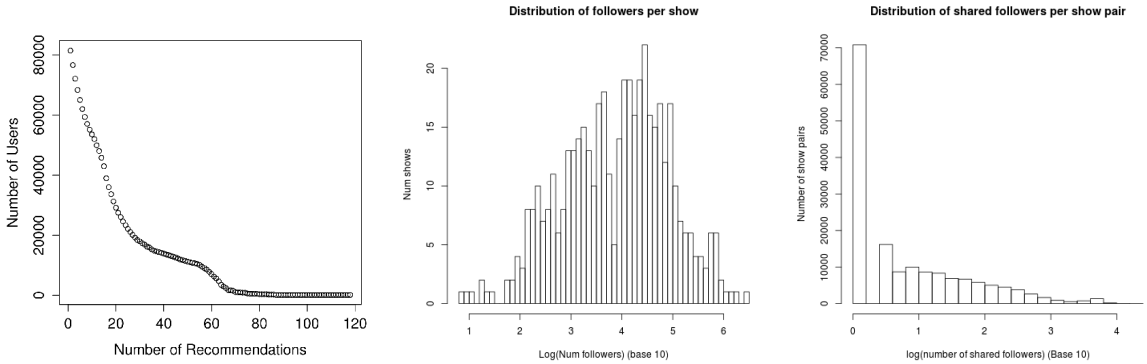


Figure 2 **Distributions of Followers of Shows and Show Pairs: The Left Graph represents the numbers of shows followed by a user. The Center Graph displays the log_{10} number of users per show. The Right Graph represents the log_{10} number of users per show pair.**

In the sixth step, we estimated the demographic features of each show by accessing Facebook advertising at (https://www.facebook.com/ads/create/). We advertised the link to our lab page (link removed for anonymity) in order to use and get access to audience information on Facebook[1]. Facebook's advertising interface allows advertisers to specify a particular demographic to target, as well as to target users who have declared a given set of interests. Once a particular target population is specified, Facebook provides an estimated reach for the ad campaign. In creating our ad, we specified that it should target users who lived in the United States and who had declared an interest in a show's Facebook page or the show's topic. This was repeated for each show. We used this as a proxy for how many online users were interested in the show. A screenshot of the Facebook advertising interface is shown in Figure 3.

We then estimated the proportion of users interested in this show who fell into different demographic, geographic, and interest categories by filtering all users according to those categories, and dividing the resulting number by the total number of those interested in the shows. Note that these values are estimated by Facebook and may not necessarily be representative of the TV-watching audience as a whole or of the Twitter follower network. However, as Facebook and Twitter share a similar demographic audience based on reports from online media measurement companies like Comscore, we believe it serves as a reasonably good proxy for the entire audience. Table 1 lists which demographics we sampled as well as the granularity at which we sampled them. We were able to collect demographic information for 430 of the 457 total shows. Owing to their lack of a Facebook presence, we were unable to collect demographic data for the missing 27 shows.

We also collected aggregate-level data on user interests, using the same method of querying Facebook advertising for the estimated reach within different populations. We calculated the pro-

[1] The Facebook terms of service state that these data can be reported at an aggregate level.

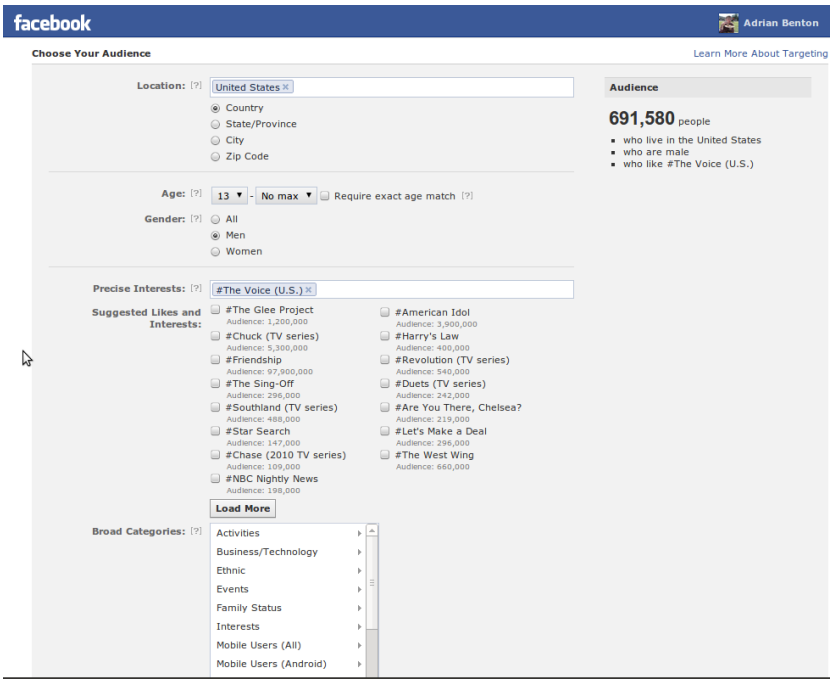


Figure 3 **Screenshot of the Facebook advertising interface. A potential advertiser is able to estimate the reach of their advertisement across a variety of interests as well as demographic characteristics such as age, gender, education level, and location.**

Table 2 **Different demographic categories collected from Facebook advertising for each show's audience.**

Demographic type	Demographic categories
Gender	male, female
Age	$<$ 17 yrs, 18-20 yrs, 21-24 yrs, 25-34 yrs, 35-49 yrs, 50-54 yrs, 55-64 yrs, $>$ 65 yrs
Hispanic	hispanic
Parents	parents (have children of any age)
Education level	in high school, in college, graduated college

Table 3 **Aggregate proportion of users interested in a particular topic or activity by show.**

Interest	Categories
Political opinion	conservative (binary), liberal (binary)
Cooking	binary
Gardening	binary
Outdoor fitness	binary
Traveling	binary
Gaming	console (binary), social/online (binary)
Pop culture	binary

portion of each show's followers on Facebook that showed an interest in each category. The user interests that we considered are listed in Table 2.

In addition, we estimated the proportion of users located in both the Northeast and Southeast by querying Facebook advertising for the estimated reach of users living in either New York, Pennsylvania, or New Jersey (for the Northeast), and South Carolina, Georgia, or Alabama for the southeast.

We then attempted to capture fine-grained audience demographic categories that might not be readily available in standard surveys. To do so, we again used the Facebook advertising platform

Table 4 Data description and original size of various Tweet subsets used

Data Description	Num Tweets	Num Unique Tokens
All Tweets	27114334	4075178
Show-Related Tweets	376216	75768
No Show-Related Tweets	26738118	4038483
English Tokens Only	27114334	20898

to query for the estimated reach of the intersections of gender and political opinion and of gender and age group (less than or equal to 30 years old versus 31 years or older). This demographic, geographic and interest data is used to describe the audience of shows.

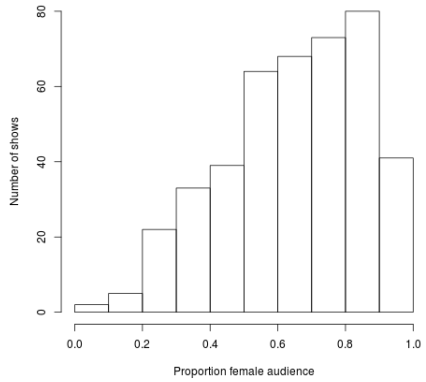
To provide various sanity checks, further discussed in the Methods section, we restricted our data in a number of ways. We firstly restricted the collected tweet text to include only tweets directly relevant to the TV shows in our sample, containing both show handle and hashtag mentions. We also operated in the opposite manner, removing any tweet mention of the TV shows to avoid the criticism that the tweets are just picking up people talking about the TV shows. Approximately 360,000 tweets were selected from the total set of 27 million, generating a training set of similar size to the set of show mentions (approximately 370,000 mentions). This set was generated by taking each user in our training set and randomly selecting 1.25% of their tweets to be included. By generating the set in this way, we ensured that each TV show follower's tweets were representative of the full set of 27 million messages. We constructed several subsets of the total set of tweets. In the first we removed show-related tweets. In the second we included only show-related tweets. The third is a data set constructed using only the WordNet English dictionary Fellbaum (1998) that appear in the show word feature vector (or bags of words). In this data only words in the English dictionary were included. This reduced the maximum number of unique tokens in our bag of words vectors from over 4 million to roughly 20,000. A description of various subsets of tweets used to build models can be found in Table 4. Note that when we later compare models built on these subsets, we limit all datasets to the smallest sized set.

3.2. Other testbeds

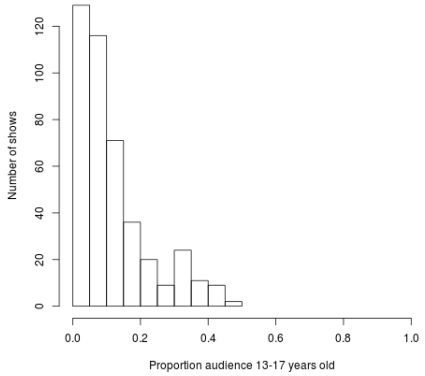
To evaluate the robustness of our text-based approach, we selected two other product domains to which to apply it. We chose a domain of automobile manufacturers and one of clothing retailers/brands. We applied the same method for collecting Twitter data as in the set of TV show handles - specifically, for each of the seed products we collected the user IDs of their Twitter followers and up to the last 400 tweets that a subset of these followers posted. These followers met the same criteria as those of the TV show data. Descriptive statistics for each of these datasets are included in Table 4, and distribution of demographic attributes across each of these datasets is included in Figure 4.

Table 5 Basic statistics for additional datasets. These domains contained fewer products to make recommendations for; however, we collected enough training data to allow for a comparison of the TV show set evaluation.

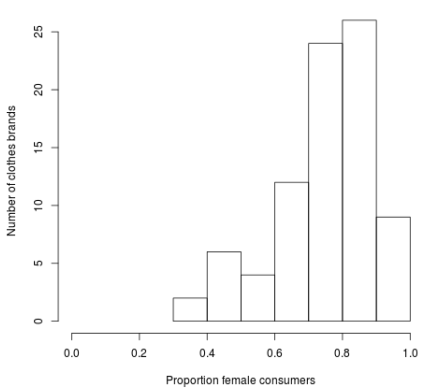
Dataset	♯ seed handles	♯ unique followers	♯ users in training/test folds	♯ tweets from in-fold users
Auto	42	1789399	68516	14912886
Clothing	83	8856664	110847	26993874



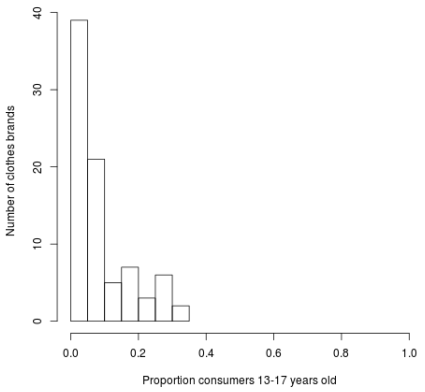
(a) TV shows female proportion



(b) TV shows proportion < 18 years



(c) Clothes proportion female



(d) Clothes proportion < 18 years

Figure 4 Distribution of selected demographic attributes over the TV show and clothes datasets. Note that although TV shows and clothing brands skew more towards female fans, the distribution of female fans across automobiles is more Gaussian. In addition, the Automobile brands skew older.

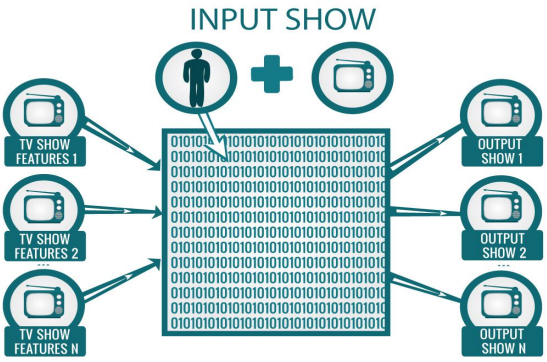
4. Method

In this section, we describe a set of RSs that we built based on different methods of using user-generated content. For each method, we assume one input TV show per user, picked at random from the shows the user follows, and use that input show to predict what other shows the user might like by calculating the similarity between the input show and other potential shows according to various metrics. For each approach, we calculate the similarity between shows by using a

Algorithm 1 Recommendation Evaluation Process

Input: A recommendation engine e, 10 sets of users (with the shows list they followed) based on cross-validation {test[1], train[1],⋯,test[i], train[i],⋯,test[10], train[10]}.
for (i IN 1:10) **do**
  Prepare to list the results for each set:
  results[i] = [ ]
  Train a recommendation metric based on the training set:
  Metric[i] = TRAIN(e, train[i])
  Test on each user u_j in the test set
  for (u_j IN test[i]) **do**
    Randomly choose a show from u_j shows list:
    randshow(j) = GET_RANDOM_SHOW(u_j)
    Use the trained metric to recommend show for user u_j:
    recommended(j) = PREDICT(Metric[i], u_j, randshow(j))
    Evaluate the performance of recommendation:
    results_byuser[j]=EVALUATE(recommended(j), u_j, randshow(j))
  end for
  Get the average performance for each test set:
  results[i]=average(results_byuser)
end for
Output: (SUM(results)/10)

(a) Evaluation Process



(b) Recommendation System Design

Figure 5 **a) Algorithm for evaluating each of the recommender system models. For each test user, a show that they follow is selected at random as input. Features of this show and features of the input user are used by the model to make a set of predictions. Given this set of predictions and the true set of other shows that the user follows, the performance of the model is evaluated and averaged across all users in the test set. These average performance metrics are then averaged across all folds. b) illustration of input show to output show recommendation and evaluation process – affinity network is built on training data then one input show is picked at random for a user and M recommendations are made.**

training set of approximately 90,000 users and apply the similarity matrix to a set of approximately 9,000 test users. We then perform 10-fold cross validation for all methods to report results on 10 training/test data pairs. Figure 5a shows the algorithm that describes the general approach for calculating metrics. We evaluate our predictions using standard RS measures of precision and recall. In this, precision is the number of correct predictions over the total number of predictions made, $\frac{|r \in pred | r \in actual|}{|pred|}$ and recall is the number of correct TV show predictions over the number of shows the viewers actually follow, $\frac{|r \in pred | r \in actual|}{|actual|}$. We further evaluate the methods using other metrics of diversity, but due to space constraints we will present only the precision and recall results in depth in the Results section, providing the additional analysis in the Appendix.

Figure 5b illustrates our method's input and output. For each tested RS method, we take in a show for a user and use a similarity matrix built using the method to make predictions by returning the most similar shows using the method.

4.1. Evaluating text-based model against baselines

We compared multiple RSs based on these social media data as baselines for our text-based model. All of the models were evaluated using the same training and test sets, and their precision and recall were evaluated in the same way.

4.1.1. Content-Based Approach For the content-based approach, we collected the features of 457 recent TV shows from IMDb.com and computed the similarity between all of them with

Table 6 The different similarity functions used for each of the content-based feature dimensions. If f_1 and f_2 are numerical values for shows 1 and 2 along feature f, then difference is defined as $\frac{max_i(f_i)-|f_1-f_2|}{max_i(f_i)}$. If f_1 and f_2 are sets, then intersection is defined as $\frac{|f_1 \cap f_2|}{|f_1 \cup f_2|}$. Exact similarity is simply the indicator function of equality.

Feature	Similarity metric
Year first broadcast	difference
Content rating (G=0, MA=5)	difference
Episode length in minutes	difference
Genres show falls under	intersection
Average user rating	difference
Number of non-critic reviews	difference
Number of critic reviews	difference
Creators of TV show	intersection
Major actors in TV show	intersection
User-generated plot keywords	intersection
Country of origin	exact
Languages broadcast in	intersection
Production companies associated with TV show	intersection
States/provinces TV show was filmed in	intersection
Network TV show was broadcast on	exact

respect to each of these separate features. We then applied a linear weighting of these features from a reserved set of users to combine these features appropriately. We used ordinary linear regression, in R, to determine this weighting. For two shows' feature vectors a and b, a learned weighting of similarity scores w, and a vector of scalar input similarity functions s, where $|a| = |b| = |s| = x$, the similarity between the two show feature vectors is defined as: $SIM(a,b) = \sum_{i=1}^{x} w_i * s_i(a_i, b_i)$. The features and similarity functions used are listed in Table 5.

4.1.2. Text-based Approach To compute user-generated text-based similarities between all shows, we used the tweets collected from followers of these TV shows to build a bag of words models for each of the seed shows. Although we were only able to collect tweets for a random sample of each show's follower network, models built using reduced data suggested that additional training data would not significantly improve the performance of the models.

4.1.3. Text-based All Tweets The user sampling resulted in a total of over 27 million tweets. If a user was known to follow a given show, then all of his/her tweets were added to that show's tweet corpus. Each show's tweet corpus was then tokenized by whitespace and non-alphanumeric characters. Twitter-specific tokens such as handles (Twitter usernames), URLs, and "RT" or retweet tokens were removed, and a "bag of words" was built for each show, along with counts for each token.

The similarity between two shows was generated using the cosine similarity between their bags of words, after transforming the show bags of words using term frequency * inverse document frequency (TFIDF). We calculated TFIDF in its typical form as well as taking the log of the numerator. This transformation was used to discount highly frequent words from overwhelming the bag of words vectors. The TFIDF value for a token t in a particular bag of words v_i, where J is

the set of show handles, was defined as follows: $\frac{v_{it}}{log(|J|/|j|v_{jt}>0|)}$. Cosine similarity was implemented, in the standard way as follows: $sim_{ij} = \frac{v_i \cdot v_j}{|v_i||v_j|}$

4.1.4. Text-based only English tokens We constructed a model using all show follower tweets that only considered tokens that appeared in WordNet's English dictionary. As mentioned in section 2.2, the bags of words vectors were reduced from over approximately 4 million unique tokens to about 40,000 tokens. Using the metrics of precision and recall, we evaluated this model against our original model using all unique tokens that appeared in tweets, as we did for all of the trained models.

4.1.5. Text-based TV Show Mention Tweets We also used an alternative approach in which only tweets that mentioned a show's Twitter handle were included in its bag of words. This resulted in a significantly smaller corpus of just over 370,000 tweets. The similarity was computed as the cosine similarity between the TFIDF-transformed vectors of the two shows.

4.1.6. Product Network In the TV show network approach, we measured the association between pairs of shows using the association rule metric of confidence. For two sets of users A and B, where A is the set of users who follow show a, and B is the set of users who follow show b, we defined directional confidence from show a to show b as $C(a,b) = \frac{|A \cap B|}{|A|}$. In other words, the total number of users who happen to follow both a and b divided by the total number of users who follow show a.

4.1.7. Other baselines Categorical Popularity-based Method: The categorical popularity-based method is introduced as a supplementary baseline method. As it is a low-quality similarity, the category information is combined with the overall popularity ranking of shows to make the recommendation. When a user provides a past-liked show, the recommender engine returns the most popular shows in the same category as the past-liked show.

Geography-based Method: Geographical information is always a popular way of making recommendations, since evidence suggeststhat geographic neighbors tend to share a background on cultural, academic, and economic levels. By grabbing the available location information of users from the Twitter free-text "location" field, the system will return the TV show with the largest number of followers in that area. We either use the user's latitude and longitude data (available for about 1-2% of users in our data), or, if this is unavailable, we attempt to infer the user's state and city based on the free-text location field in their user profile. We infer their location using a dictionary of locations in the United States and by attempting to match their self-reported location with entities in this dictionary. By inferring geographic location, we were able to infer state-level location data for about 10% of users in our set. The geography model learned from training data

predicts location at the US state level, and thus only applies to Twitter users located in the United States. Users for whom we were unable to infer location data, or who were located outside of the United States, were grouped into the same category when making predictions.

Gender-based Method: Similar to the geography-based method, we recommend the most popular shows by gender. Gender is inferred using a first-name lookup match from the user's personal name field to male and female dictionaries provided in the Natural Language Toolkit names corpus. First names which were not found in either dictionary or were ambiguous were classified as gender unknown.

In addition to these categorical popularity-based baselines, we also implemented a model which recommends the most popular shows of the entire training set irrespective of the input show a user is known to follow. This is a trivial version of a categorical popularity-based method, where all user-show pairs are placed into a single category. To reduce the clutter on the plots, when presenting results in the body of the paper we will compare only the text-based approach to the baselines of TV show network, popularity-based, and random approaches. Results from other models are presented in the Appendix.

4.2. Analyzing the performance of text-based system

In order to understand why our text-based system was performing as well as it does, we correlated the token frequencies measured by TFIDF scores in the shows' bags of words with audience and show features measured by the proportions calculated using the Facebook advertising interface. In addition, we evaluate the performance of our system when only including those tokens that are highly correlated with any of these features.

4.2.1. Linking text to demographics and show features We first constructed a table in which each row corresponded to a particular show, where the proportion of users in each demographic category was considered as a dependent variable, and the token frequency measured by the TFIDF score of each token in the show's bag of words was considered the independent variable. We then correlated the token's frequency with each of the dependent variables, one at a time, using ordinary linear regression in R, recording the estimated weight for the token frequency and the R^2 fit of the model. Filtering only those tokens with a learned positive weight, we ranked them in descending order by fit. In other words, for the proportion of a single demographic category in a show's audience, d, and a single token frequency t_i, the intercept c_0 and coefficient c_i were learned for the model $d = c_0 + c_i * t_i$ using least-squares estimation. t_j was retained if and only if $c_j > 0$, and t_j was ranked based on R_j^2.

Similarly, we correlated token frequencies with the show's genre, using logistic regression. Each genre that a show might possibly be classified under was treated as a binary variable indicating

whether or not the show was classified as this genre. We did not consider genre to be a multi-valued categorical variable, since shows can be assigned up to three different genre tags on IMDb. Just as in the correlation to aggregate demographic features of show audience, we filtered only those tokens with positive weight, and ranked the fit of each of the models using the Akaike Information Criterion in ascending order. In other words, given the binary variable representing whether or not a show falls under a particular genre label g, and a single token frequency t_i, the intercept c_0 and coefficient c_i were learned for the model $g = \frac{1}{1+e^{-(c_0+c_i*t_i)}}$, using maximum likelihood estimation over the set of shows.

4.2.2. Linking text-based features to user interests For the user interest variables collected from Facebook advertising, we also correlated the frequency of a token in a show's bag of words with the proportion of users who follow a given TV show and also follow a specific interest or activity. This was done using the same method of ordinary linear regression, retaining only those tokens with positive weights, and then ranking them by descending R^2 fit of the models learned.

4.2.3. Linking text-based features and aggregate geographic-level data Using the geographic features of the proportion of show fans living in the Northeast and Southeast United States, we were also able to correlate token frequency with the proportion of users living in particular regions of the United States. This was again done by fitting linear regression model toeach token, using least-squares estimation.

4.2.4. Analyzing the ability of text-based features to generalize and capture fine-grained demographic categories We then demonstrated that tokens in a show's bag of words are not only correlated with coarse demographic audience information, but also with more fine-grained demographic categories. We did this by correlating token frequencies with specific cross-sections of demographics, namely gender cross political opinion and gender cross age group. Token frequencies were correlated with these dependent variables, again using ordinary linear regression.

Finally, we consider the top K tokens most correlated with any of our demographic attributes, and evaluate the performance of the text-based model when only using these K tokens to calculate similarity. We then compare the performance of this reduced feature set model to the baselines described in section 3.1.

In order to determine whether the tokens found to be predictive of a particular TV show demographic could be generalized to other product types, we selected approximately 80% of the TV shows in our set, learned a ranking of tokens based on their level of correlation with a demographic attribute, then trained a linear model to predict this attribute based on the top N most-correlated tokens. The R^2 of this model was then evaluated on the training set (347 show brands the model was learned on), a holdout set (83 show brands disjoint from the training set), and a clothes handles

(83 clothes brands) set. We compared the performance of this model to that attained by randomly choosing N tokens over each of these sets.

4.2.5. Performance of text-based method as a function of show "nicheness" Given a distribution of a show's viewers over a series of demographic feature bins (e.g., different age groups, gender, education level), we defined the "nicheness" of a show as the symmetric KL divergence from that show's audience demographic distribution to the average demographic distribution across all shows. Given an average demographic distribution A, the input show's demographic distribution S, and a space of possible demographic bins D, the symmetric KL divergence was calculated as follows: $KL(A,S) = \sum_{d \in D}(\log(\frac{A[d]}{S[d]})) + \sum_{d \in D}(\log(\frac{S[d]}{A[d]}))$. By ranking them in this way, shows with high KL divergence - an atypical demographic audience distribution - were considered to have a more specialized, niche audience, whereas those with a low KL divergence - a more typical demographic audience distribution - were considered to have a more typical audience. All shows were then ranked by their KL divergence score and placed into five bins, based on their rank. For each set of shows in a bin, we evaluated the performance of the text-based method when considering only those cases where the top 1 recommendation made by this method belonged to this particular bin. Performance of the system was recorded and then compared to the product network baseline mentioned in section 3.1.

4.3. Other analyses

To test the generalizability of our model, we employed the same methodology used with the TV-show data-set on the automobile and clothing data sets, evaluating the performance of the product-similarity calculations by analyzing tweets from the products' followers. The major difference between these two corpora and our original testbed was that there were far fewer products that our model could make predictions for.

We also attempted to provide cross-product type recommendations. Looking at the set of users that followed at least one TV show and one clothing retailer/brand, we took as input one product from one of the product type. Using our methods, we then attempted to predict what products of the other type the user would also like. For example, given a fan of the TV show "The Voice," the system attempted to predict which clothing brands would show up among the user's interests. Here too we evaluated our text-based RS against the aforementioned baselines. The implementation of these systems were the same as the implementation of the within-product type recommendations, except that the predictions were ensured to be of a different product type as the input.

5. Results

To recapitulate, our method makes TV show recommendations to Twitter users who already follow more than one show on Twitter. We take in one show that each given user follows and try to predict what other shows they follow, making a prediction for each user-show pair.

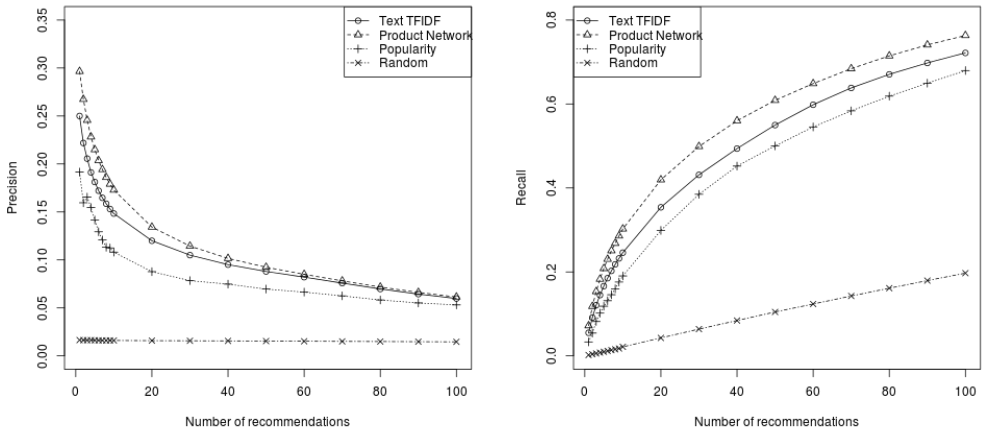


Figure 6 **The left plot (a) displays the precision of the different models, while the right plot (b) displays their recall. These are both a function of the number of recommendations each method makes, from 1 to 100. The rankings of the different methods are the same for each metric. The TV show network model performs best, followed by text-based and popularity-based models. All three perform markedly better than the random baseline.**

5.1. Evaluating the text-based prediction model

This section presents the results of comparing the aforementioned recommendation strategies by precision and recall. The Appendix provides results for all of the additional methods we tested. For ease of reading, in the body of the paper we compare only the text-based TFIDF methods, the TV show network method, popularity based, and random for ease of reading the documents. Figures 5a and 5b demonstrate that the TV show network approach outperforms all of the individual recommendation engines, with the text-based TFIDF-transformed similarity method also performing well. The results make it clear that calculating similarity between shows by only considering show mentions, a method closest to those used in prior work incorporating text data to make product recommendations, performs poorly in comparison to considering tweets posted by TV show followers.

In Figures 6a and 6b, we present averages over 10 folds of cross validation, which demonstrate that user-generated content alone, in the form of tweets and Twitter follower behavior, can be used to make highly reliable recommendations. These averages further show that the different types of social media content, although based on the same set of users, yield different types of predictions.

5.1.1. Comparison to show mentions method As mentioned in the Methods section, our set contains far fewer tweets offering mentions of the shows, and the show mentions model therefore has a much smaller training set. To compare our all-follower text-based model with the show-mentions model, we reduced the size of the training data for our model to slightly below the number of tweets that were given to the show-mentions model (about 360,000 tweets). Figure 7 displays the

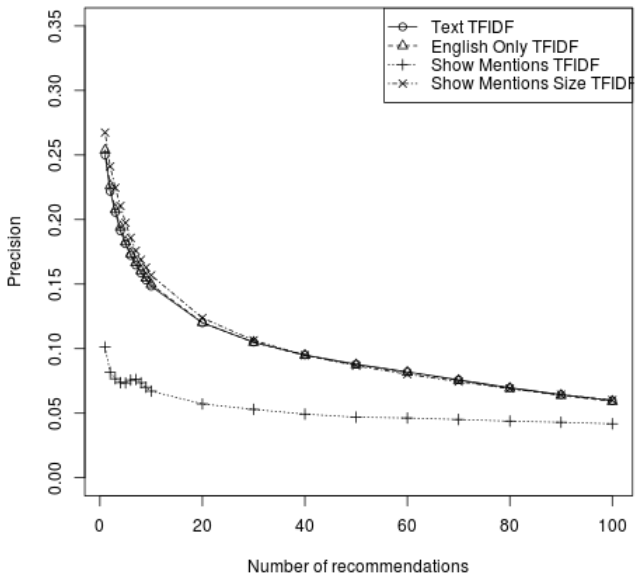


Figure 7 **A plot of the precision of various text-based methods against the number of recommendations made. It is clear from this figure that considering only tokens in our English language dictionary results in the same or slightly increased performance as using all of the tokens in the show bags of words. In addition, constraining the number of tweets used to a very small size (approximately 370,000 tweets, the same as the number of tweets in our corpus with show mentions) results in similarly high performance. However, calculating cosine similarity between shows based on tweets that mention each show does not result in very high performance compared to using all tweets generally posted by users.**

precision of each of these systems; it is evident that our method outperforms the show-mentions model by a wide margin, even with a reduced training set.

5.1.2. English-only bags of words We found that considering only a small set of English tokens to include in the show bags of words resulted in similar performance by our model. The precision of this model in relation to the full 4-million token bag of word vectors is displayed in Figure 6. From this graph, it is clear that by only including this small set of tokens, the model is able to achieve a similar performance. This suggests that the predictive power of our method does not rely on strange, difficult to interpret, Twitter-specific tokens, or on misspellings, and that it can be captured by natural English tokens. If one ranks the tokens from each show's TFIDF-transformed bag of words, the results are also promising. Table 6 lists the top-ranked tokens for a selection of shows in our set. Highly ranked tokens seem to be describing features of the shows as well as the audience of those shows.

5.2. Analyzing the performance of text-based system

Our text-based method is thus revealed as an instrument of surprising precision compared to previous methods. By analyzing the relationship between features of a show's audience, features

Table 7 Top-ranked TFIDF tokens for different shows. The language seems to be indicative of qualities of the shows and of the show audience.

American Idol	Amsales Girls	Colbert Report	RuPaul's Drag Race	Thundercats Now	Beavis and Butthead
idol *44659*	bridal *2984*	petition *20906*	gay *47358*	samurai *727*	f*** *97969*
birthday *199654*	wedding *39125*	bullying *20259*	lesbian *7681*	marvel *5289*	s*** *115609*
snugs *1537*	gown *2461*	newt *5938*	drag *7156*	barbarian *469*	f***ing *66297*
god *187816*	bride *4168*	republican *5801*	equality *6228*	cyborg *266*	loco *3387*
recap *27612*	curvy *683*	tax *14040*	marriage *28252*	batman *10972*	b**** *66153*
finale *75768*	meditation *1653*	president *37588*	maternal *608*	comic *14578*	ass *84656*
bullying *20259*	fortune *6198*	f*** *97969*	cuckoo *569*	wars *20389*	hate *184516*
love *1212244*	coziness *22*	debate *9507*	s*** *115609*	watchmen *469*	damn *88485*
excited *126069*	respectable *521*	freedom *17209*	b**** *66153*	spiderman *6993*	smoke *14896*
happy *474147*	hopefulness *26*	unsigned *991*	jewelry *11851*	extermination *39*	stupid *60120*

of the show itself, and token frequency within the text-based bag of words, we are able to isolate the method's power. Intuitive tokens are correlated with aggregate-level demographic features of the shows' audiences; finer-grained demographic categories that are likely to be overlooked in traditional surveys are captured by different sets of tokens. By considering only a small set of tokens that are correlated with demographic features of the shows, we are able to attain performance approaching that of the text-based method using the full bag of words. Finally, we show that input shows whose audience is skewed to a particular demographic category allow the text-based model to make more accurate predictions. In each of the results tables, we present a set of best-ranked words and their associated R-squared values at predicting the target proportion variable of interest (for example the proportion of female followers, proportion of cooking followers, proportion of southerners, etc.).

5.2.1. Linking text to demographics and show features By correlating the token frequencies within each show's bag of words to demographic features of its audience, we generated a ranking of tokens based on their correlation with the dependent variable of interest. Table 7 displays the top 10 tokens found for a selection of demographic categories using this ranking. Noticeably, these very telling rankings agree with prior intuitions about which words these particular demographics would use. While this work is similar to work by Schwartz, et. al., (2013), it is distinct from the latter in that we are attempting to correlate text features with demographic attributes at the aggregate level rather than the user level. One of the most surprising results of our research is the discovery that these correlations hold true at the aggregate level.

As mentioned in section 3.2.1, we also correlated token frequency with the shows' genres. Table 8 displays the top-ranked tokens according to this method for a selection of genres. The highly-ranked tokens also confirm intuitions as to which topics these shows might focus on. Together, these analyses suggest that this method allows the model to capture not only the demographic features of the show audience, but also the features of the show. The predictive power of this model, which has a level of effectiveness unparalleled by any previous methods, is clearly due to its use of all tweets from the shows' followers, rather than only tweets selected for their mention of the show.

Table 8 Best-fitting tokens predicting a particular proportion demographic. Note that some tokens have relatively high correlation with the proportion of a particular demographic (e.g., "love" has a fit of 0.36 with female, "school" has a fit of "0.23" with "less than 17 years old"). The R^2 value of the regression fit is in parentheses.

proportion female	proportion male	proportion < 17 yrs old	proportion 21-24 yrs old	proportion 25-34 yrs old	proportion 35-49 yrs old	proportion parents	proportion college grads
love *1212244* (0.38)	game *216034* (0.19)	ariana *4183* (0.24)	f*** *97969* (0.11)	work *276534* (0.09)	great *481832* (0.21)	hubby *15733* (0.19)	gop *12088* (0.19)
beautiful *148583* (0.21)	league *16740* (0.17)	school *150580* (0.23)	f***ing *66297* (0.10)	women *70392* (0.09)	service *32196* (0.17)	morning *201815* (0.15)	office *44718* (0.18)
cute *88419* (0.20)	hulk *6756* (0.14)	liam *19987* (0.20)	b**** *66153* (0.07)	daily *45143* (0.08)	taxpayer *574* (0.14)	blessed *21918* (0.14)	political *7501* (0.18)
happy *474147* (0.18)	battlefield *1977* (0.13)	direction *40141* (0.20)	s*** *115609* (0.07)	husband *22882* (0.08)	market *22733* (0.13)	husband *22882* (0.11)	media *52297* (0.17)
amazing *212601* (0.16)	comic *14578* (0.12)	victorious *3423* (0.19)	hate *184516* (0.06)	lounge *6104* (0.08)	pres *4428* (0.13)	family *153142* (0.10)	daily *45143* (0.17)
miss *177808* (0.15)	players *19295* (0.12)	follow *422471* (0.18)	boyfriend *30321* (0.05)	hire *5472* (0.08)	wine *25948* (0.12)	day *758441* (0.10)	st *72497* (0.17)
mom *112148* (0.13)	wars *20389* (0.12)	awkward *58774* (0.17)	song *173029* (0.05)	st *72497* (0.08)	recipe *14999* (0.12)	loving *52844* (0.10)	cc *14076* (0.16)
heart *125241* (0.13)	beer *25870* (0.12)	harry *49969* (0.15)	tenia *1263* (0.05)	interested *17493* (0.08)	media *52297* (0.12)	pray *23237* (0.09)	pres *4428* (0.16)
loving *52844* (0.13)	batman *10972* (0.11)	jonas *11110* (0.15)	bored *44822* (0.05)	drinks *11234* (0.07)	political *7501* (0.12)	bless *28137* (0.09)	service *32196* (0.15)
smile *62850* (0.13)	shot *35565* (0.11)	bored *44822* (0.13)	n**** *11847* (0.05)	keeping *17769* (0.07)	wealth *3083* (0.12)	happy *474147* (0.09)	homeland *3268* (0.15)

Table 9 Top-ranked tokens most correlated with genre of show. AIC of the logistic regression model fit is in parentheses. Many words pertaining to the program type are highly ranked.

animation	fantasy	horror	sports	mystery
animation *1953* (194.8)	moslem *17* (93.1)	moslem *17* (91.9)	champs *3877* (61.2)	mindedness *28* (218.3)
cartoon *5138* (201.8)	vampire *32962* (93.4)	volgograd *6* (98.8)	hill *21626* (64.3)	supernatural *19798* (219.8)
wobbling *31* (205.3)	demoniac *8* (94.7)	noisemaker *20* (101.0)	triple *6532* (66.0)	nostra *318* (221.7)
restrict *122* (207.5)	fesse *13* (95.6)	vampirism *27* (106.0)	intervening *39* (66.2)	axon *79* (222.2)
spelunker *9* (207.8)	noisemaker *20* (95.9)	supernatural *19798* (109.6)	heavyweight *1378* (68.6)	reunify *19* (223.3)
diabolic *6* (208.4)	pacifically *5* (96.4)	poetess *33* (110.5)	allen *7534* (69.2)	bankable *29* (223.3)
chainsaw *800* (209.5)	rattan *25* (97.0)	dekker *78* (110.6)	ahead *20707* (69.3)	paralyse *29* (223.3)
anime *2682* (211.0)	veronese *16* (97.2)	blackheart *34* (110.9)	racket *320* (69.3)	stabilisation *29* (223.3)
characters *15658* (211.6)	tabuk *7* (97.3)	garish *20* (111.9)	title *16106* (69.8)	quantal *29* (223.3)
comic *14578* (213.0)	viscera *12* (97.8)	calamita *37* (112.7)	bantamweight *35* (70.4)	oscan *29* (223.3)

5.2.2. Linking text-based features to user interests Similarly, when correlating token frequency with user interests, the tokens highly correlated with these outcomes tend to be intuitive. This suggests that the language of show followers is also predictive of user interests. Not only that, but the highly correlated tokens tend to be words that are indicative of that particular interest. Table 9 displays the 10 most correlated tokens for a selection of interests.

Table 10 Top-ranked tokens most correlated with user interests. The tokens all had positive weights and are ranked by R^2 fit.

cooking	gardening	travelling	pop culture
preservative *33* (0.08), oafish *13* (0.07), crockery *13* (0.07), terrine *35* (0.07), cherimoya *8* (0.07), food *91048* (0.06), restauranteur *1* (0.06), irrevocably *14* (0.06), compote *119* (0.06), padus *3* (0.06)	great *481832* (0.11), recipe *14999* (0.11), lots *38501* (0.09), market *22733* (0.09), puree *143* (0.09), organic *4981* (0.09), dinner *60313* (0.09), enjoy *78335* (0.09), meditation *1653* (0.08), handmade *3203* (0.08)	gop *12088* (0.10), bistro *1230* (0.10), candidate *4338* (0.10), latest *32903* (0.09), neil *5341* (0.09), campaign *20069* (0.09), government *12876* (0.08), reference *3559* (0.08), pilot *14436* (0.08), film *55699* (0.08)	love *1212244* (0.18), liam *19987* (0.15), direction *40141* (0.14), boyfriend *30321* (0.13), awkward *58774* (0.13), hate *184516* (0.13), school *150580* (0.12), girl *211081* (0.12), follow *422471* (0.12), malik *4413* (0.11)

Table 11 Top-ranked tokens most correlated with geographic region of the United States.

Northeast	Southeast
oread *1* (0.08), rathskeller *1* (0.08), naqua *1* (0.08), littre *1* (0.08), hopkinson *2* (0.08), squiffy *2* (0.08), porcine *2* (0.07), psilocybin *2* (0.07), cloisonne *3* (0.07), cloaca *2* (0.07), comber *2* (0.07), eero *3* (0.06), saarinen *3* (0.06), meridiem *3* (0.06), tacitus *3* (0.06), cepheus *11* (0.06), tuvalu *4* (0.06), scantling *6* (0.06), censer *3* (0.05), goncourt *2* (0.05)	blessed *21918* (0.12), interjection *33* (0.10), redouble *25* (0.10), god *187816* (0.10), birdseed *2* (0.09), rachet *223* (0.09), dis *7983* (0.09), shuffler *8* (0.09), nonjudgmental *26* (0.09), americus *5* (0.07), prayerful *100* (0.07), boo *18787* (0.07), fineness *6* (0.07), anthropocentric *8* (0.07), wit *22941* (0.07), scallion *44* (0.07), eleuthera *25* (0.07), evelyn *1158* (0.06), adverb *40* (0.06), n***a *11847* (0.06)

5.2.3. Linking text-based features and aggregate geographic-level data To see whether our model would also be able to predict geographic preferences, we also correlated token frequency with proportion of users living in the Northeast and Southeast of the United States. We filtered and ranked tokens in the same way, by positive weight and R^2 fit of the linear regression. Tokens suspected of being associated with these regions are also correlated with these geographic features. Table 10 displays the 20 most highly correlated tokens for geographic features of a show's audience.

5.2.4. Analyzing the ability of text-based features to generalize and capture fine-grained demographic categories Based on the success of our model, we claim that analysis of the language use of a show's followers can also capture fine-grained demographic categories, categories which it is uncommon to find defined in standard surveys. Table 11 shows that words most strongly correlated with a demographic cross-product category are better able to predict that subcategory than would a coarser demographic category. Table 12 displays the top 5 tokens most strongly correlated with gender and political opinion together.

To further test our contention that our proposed method allows for capturing these unusual fine-grained demographic categories unavailable in standard surveys, we returned to the Facebook advertising platform, looking at the intersections of gender and political opinion and of gender and age group (less than or equal to 30 years old versus 31 years or older). For each paired category

Table 12 Top 20 tokens learned for gender combined with young and old.

	Young	Old
Female	love *1212244* (0.37), direction *40141* (0.24), girl *211081* (0.23), cute *88419* (0.22), malik *4413* (0.21), boyfriend *30321* (0.21), liam *19987* (0.20), awkward *58774* (0.19), hate *184516* (0.18), school *150580* (0.17), eleanor *4139* (0.16), follow *422471* (0.16), moment *108532* (0.16), swaggie *711* (0.16), sister *37681* (0.15), harry *49969* (0.15), amazing *212601* (0.15), song *173029* (0.15), ariana *4183* (0.15), mom *112148* (0.14)	great *481832* (0.19), hubby *15733* (0.17), recipe *14999* (0.15), service *32196* (0.13), healthy *24287* (0.12), handmade *3203* (0.12), morning *201815* (0.12), wonderful *49184* (0.11), dinner *60313* (0.11), savory *433* (0.11), casserole *785* (0.11), blessed *21918* (0.11), meade *187* (0.10), prayer *11315* (0.10), scallop *187* (0.10), discipline *1675* (0.10), coffee *44093* (0.10), market *22733* (0.10), cardamom *101* (0.09), foodie *1007* (0.09)
Male	dude *49104* (0.11), game *216034* (0.10), battlefield *1977* (0.10), league *16740* (0.10), zombie *10965* (0.09), c*** *3196* (0.09), batman *10972* (0.08), cyborg *266* (0.08), metal *8062* (0.08), silva *1070* (0.08), play *123643* (0.08), megadeath *32* (0.08), gaming *4560* (0.08), comic *14578* (0.08), icehouse *24* (0.08), hulk *6756* (0.07), f***ing *66297* (0.07), ops *3353* (0.07), miller *6397* (0.07), beer *25870* (0.07)	war *35612* (0.14), game *216034* (0.13), league *16740* (0.12), hulk *6756* (0.12), field *16623* (0.12), newt *5938* (0.12), players *19295* (0.11), devils *7007* (0.11), occupy *6829* (0.11), conservative *3904* (0.11), officials *4613* (0.11), column *2905* (0.11), analyst *1465* (0.11), pitch *5718* (0.11), comedy *24384* (0.10), political *7501* (0.10), pentagon *993* (0.10), striker *484* (0.10), shark *8049* (0.10), jones *17190* (0.10)

(for example Young and Female) we used the proportion of followers on Facebook as the dependent variable. Table 13 shows the results: when we build a model on the basis of the top 3 female words and their associated weights for each show only, as opposed to all of the words that represent the audience of the shows, we predict the proportion females in the audience of shows better than we predict the young female proportion, and when we build a model on the top young female tokens we do a better job of predicting young female proportion, as evidenced by R-squared values on a holdout set of shows. In other words, words can be predictive of the audience demographics. To calculate the results in Table 13, we learned the top words on a training data, figured out how many top words we should consider in a validation data set, and applied it to a test set; we then performed this process 10 times, yielding Table 13's results.

Though these various results provide clear evidence of our model's superiority over previous attempts in this area, it is clear that the method can only be of true value if it can be demonstrated to be generalizable to areas other than television shows. Accordingly, we attempted to determine if the language used by TV show followers that was predictive of a demographic attribute could

Table 13 Top 20 tokens learned for gender combined with political opinion.

	Liberal	Conservative
Female	bachelorette *4033* (0.12), hubby *15733* (0.11), amazing *212601* (0.10), umbria *114* (0.10), monogram *157* (0.10), happy *474147* (0.10), floral *1509* (0.10), excited *126069* (0.09), silhouette *354* (0.09), love *1212244* (0.09), yay *72571* (0.09), braid *855* (0.09), batch *1758* (0.09), yummy *18518* (0.08), cute *88419* (0.08), dixie *6304* (0.08), capiz *8* (0.08), nape *190* (0.08), idol *44659* (0.08), rochelle *413* (0.08)	evelyn *1158* (0.21), blessed *21918* (0.20), interjection *33* (0.19), morning *201815* (0.18), redouble *25* (0.18), god *187816* (0.18), braxton *757* (0.16), thirdly *17* (0.16), boo *18787* (0.15), zambian *17* (0.15), scallion *44* (0.14), nonjudgmental *26* (0.13), adverb *40* (0.13), salaried *24* (0.13), transferee *24* (0.13), yaw *141* (0.13), rachet *223* (0.12), benet *110* (0.12), love *1212244* (0.12), authentically *48* (0.12)
Male	tactical *203* (0.17), game *216034* (0.16), battlefield *1977* (0.13), league *16740* (0.13), ops *3353* (0.12), survival *3633* (0.12), players *19295* (0.12), midfield *198* (0.12), fullback *148* (0.11), warfare *2761* (0.11), hockey *13883* (0.11), duty *8044* (0.11), shot *35565* (0.11), preseason *1087* (0.10), conservative *3904* (0.10), tourney *2092* (0.10), championship *10188* (0.10), war *35612* (0.10), strikeout *191* (0.10), saints *9296* (0.10)	comedy *24384* (0.15), hulk *6756* (0.13), coxswain *5* (0.12), comic *14578* (0.11), inaudible *3* (0.11), automatism *21* (0.11), marsupium *21* (0.11), stenosis *11* (0.10), pitchfork *263* (0.10), game *216034* (0.10), mangold *16* (0.10), anthropomorphic *25* (0.10), hornblower *17* (0.10), agitating *25* (0.10), theorize *2* (0.10), driveshaft *2* (0.10), feasibly *2* (0.10), toklas *2* (0.10), argot *2* (0.10), chicanery *2* (0.10)

Table 14 In table a, each row corresponds to the models learned when considering the top 5 words most strongly correlated with Female, Young Female, Old Female, Male, Conservative Male and Liberal Male viewers. The columns correspond to the dependent variables that are being predicted by these tokens. The values in each of the cells are the R^2 fits of each of these models. Similarly, In table b we have fits for models learned considering gender cross political opinion.

	Female	Young female	Old female		Male	Conservative male	Liberal male
Female	0.38	0.33	0.12	Male	0.40	0.36	0.34
Young female	0.41	0.44	0.25	Conservative male	0.38	0.65	0.20
Old female	0.05	0.11	0.31	Liberal male	0.40	0.15	0.74

be generalized to followers of clothing brands, as a suitable example of another product where both consumers and businesses could benefit from enhanced recommendations. Figure 8 displays the results of this analysis when including only the top 5 most correlated tokens in the learned model. Similar results were observed when varying the number of tokens from 1 to 10. From this plot it is clear that over all the sets, tokens learned on the training set are more predictive of all demographic attributes considered than are a randomly selected set of tokens.

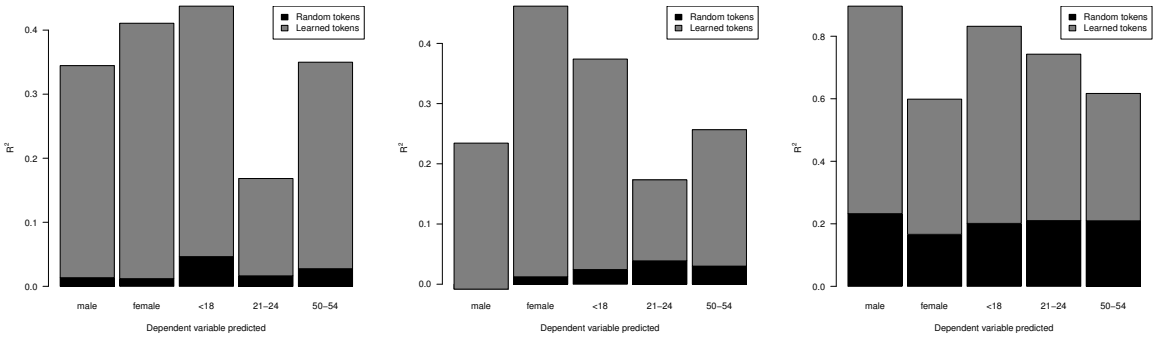


Figure 8 **(a) R^2 attained by the learned model on the training set, (b) the held-out set of TV show handles, and (c) the set of 83 clothes brands, against the dependent variable predicted and a model using a randomly-chosen feature set. These results are when considering the 5 most-correlated tokens in the model. It is clear that the tokens learned in the show domain generalize to the domain of clothing brands.**

5.3. Words highly correlated with demographics are driving the text-based results

We further determined that by only considering a small set of tokens, those most strongly correlated with demographic attributes, we attain similar performance to the text-based model using all tokens. From Figure 9, it is clear that considering only demographically-correlated tokens results in similar performance to the full text-based model. It also significantly outperforms a system where similarity is defined as the cosine distance between proportion male and proportion female viewers between shows, showing that the increased flexibility of the tokens allows us to outperform a content-based model with fewer degrees of freedom. This also shows that demographic features are driving the text-based TFIDF results.

Given that demographic text features appear to be driving the text results, we wanted to make sure it is not demographics alone. Figure 9 therefore also shows the results attained when we calculate the similarity of shows based on the Facebook demographic features we collected. As can be seen, this method performs poorly, indicating that in fact the text features possess a value beyond merely learning, for example, the proportions of certain demographics from Facebook.

5.3.1. Performance of text-based method as a function of show "nicheness" Given our binning of shows based on how homogeneous their viewership is by gender, we evaluated the performance of our text-based approach as a function of how much the method's input shows audience demographic makeup differs from the average demographic makeup over all shows. The results of this evaluation are displayed in Figure 10.

5.4. Other analyses

We also validated our method on two other data sets in order to assess its generalizability. Just as we did with TV shows, we collected the networks and tweets of followers for a selection of 42 car

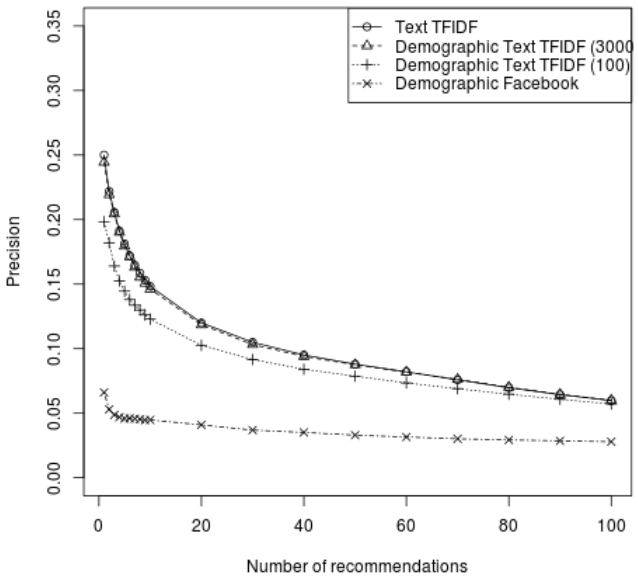


Figure 9 **Precision for a set of text-based methods, considering only the top K(100/3000) English tokens most correlated with audience demographic, given the number of recommendations. The top 3000 English tokens perform at a level comparable to considering the more than 20,000 total English tokens in our data, whereas considering only the top 100 results in some reduction in performance. However, both of these methods outperform computing similarity by considering only the aggregate-level demographic features of each show.**

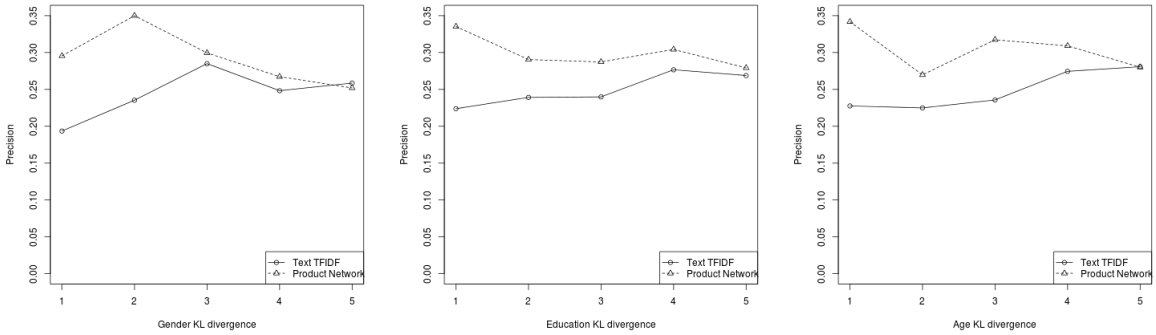


Figure 10 **Precision of our text-based method compared to the baseline product network method given number of recommendations made by the systems. Different lines correspond to successively higher bins of KL divergence of the recommended output show from the average demographic distribution over all TV shows. From this figure, it is clear that our text-based method makes more accurate recommendations when recommending shows with some demographic bias, and is outstripped by the product network method on those that have a more typical demographic mix of consumers. This confirms suspicions that the performance of our text-based method is driven by its ability to make recommendations based on demographically-correlated tokens. (a) Performance of methods when binning by KL divergence of gender distribution, (b) education level distribution, and (c) age group distribution.**

brands, and evaluated the performance of the product-network model against the product-follower text-based method and the popularity baseline. Figure 11 displays the precision and recall of these

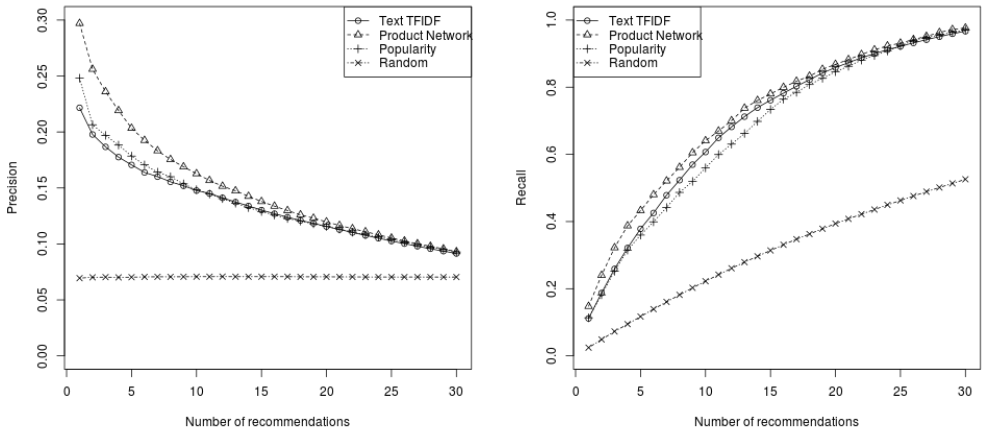


Figure 11 **The precision (a) and recall (b) of our text-based method against baselines on the auto dataset, given the number of recommendations made. Although the text-based approach initially underperforms the popularity-based approach, it exceeds it in recall after a few recommendations. This is likely due to the low number of recommendations that our models are able to make.**

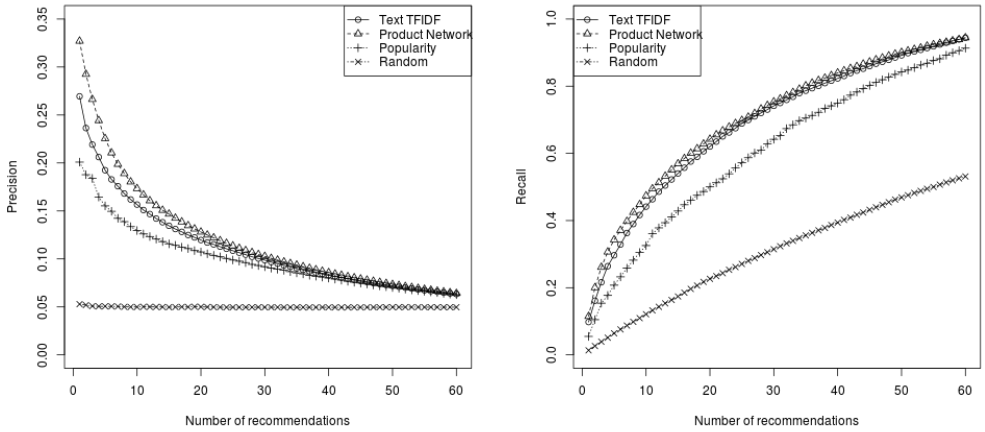


Figure 12 **The precision (a) and recall (b) of our text-based method against baselines on the clothes dataset, given the number of recommendations made. In this case, the text-based method does not perform quite as well as the product network, but consistently outperforms the popularity baseline, similar to its performance in the TV show testbed.**

systems, averaging over 10-fold cross-validation. Even though there are far fewer recommendations available for the systems to make, the text-based model still seems to outperform the popularity-based method in recall. Applying these same three methods over a collection of 83 clothing retailers and brands, we see (Figure 12) that there is similar ranking in performance for these methods. This consistently high performance across three very different product types suggests that the method will prove generally applicable over a wide variety of other types of products and services. This generalizability underlines the value this method can have for businesses and firms of all sorts.

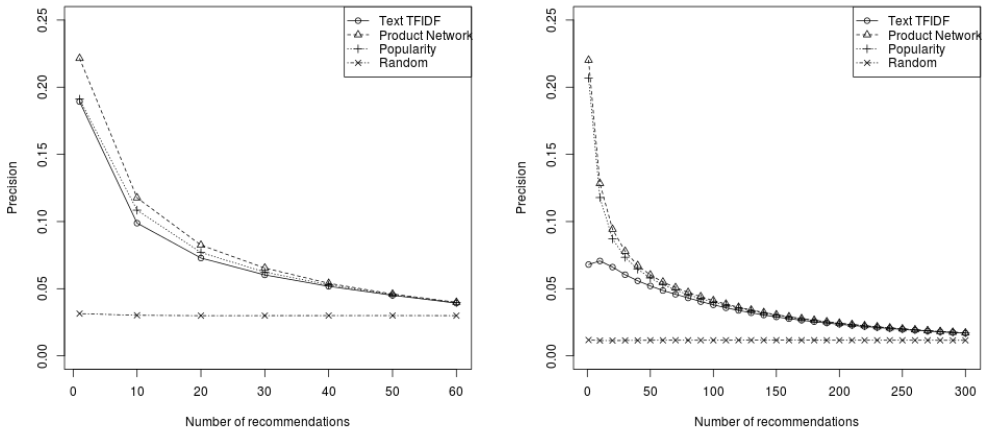


Figure 13 **(a) Precision of our text-based system against the product network, given number of recommendations, when considering a TV show as input and predicting which clothing brand a user also follows. (b) Likewise, the precision of these systems when using a clothing brand as input, and predicting which TV show the user follows.**

We then determined to see whether our model would be able to take the next step that would confirm it as the leading model for businesses to adopt, namely the ability to offer suitable product recommendations based on users' tastes in a separate product type. Specifically, we attempted to predict a TV show that a given user would like on the basis of a clothing brand they liked, or vice versa. In this case, we find that our text-based method still performs well. Figures 13a and 13b are plots of the precision of our system, given a TV show and clothing brand respectively as input.

As an additional step to compose the text-based and product network methods, we created a recommendation system that only makes text-based recommendations once the product network is unable to make additional recommendations. Figures 14 a-c show that, by combining these two methods in this way, we are able to make a far greater number of accurate recommendations even when the product network method is unable to make additional recommendations. They further show that for unpopular input shows, the gain over just using a collaborative filtering approach is much greater.

6. Conclusion

Using a data-collection approach we designed, we have collected a large and unique dataset to make and evaluate recommendations for products - in this case, TV shows, clothes and automobiles.

In this work, we capitalize on what we can learn about people's preferences for TV shows (and other brands) by what they freely reveal in the public forums of the social networking sites Twitter and Facebook. Additionally we capitalize on the aspects of their daily lives that these TV show followers mention on social media. Mining both the follower network and text data from this user-generated content, we both create and evaluate affinity networks for shows in the context of

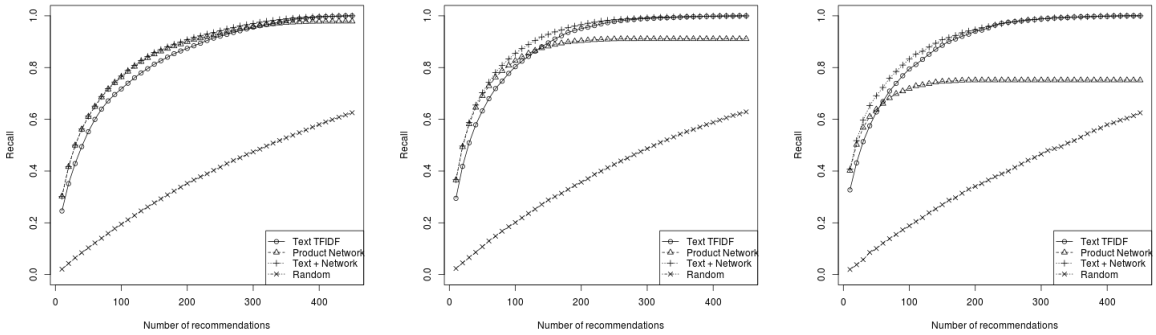


Figure 14 **The recall of combining both the product network and text-based method against the two alone, given the number of recommendations made. The plots correspond to (a) over all input shows, (b) the 50% least popular shows, and (c) the 25% least popular shows, respectively. Combining the two methods results in a greater improvement in recall when the number of recommendations is large and the popularity of the input show is low.**

using a novel RS approach. We show that the text and network data that users reveal is useful in predicting what shows users like as well as useful in aggregate for describing shows' viewing audiences. We show that words are indicative of both geographics and demographics as well as of viewers' interests, and that when extracted from training data sets the words of this UGC can be used to predict the demographic features of hold out sets of shows. We show that the text-based approaches we develop perform remarkably well against other RSs baselines often used in the literature. Finally, we demonstrate that the approach is easily extendable to other product contexts, specifically automobile and clothing retailers.

Extant research on recommender systems (RS) focuses mainly on improving recommendation accuracy. Little attention has been paid to using user-generated content to explain the affinity between brands, products, and services. We show that publicly available data can enable researchers and firms to both build models and evaluate their results against publicly available preferences, potentially for all brands and services that have an Internet presence, particularly on social media sites. By collecting data on hundreds of TV shows and millions of Twitter users, their tweets, and their social networks, we were able to build talkographic profiles for given brands. Proposing a privacy-friendly approach to extracting meaning from user-generated content, we show that user-generated content represents the interests, demographics, and geographics of the user base. This enables us to construct a talkographic profile of customers, viewers, and followers of TV shows and brands.

We highlight the fact that user-generated content has value for both consumers and firms, answering the open question of whether it could possess any at all. For consumers, the advantage lies in that user-generated content can provide them with better recommendations of products and services they might enjoy. Firms, meanwhile, can identify and quantify features of their consumer

base and use the aggregate-level profiles to calculate the affinity between their brand and others, allowing them to find new methods to build customer loyalty and differentiate themselves from their competition.

A major distinction between our work and previous work in this field is that we need neither an ontology for brands nor a set of pre-specified product-related keywords to mine the user-generated text. Our approach is both general and flexible, able to be extended across all brands, products and services in all areas. We have demonstrated that with our method, it is possible to apply features learned in one domain to another one - for instance, to calculate the similarity between a TV show and a clothing retailer. To the best of our knowledge, this is the first work to represent the audience of a brand or service with the use of talkographic profiles incorporating features of all aspects of consumers' daily lives. This all-encompassing approach has rich implications for all industries with customers who freely reveal their association with products and the details of their daily lives online, offering a method to use this new source of information in countless ways that will benefit both provider and user, while guaranteeing the latter's privacy. This is not to say that our findings are without limitations; but these provide fertile ground for further research. The results we present are based on three very specific contexts: the Twitter presence of TV shows, automobile manufacturers, and clothing retailers. They further rely on users' propensity for revealing details of their true preferences, essentially assuming their honesty. In environments where purchases take place less often (for instance, acquiring major household appliances)or where item prices are noticeably higher (luxury items), it is possible that consumer-friending behavior may be different. It is also possible for consumers to friend high-status brands to heighten their standing among acquaintances, rather than brands they can actually afford. Obviously, further research is needed into the implications of these and similar patterns of behavior for the creation of talkographic profiles.

Despite the central role played by the development of a new recommendation system in this paper, the true focus of this research was not the creation of a new RS but on demonstrating the value of user-generated content to the construction of viewer base profiles. In future work we plan to optimize the RSs' performance based on both user-generated content and personalized information. We also intend to test our approach in a laboratory setting in order to determine whether our approach does in fact yield recommendations better appreciated by consumers than those presented by crowdsourcing on Twitter, primarily through co-occurring links between followers. This lab setting will also allow us to inject further details into the model, for instance by taking the time of day into account to perform context-aware recommendations such as recommending TV shows currently playing.

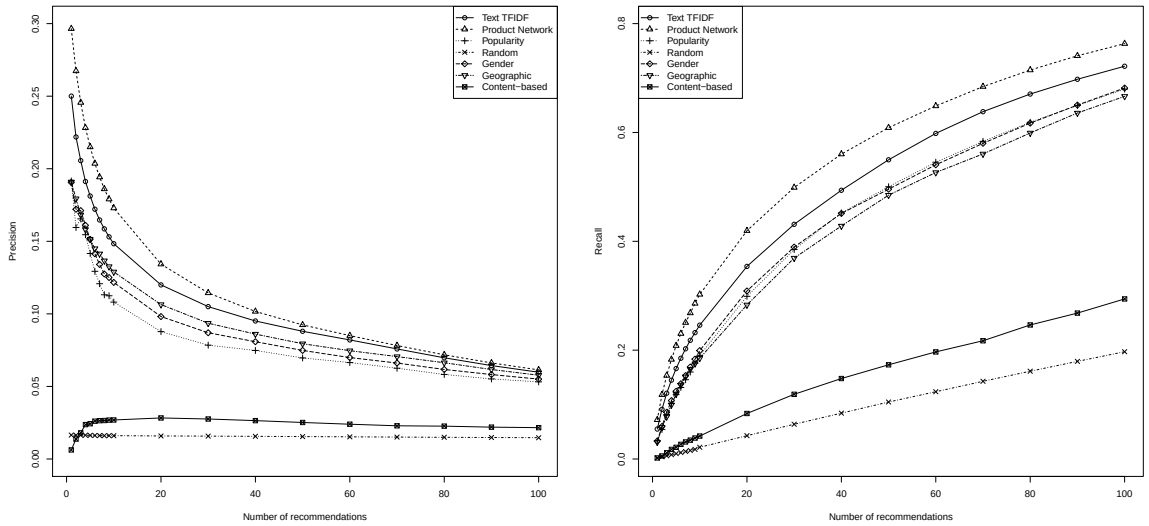


Figure 15 **(a) The precision of the proposed methods against all baselines, and (b) the recall of these methods against each other. Content-based refers to the RS where similarity between TV shows was defined as the weighted product of similarity scores along features of TV shows as listed on IMDb. This method only performs marginally better than the random baseline. Geographic refers to an RS similar to the popularity-based method, but recommends the most popular shows that users from the same state as the input user follow. Similarly, Gender, recommends the most popular TV show that users of the input user's gender follow. Although Gender seems to perform slightly better than Popularity, Geographic's results are mixed. This is likely due to the small number of users we are able to accurately infer their state for.**

However, this further research turns out, whatever results are found and incorporated to refine both model and method, we believe that we have demonstrated that when properly analysed and used, user-generated content can provide immense value to firms and businesses, allowing for greater targeting of audiences, differentiation from the competition, and increased loyalty from customers. We know of no other method that approaches the completeness and range of ours, and suggest that talkographics will come to be recognized as an essential, integral part of marketing for all companies and industries possessing an internet presence.

Appendix A: Baseline Performance

We also compared our text-based method against other baselines RSs. The precision and recall of these baselines are included in Figure 15. Note that none of the baselines perform as well as the proposed text-based method.

Appendix B: Bigram Performance

We performed our approach using bigrams (frequency of two words in the text) instead of unigrams with limited success. The relative performance can be found in Figure B.

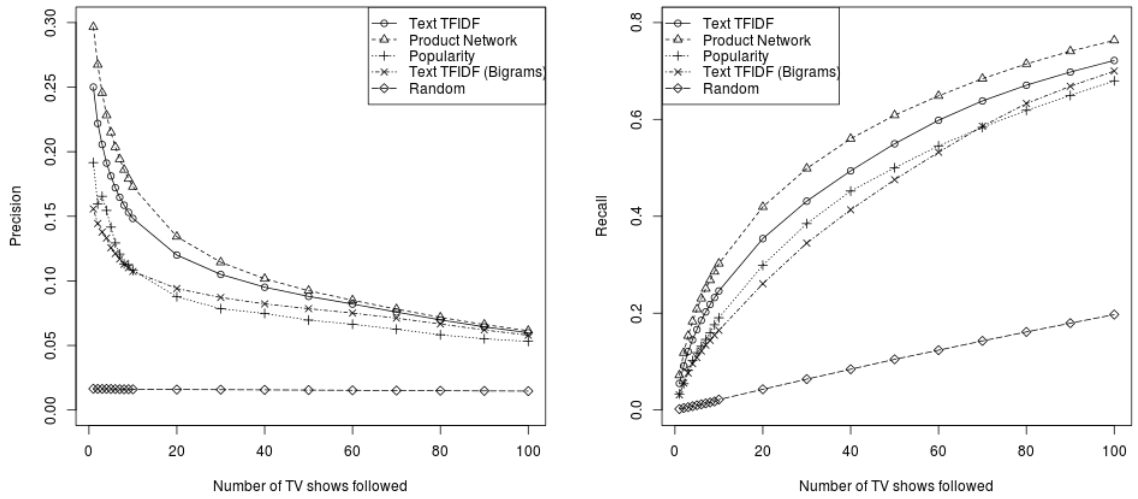


Figure 16 **(a) The precision of bigrams against our methods, and (b) the recall of bigrams compared to our method.**

Appendix C: RS Similarity Matrix Visualization

An interactive visualization of the similarity matrices learned by the product network and text-based methods can be found at: `http://108.167.179.169/\~shawndra/jp/adrian/network_vis/interactive_network_recommender/` The visualization has been tested under the Firefox and Chrome browsers.

Acknowledgments

The authors gratefully acknowledge the Google and WPP research grant, Wharton Junior Faculty Dean's Award.

References

Abel, Fabian, Qi Gao, Geert-Jan Houben, Ke Tao. 2011. Analyzing user modeling on twitter for personalized news recommendations. *User Modeling, Adaptation, and Personalization*. Springer, New York, 1–12.

Adomavicius, Gediminas, YoungOk Kwon. 2012. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* **24**(5) 896–911.

Adomavicius, Gediminas, Ramesh Sankaranarayanan, Shahana Sen, Alexander Tuzhilin. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems* **23**.

Adomavicius, Gediminas, Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* **17**(6) 734749.

Ansari, Asim, Skander Essegaier, Rajeev Kohli. 2000. Internet recommender systems. *Journal of Marketing Research* **37**(3) 363–375.

Ansari, Asim, Carl F. Mela. 2003. E-Customization. *Journal of Marketing Research* **40** 131–145. doi: 10.1509/jmkr.40.2.131.19224.

Archak, Nikolay, Anindya Ghose, Panagiotis G. Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science* **57**(8) 14851509.

Atahan, Pelin, Sumit Sarkar. 2011. Accelerated learning of user profiles. *Management Science* **57**(2) 215239.

Balabanovic, Marko, Yoav Shoham. 1997. Fab: Content-based, collaborative recommendation. *Communications of the ACM* **40**(3) 6672.

Breese, John S., David Heckerman, Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 4352.

Chen, Jilin, Rowan Nairn, Les Nelson, Michael Bernstein, Ed Chi. 2010. Short and tweet: Experiments on recommending content from information systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, New York, 1185–1194.

Chorianopoulos, Konstantinos, George Lekakos. 2008. Introduction to social tv: Enhancing the shared experience with interactive tv. *International Journal of HumanComputer Interaction* **24**(2) 113120.

Das, Sanjiv R., Mike Y. Chen. 2007. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science* **53** 1375–1388. doi:10.1287/mnsc.1070.0704.

De Bruyn, V, John C. Liechty, Eelko K. R. E. Huizingh, Gary L. Lilien. 2008. Offering online recommendations with minimum customer input through conjoint-based decision aids. *Marketing Science* **27**(3) 443460.

Decker, Reinhold, Michael Trusov. 2010. Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing* **27**(4) 293–307. doi:10.1016/j.ijresmar.2010.09.001.

Dellarocas, Chrysanthos. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science* **49**(10) 1407–1424.

Dörre, Jochen, Peter Gerstl, Roland Seiffert. 1999. Text mining: Finding nuggets in mountains of textual data. *Proceedings of the Fifth ACM SIGKDD International Conferences on Knowledge Discovery and Data Mining*. ACM, 398401.

Ducheneaut, Nicolas, Robert J. Moore, Lora Oehlberg, James D. Thornton, Eric Nickell. 2008. Social tv: Designing for distributed, sociable television viewing. *International Journal of HumanComputer Interaction* **24**(2) 136154.

Eliashberg, Jehoshua, Sam K. Hui, Z. John Zhang. 2007. From story line to box office: A new approach for green-lighting movie scripts. *Management Science* **53** 881–893. doi:10.1287/mnsc.1060.0668.

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, Benjamin Segal. 2010. Managements tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* **15**(4) 915953.

Feldman, Ronen, James Sanger. 2006. *The Text Mining Handbook*. Cambridge University Press, New York.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fleder, Daniel, Kartik Hosanagar. 2009. Blockbuster cultures next rise or fall: The impact of recommender systems on sales diversity. *Management Science* **55**(5) 697712.

Geerts, David, Dirk De Grooff. 2009. Supporting the social uses of television: Social heuristics for social tv. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 594–604.

Ghose, Anindya, Panagiotis G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering* **23**(10) 1498–1512. doi:http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.188.

Ghose, Anindya, Panagiotis G. Ipeirotis, Beibei Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* **31**(3) 493520.

Gross, Tom, Mirko Fetter, Thilo Paul-Stueve. 2008. Toward advanced social tv in a cooperative media space. *International Journal of HumanComputer Interaction* **24**(2) 155–173.

Hannon, John, Mike Bennett, Barry Smyth. 2010. Recommending twitter users to follow using content and collaborative filtering approaches. *Proceedings of the Third ACM Conference on Recommender Systems*. New York, New York, 199–206.

Hannon, John, Kevin McCarthy, Barry Smyth. 2011. Finding useful userson twitter: Twittomender, the followee recommender. *Advances in Information Retrieval* **6611** 784–787.

Hill, Shawndra, Adrian Benton. 2012. Social tv: Linking tv content to buzz and sales. *International Conference on Information Systems*.

Hu, Nan, Ling Liu, Jie Jennifer Zhang. 2008. Do online reviews affect product sales? the role of reviewer characteristics and temporal effects. *Inf. Technol. and Management* **9**(3) 201–214. doi: 10.1007/s10799-008-0041-2. URL `http://dx.doi.org/10.1007/s10799-008-0041-2`.

Lee, Thomas Y., Eric T. BradLow. 2011. Automated marketing research using online customer reviews. *Journal of Marketing Research* **48**(5) 881–894. doi:10.1509/jmkr.48.5.881.

Liu, Bing. 2011. Opinion mining and sentiment analysis. Bing Liu, ed., *Data Centric Systems and Application: Web Data Mining, 2nd ed.*. Springer-Verlag, Berlin, 459526.

McGinty, Lorraine, Barry Smyth. 2003. On the role of diversity in conversational recommender systems. *Proceedings of the Fifth International Conference on Case-Based Reasoning*. Springer-Velag, 276290.

Michelson, Mathew, Sofus A. Macskassy. 2010. Discovering users' topics of interest on twitter: A first look. *Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data*.

Mitchell, Keith, Andrew Jones, Johnathan Ishmael, Nicholas J.P. Race. 2010. Social tv: Toward content navigation using social awareness. *Proceedings of the 8th International Interactive Conference on Interactive TV and Video*. ACM, 283292.

Montgomery, Alan L., Kannan Srinivasan. 2002. *Learning about customers without asking*. eBRC Press.

Mooney, Raymond J., Loriene Roy. 1999. Content-based book recommending using learning for text categorization. *Proceedings Of The Fifth ACM Conference On Digital Libraries*. ACM Press, 195–204.

Morales, Gianmarco De Francisci, Aristides Gionis, Claudio Lucchese. 2012. From chatter to headlines: Harnessing the real-time web for personalized news recommendation. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. New York, New York, 153–162.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, Mashe Fresko. 2012. Mine your own business: Market-structure surveillance through text mining. *Marketing Science* **31**(3) 521–543.

Palmisano, Cosimo, Alexandrer Tuzhilin, Michele Gorgoglione. 2008. Using context to improve predictive modeling of customers in personalization applications. *IEEE Transaction on Knowledge and Data Engineering* **20**(11) 15351549.

Pang, Bo, Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* **2**(1-2) 1135.

Pankong, Nichakorn, Somchai Prakancharoen. 2011. Combining algorithms for recommendation system on twitter. *Advanced Materials Research* **403-408** 36883692.

Panniello, Umberto, Michele Gorgoglione. 2012. Incorporating context into recommender systems: An empirical comparison of context-based approaches. *Electronic Commerce Research* **12**(1) 1–30.

Pazzani, M. J., D. Billsus. 2007. Content-based recommender systems. *The Adaptive Web: Methods And Strategies Of Web Personalization*, *Lecture Notes in Computer Science*, vol. 4321. Springer-Verlag, 325341.

Phelan, Owen, Kevin McCarthy, Mike Bennett, Barry Smyth. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. *Proceedings of the 33rd European Conference on Advances in Information Retrieval*. Springer-Verlag, 448–459.

Phelan, Owen, Kevin McCarthy, Barry Smyth. 2009. Using twitter to recommend real-time topical news. *Proceedings of the Third ACM Conference on Recommender Systems*. New York, New York, 385–388.

Sahoo, Nachiketa, Ramayya Krishnan, George Duncan, Jamie Callan. 2008. On multi-component rating and collaborative filtering for recommender systems: The case of yahoo! movies.

Schwartz, Andrew H., Johannes C. Eichstaedt, Lukasz Dziurzynski, Eduardo Blanco, Margaret L. Kern, Michal Kosinski, David Stillwell, Lyle H. Ungar. 2013. Toward personality insights from language exploration in social media. *Proceedings of the AAAI 2013 Spring Symposium on Analyzing Microtext*.

Shardanand, Upendra, Pattie Maes. 1995. Social information filtering: Algorithms for automating word of mouth. *CHI 95 Proceedings*. ACM Press, 210217.

Soboroff, Ian, Charles K. Nicholas. 1999. Combining content and collaboration in text filtering. *Proceedings of the IJCAI99 Workshop on Machine Learning for Information Filtering*. 8691.

Sun, Aaron R., Jiesi Cheng, Daniel Dajun Zeng. 2010. A novel recommendation framework for micro-blogging based on information diffusion.

Ying, Y., F. Feinberg, M. Wedel. 2006. Leveraging Missing Ratings to Improve Online Recommendation Systems. *JOURNAL OF MARKETING RESEARCH* **43**(3).