Identifying Fashion Accounts in Social Networks

Doris Jung-Lin Lee, Jinda Han, Dana Chambourova, Ranjitha Kumar
Department of Computer Science
University of Illinois, Urbana-Champaign
jlee782,jhan51,dchambourova,ranjitha@illinois.edu

ABSTRACT

Fashion and style are characterized by the ebb and flow of trends. With the rise of social media, fashion blogs, and the fast-fashion movement, bottom-up fashion trends are emerging at an ever-increasing rate. Recognizing these trends as they happen — and the influencers that create them — is challenging for retailers and consumers alike. As a first step, this paper presents a classifier for identifying fashion-related accounts on social media. To develop this classifier, we collected a dataset of 10k Twitter accounts using a snowball sampling approach, and crowdsourced ground-truth labels for them. Based on this training data and a set of content-based features, we trained a classifier that identifies whether or not a Twitter account is fashion-related. In the future, we hope to leverage this classifier to identify key fashion influencers and conduct large-scale monitoring of fashion trends.

KEYWORDS

fashion, social networks, trendsetters

ACM Reference format:
Doris Jung-Lin Lee, Jinda Han, Dana Chambourova, Ranjitha Kumar. 2017. Identifying Fashion Accounts in Social Networks. In *Proceedings of ACM KDD 2017 Machine learning meets fashion Workshop, Halifax, Nova Scotia, August 2017 (ML4Fashion'17),* 5 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

The rise of social media and fashion blogs has affected the creation and diffusion of fashion trends. These new networks distribute influence across a broader set of tastemakers, begetting rapidly changing, bottom-up trends. Designers, retailers, and consumers find it challenging to recognize trends — and the trendsetters that create them — in this new, fast-paced industry.

As a first-step to tracking trends and discovering trendsetters, this paper introduces a content-based classifier for identifying fashion-related user accounts on Twitter. To train this classifier, we created a training dataset by snowball-sampling 10K accounts from Twitter, starting from a seed set of 10 influential fashion accounts [16]. Via Amazon Mechanical Turk, we crowdsourced ground-truth labels for this dataset (i.e., whether or not an account was relevant to fashion). Leveraging this training dataset, we trained support-vector machine and naive Bayes classifiers with content-based, interpretable features.

From the crowdsourcing task, we identified 2,734 fashion-related Twitter user accounts (out of 10,230 accounts). We demonstrate that on average our classifier has precision and recall rates of 75% and 72%, respectively. We show that many Twitter fashion accounts contain media (images, video) or external links to media. Future work should incorporate media-based features to improve classifiers.

ML4Fashion'17, August 2017, Halifax, Nova Scotia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.475/123_4

2 METHODS

We are building a classifier for fashion. First, we are going to create a training dataset. As part of that, we have to crawl a bunch of accounts from Twitter. Then we will crowdsource ground-truth labels for these accounts on Amazon Mechanical Turk. Moreover, we have to figure out the feature set, and compute the feature set over the fashion accounts. After we have the training dataset, then we will train the classifier.

2.1 Data Collection

Our goal for sampling the social network is to obtain a subgraph of fashion-related accounts and their associated information as a raw dataset for account discovery. Since the number of fashion Twitter account is much smaller than the total network size of Twitter and due to the rate-limiting nature of the Twitter API[1], we decided to take a content-based, snowball sampling technique to collect our initial raw dataset.

We are interested in developing a sampling approach that depends on the amount of "fashion information" each node contains. Most existing sampling algorithm tries to preserve some network structure of the original network (e.g. degree, centrality, cluster, etc.), but are independent of node information[3, 5]. In order to capture whether an account is fashion-related or not, we use two criterion: 1) number of fashion words from all the tweets posted by a user exceeds a threshold and 2) whether the word 'fashion' is contained in their profile description. Our sampling algorithm starts from a small set of fashion accounts as seeds based on an expert-curated list by retail-marketing experts [2]. For each account our classifier examines whether the accounts that they follow are fashion accounts. A fashion account is defined as a user whose tweets contain more fashion words than a threshold value or have the word fashion in their profile. Our dataset consist of the first level of this crawl.

We ran experiments to fine tune the fashion measure that will be used as a threshold in the graph crawling, then we are able to create

[1] One can only request a maximum of 3200 tweets for an individual user, which for very active users only allows us to collect data from only a couple months of activity. We bypass the problem by storing the necessary meta-data on disk and creating an architecture that allows for cron jobs.
[2] https://www.ometria.com/blog/top-fashion-twitter-accounts

the user graph through the snowballing approach. The Twitter API allows to search for the 15 most relevant users when provided a topic[3] . We use fashion and non-fashion related keywords to search for a list of fashion and non-fashion users as ground-truth. Then we conducted an experiment where we compared the performance of our crawler on different parameter settings to improve the precision and recall of these ground-truth fashion accounts.

We use modified version of fashion vocabulary developed by Vaccaro et al.[9] to determine whether a word is fashion-related or not. Our dataset combines the Vaccaro et al. style and element vocabulary collected from Polyvore, a popular fashion-based social network. We filter out generic words that are found to be highly overlapping amongst both the fashion and non-fashion users' tweets, including stop words and non-discriminative terms that could be used in generic words.

In order to do this, we collected a list of 77 ground truth fashion users[4] and a list of 75 non-fashion users gathered from curated list of top Twitter accounts for various non-fashion topics, such as sports[5], technology[6], science[7], and politics[8]. We scraped the tweets from these account using the method described in Section 2.1. After filtering out the stopwords, we selected the top 200 words from the non-fashion tweet data as non-discriminative words. We checked that the top 50 words from both the fashion and non-fashion groups are fairly discriminative when we used the non-discriminative words to filter the fashion vocabulary. Our filtered fashion vocabulary contains 9806 words that are highly relevant to fashion, as show in Figure 1.

FILTERED FASHION VOCABULARY (Total 9806 words)
stylish, pajama, party-costume, fancy-footwear, bow-belt, bathing-suit-swimwear, clothes, elegant, summer, boyfriendjeans, colourful, blouse, scrunch-bikini, overknee, animated, reversible, dungaree, streetstyle, guilloche, hoodie, peacoat, modern...

NON-DISCRIMINATIVE WORDS (Total 200 words)
big, hear, work, game, life, play, man, long, keep, world, real, another, watch, image, post, list, season, national, beautiful...

STOP WORDS (Total 75 words)
the, to, of, and, in, for, on, you, my, at, all, think, or, see, time, think, look, only...

Figure 1: Example non-discriminative words, stopwords, and filtered fashion vocabulary.

2.2 Crowdsourced Dataset

We conduct a crowdsourced data collection on Amazon Mechanical Turk where we ask workers to classify whether an given account is fashion, non-fashion, or inaccessible. An inaccessible account is one that is either a deleted or private account. For each task, we show the worker a list of 10 Twitter accounts as shown in Figure 2 and the workers were asked to classify whether the account is fashion-related or not. The workers were compensated 10 cents for each task.

We regard an account is deemed a fashion account if at least two out of three crowdworkers classify it as a fashion account. Also, we inserted an attention check question with known responses in each HIT for quality evaluation. Finally, we collected a total of 30510 responses. Out of the 10230 unique labeled accounts, 26.72 %(2734) of the dataset is labeled as fashion accounts and the rest labeled as non-fashion.

2.3 Classification

To classify the Twitter accounts as fashion or non-fashion, we use the features based on the data collected from the crawler which consists of all the recent tweets posted by the account[9].

From the account information, we define fashion counts divided by total number of words in all tweets as normalized fashion counts, and we use normalized fashion counts and number of tweets and the user's profile description as an indicator of how much fashion content an account contains. Another feature in used for classification is the normalized fashion counts, which is computed as the total number of fashion words divided by the total number of words over all tweets. Both the denominator of the normalized fashion counts and number of tweets is used as an indicator of the verbosity and posting-frequency of the account. The Twitter user profiles are short (maximum 160-character) descriptions, where users often describe their interest (sports,fashion) and occupational description (blogger, editor-in-chief). We binaries this feature by checking whether the word 'fashion' is contained in their profile description.

We use two separate machine learning algorithms for account classification: Naive Bayes (NB) and Support-Vector Machines (SVM). For Naive Bayes, we use LaPlace smoothing for regularization in the rare cases where the feature and class does not occur together. A linear parameter search shows that any non-zero smoothing parameter is sufficient in improving the model's performance. We use a SVM with an RBF kernel (with a coefficient γ) and a regularization constant C that controls the degree-of-freedom of the decision boundary. We perform grid search to determine the best parameter settings for C and γ. Our results is based on a setting of C = 4 and γ = 0.05.

3 RESULTS

3.1 Evaluation

We evaluated the performance of the classification algorithms using 10-fold cross validation using the best parameters setting described in the previous section. We sampled 5500 non-fashion accounts and 2734 fashion accounts in performing these evaluations. Table 1 summarizes these results. We will discuss the inherent reason for the low recall of fashion accounts in the following section.

[3]https://dev.twitter.com/rest/reference/get/users/suggestions
[4]goo.gl/aPv7WW,goo.gl/LZeFkv,goo.gl/3dor03
[5]https://www.si.com/sports-illustrated/twitter-100-2014
[6]goo.gl/45tLav
[7]goo.gl/5QtcEa
[8]goo.gl/eNXDhE

[9]The Twitter API limits the capability of its backwards history search to a maximum of 3,200 most recent Tweets for each user.

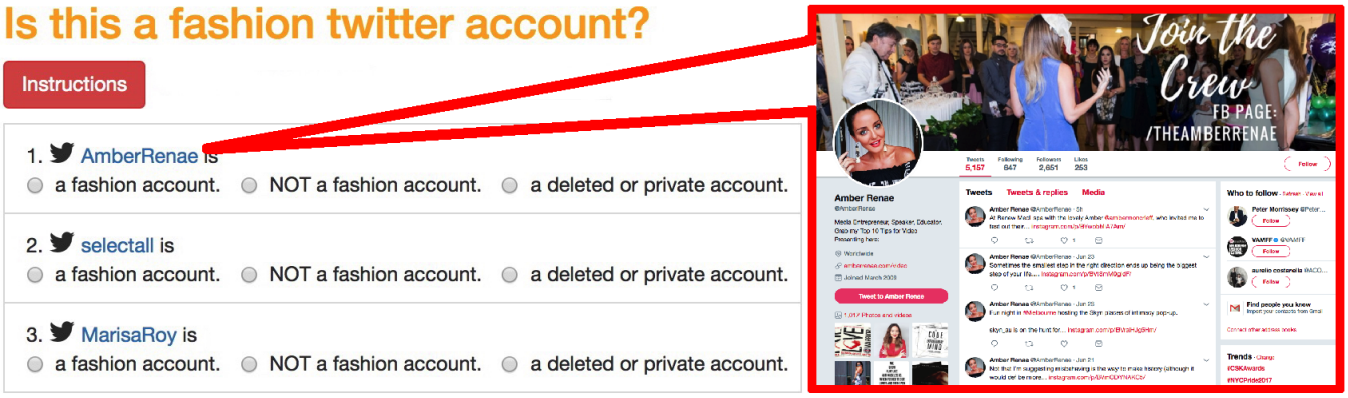

Figure 2: Screenshot of sample web interface used for crowdsourced data collection showing a sample Twitter account.

Table 1: Performance of classification algorithm.

	Precision	Recall	F1
Non-fashion (SVM)	0.72	0.97	0.82
Fashion (SVM)	0.78	0.24	0.37
Average (SVM)	0.74	0.72	0.67
Non-fashion (NB)	0.71	0.98	0.82
Fashion (NB)	0.82	0.20	0.32
Average (NB)	0.75	0.72	0.66

3.2 Findings

By analyzing the accounts where classifier makes a wrong prediction, we highlight the challenges for studying social media account discovery in fashion using our approach and how our work takes a first step in this direction in tackling these problems, as summarized in Figure 3.

Category	Examples
Image-heavy	@LaurieTrott, @jacvanek
Link-heavy	@annadellorusso,@wanderlustandco
off-topic fashion workers	@adelefashionmag,@JessC_M
Fashion+X	@ALTOmagazine,@ProjectMMNYC

Table 2: Example media-heavy, off-topic, multi-topic fashion accounts.

Media-heavy Account: There were many users that had low fashion word counts even though they belonged to the fashion positive set. By analyzing specific instances of these accounts, we find there are users who maintain their fashion relevance through the usage of media or external links without referencing the fashion vocabulary we use. Image-heavy accounts and Instagram external links are common especially due to the visual nature of fashion. Since the media content is often self-explanatory, tweets with media content are often associated with not many descriptive tweet text related to fashion. This is especially common for twitter accounts associated with clothing or items website where an image of a new product alone is enough to stimulate discussion and start a trend. Another common usage of images is photographs of Internet fashion celebrities model with trendy clothing. These Twitter posts often contain Instagram links and non-descriptive tweet descriptions.

Links are common for fashion users that have their own publication platform and simply use Twitter as a platform for reaching a broader set of audience and attracting readers. These users include bloggers and official magazine accounts. A short descriptive text is associated with linked tweets. Since these media-heavy accounts don't often have enough fashion-related descriptive text, our classifier is unable to identify them which largely accounts for the low recall of fashion users compared to other metrics.

We conduct a post-analysis to understand how many accounts from the misclassified cases are media heavy accounts by examining 100 tweets from 100 account for both the false positive and false negative cases. We find that for the false negatives: 14.89% of the tweets of these contains an image, 0.42% contains a video, 51.71% contains an external URL and 4.5% contains an Instagram link. For the false positives, 24.41% of the tweets of these contains an image, 1.12% contains a video, 65.54% contains an external URL and 11.05% contains an Instagram link.

Off-topic fashion workers: We find that it is fairly common for a fashion user to be identified as fashion due to their occupational description on their profile descriptor. However, some of these users uses their Twitter account to discuss things related to their personal life unrelated to fashion. For example, @adelefashionmag is a fashion creative art director but her Twitter posts includes complaints about train delays and bad customer services. The noise in these cases may account for the variance in our predictions. For the purpose of monitoring the fashion activity in a social network, they would still be important to include them because they may be connected to other fashion personnel. A potential future work includes a more fine grained classification for whether someone is fashion-related by occupation or by tweet content.

Fashion+X interest page: There are many bloggers, magazines, or special interest pages on Twitter that post tweets related to multiple topics. For example, several luxury magazines (@ALTOmagazine, @ProjectMMNYC) contain topics related to fashion, art, and design. Another common topic combinations is fashion, cosmetics and

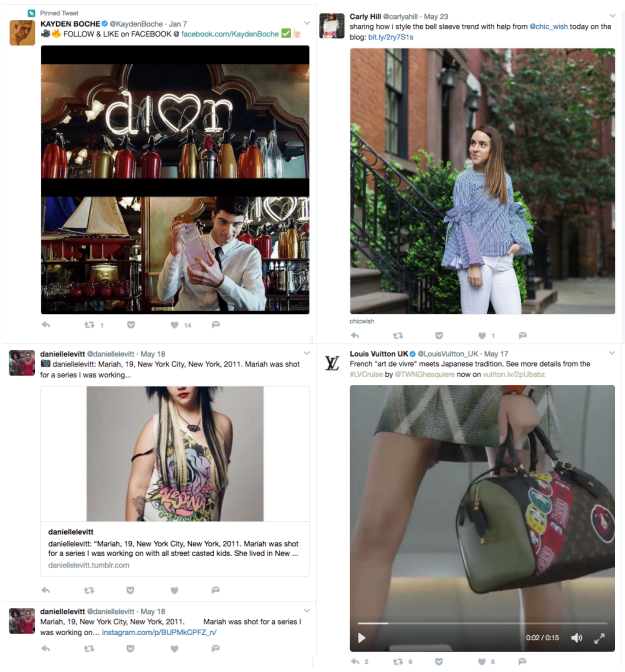

Figure 3: Due to the 140 word-limit on Twitter, a diverse set of fashion users uses media attachments as a way to further their expression, including a model for Dior TV ads and runways(@KaydenBoche), a fashion blogger (@caryahill), a high class photographer(@danielllelevitt), and a famous, high-end clothing company(@LouisVuitton_UK). This poses a challenge to our text-based classification approach.

beauty products(@CosmeticccBlog,@FierceBeauty101). These tangentially fashion accounts contributed largely to the false positive rates.

4 DISCUSSION AND FUTURE WORK

After obtaining the selected subset of influencers in the social network and understanding the types of fashion accounts and their peculiar usage on Twitter, we propose several promising directions of future work that we plan to explore.

4.1 Influencers in networks

In this paper, we are interested in discovering the influencer in a fashion-related network in order to find key personnels that generate significant changes (e.g. trend-setting). The problem of identifying influencers and information cascades in a social network has been well-studied [3, 4]. People connected by network influences other people's behaviors and decision, such as whether or not to adopt a fad in the case of fashion. A content's popularity in network is a result of network imbalance and the instability propagate through diffusion in a network to gain wide visibility. These cascading effects have been used by marketers to promote new products through a the idea of "viral marketing"[3, 11], which promotes a small number of key members of a network to adopt a new product and thereby causing a cascade of adoption at the population level. The generic problem of finding important nodes in a network have also been extensively studied in the context of finding popular web pages in a network[8, 13]. As an extension to the HITS model [8], [18] proposes CuRank for identifying curators in network by accounting for the timeliness and curatorial taste of users in the network. While our paper focuses on the identification of fashion accounts, a future direction includes developing a computational approach for ranking the importance of a user in a fashion-based network in order to further support more accurate decision-making and knowledge-discovery.

4.2 Fashion Trend Discovery

Traditionally, merchandisers and designers have used sales statistics and market surveys to understand consumer behavior and forecast upcoming trends. With the advent of social media, it is challenging for designers or marketing experts to keep up with the rich and diverse signals indicating subtle changes in a consumer's taste in fashion required for making important business decision[6]. Therefore, recent work has leveraged scalable, data mining approaches to study the multi-scaled problem of how fashion have changed over time. Visual evolution of fashion have been captured by large-scale, photo collections [7, 15, 17]. Social media signals[12], fashion-expert generated content[2, 10], search queries [1], and customer feedback [14] have also revealed patterns due to seasonal effects, help detect emerging trends, and fads. While these data-driven approaches reveal important insights, they often ignore the roles of trendsetter in how they give rise to fashion trends. Our work takes the first step towards this direction: by identifying and by understanding the common patterns of online actions and motivations of trendsetters, we hope to discover more principled approach for modeling and prediction of fashion trends.

5 CONCLUSION

In this paper, we develop a classifier for identifying whether a Twitter account is fashion-related. To create our raw dataset for classification, We use a content-based snowballing method to collect the potentially relevant Twitter accounts and used crowdsourcing to collect labels for this dataset. Using account features based on the tweet and profile information, we use support-vector machine and Naive Bayes to conduct the classification.

By understanding the inherent errors that we observe in our classifier, we discovered several interesting behaviors of fashion users on Twitter: 1) many fashion Twitter accounts are media-heavy 2) some accounts are related to fashion by occupation rather than by tweet relevancy and 3) blogs and magazines accounts can cover more than one topic that includes fashion.

After understanding the types of fashion accounts and their peculiar usage on Twitter, we propose several promising directions of future work that we plan to explore. Social media platforms such as Instagram, Pinterest or Snapchat are increasingly focusing their entire communication strategy on visual communication. As use the exploration of the word-based Twitter space as a foundation to start exploring the fashion ecosystem, we also recognize the need to extrapolate our methodology to non-verbal communication methods. The poor performance of the classifier on image-based

accounts points to the need for training classifiers that incorporates visual information. We could envision an image-text hybrid algorithm as our existing classifier still performs well for the majority of the Twitter accounts which are text-based. We believe that exploring visual dimension of fashion would uncover more interesting insights that augment the techniques proposed in this paper.

REFERENCES

[1] 2015. Google Fashion Report 2015. (2015). https://think.storage.googleapis.com/docs/google-fashion-trends-report-spring2015.pdf

[2] Samaneh Beheshti-kashi, Michael Lütjen, Lennard Stoever, and Klaus-dieter Thoben. 2015. TrendFashion - A Framework for the Identification of Fashion Trends. (2015).

[3] Jon Kleinberg David Easley. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press.

[4] Malcolm Gladwell. 2002. *The Tipping Point: How Little Things Can Make a Big Difference.* Back Bay Books.

[5] Mark Granovetter. 1976. Network Sampling: Some First Steps. *Amer. J. Sociology* 81, 6 (1976), 1287–1303. https://doi.org/10.1086/226224 arXiv:https://doi.org/10.1086/226224

[6] Kate Hart. 2015. The Future of Fashion Forecasting. (2015). https://www.notjustalabel.com/editorial/the-future-of-fashion-forecasting

[7] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. *Www* (2016), 507–517. https://doi.org/10.1145/2872427.2883037 arXiv:1602.01585

[8] Jon M. Kleinberg. 1999. Authoritative Sources in a Hyperlinked Environment. *J. ACM* 46, May 1997 (1999), 668–677. https://doi.org/10.1.1.120.3875

[9] Kristen Vaccaro, Sunaya Shivakumar, Ziqiao Ding, Karrie Karahalios and Ranjitha Kumar. 2016. The Elements of Fashion Style. *Uist* (2016), 777–785. https://doi.org/10.1145/2984511.2984573

[10] Holly Lauridsen. 2014. What's in Vogue? Tracing the evolution of fashion and culture in the media. (Sep 2014). http://news.yale.edu/2014/09/05/what-s-vogue-tracing-evolution-fashion-and-culture-media

[11] J. Leskovec, L. A. Adamic, and B. A. Huberman. 2007. The Dynamics of Viral Marketing. *ACM Transactions on the Web (TWEB)* 1, 1 (2007), 1–39. https://doi.org/10.1145/1232722.1232727 arXiv:physics/0509039

[12] Lydia Manikonda, Ragav Venkatesan, Subbarao Kambhampati, and Baoxin Li. 2015. Trending Chic: Analyzing the Influence of Social Media on Fashion Brands. (2015). arXiv:1512.01174

[13] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems* 54, 1999-66 (1998), 1–17. https://doi.org/10.1.1.31.1768 arXiv:1111.4503v1

[14] Roberto Sanchis-Ojeda, Daragh Sibley, and Paolo Massimi. 2016. Detection of fashion trends and seasonal cycles through the analysis of implicit and explicit client feedback. *KDD Fashion Workshop* (2016).

[15] Alexei A. Efros Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin. 2015. A Century of Portraits : A Visual Historical Record of American High School Yearbooks. *Extreme Imaging Workshop, International Conference on Computer Vision, ICCV* (2015). https://doi.org/10.1109/ICCVW.2015.87 arXiv:arXiv:1511.02575v1

[16] Hannah Stacey. 2017. Top Fashion Twitter Accounts: 10 Must-Follow Fashion Industry Insiders. (Jun 2017). http://blog.ometria.com/top-fashion-twitter-accounts

[17] Sirion Vittayakorn, Alexander C Berg, and Tamara L Berg. 2016. When Was That Made? *arXiv* (2016). arXiv:1608.03914

[18] Haizi Yu, Biplab Deka, Jerry O. Talton, and Ranjitha Kumar. 2016. Accounting for Taste : Ranking Curators and Content in Social Networks. *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), Pages 2383–2389. https://doi.org/10.1145/2858036.2858219