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
The fashion industry is characterized by the ebb and flow of trends. With the rise of social media, fashion blogs, and fast-fashion movement, bottom-up fashion trends are emerging at an ever-increasing rate. Identifying new influencers and trends as they happen is challenging for retailers. 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 content-based snowball sampling approach, and crowdsourced ground-truth labels for these accounts. We train a classifier that identifies whether a Twitter account is fashion-related and evaluates the efficacy of our method. 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:
https://doi.org/10.475/123_4

1 INTRODUCTION
"Clothes, like architecture and art, reflect an era.”
- Azzedine Alaia

With the rise of social media, fashion blogs, and fast-fashion movement, bottom-up fashion trends are emerging at an ever-increasing rate. Although the nature of fashion and trends has been changing, our way of analyzing and reporting them is still rooted in the times when it was only influenced by a select few. Trends are still discussed with several weeks or months delay. 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 content-based snowball sampling approach, and crowdsourced ground-truth labels for these accounts. We train a classifier that identifies whether a Twitter account is fashion-related and evaluates the efficacy of our method. We hope to leverage this classifier to identify key fashion influencers and conduct large-scale monitoring of fashion trends.

In this paper, we train a classifier for discovering fashion accounts on Twitter using support-vector machines and Naive Bayes with discriminative features that signals whether a user is fashion-related or not in a social network. We identify a list of 2734 Twitter accounts that are relevant to the understanding of fashion trends. Detecting these fashion trends requires constant monitoring of these accounts to understand how fashion ecosystem evolves and capturing dynamic trends as they happen.

In order to collect a dataset of Twitter accounts that may be relevant to fashion, we use a content-based, snowball-sampling approach. We started with a seed of 12 highly-fashion related fashion accounts and traversed through other accounts that they follow. By analyzing the tweets of these accounts and the amount of fashion-related words in their tweets as well as their profile description, we are able to collect a list of candidate fashion accounts.

We crowdsourced the task of evaluating whether an account is fashion-related on Amazon Mechanical Turk to create a set of ground-truth labels. Our dataset narrows the search space to observe and control and pick up trends earlier on as well as have a higher granularity on micro trends and fads. We demonstrate that on average our classifier has a 72% recall and 75% precision. Our analysis shows that many Twitter fashion accounts are multifaceted, employ a heavy use of media attachments, and are fashion-related by occupation but not by tweet content.

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 due to the rate-limiting nature of the Twitter API. We started with a seed of 12 highly-fashion related fashion accounts and traversed through other accounts that they follow. By analyzing the tweets of these accounts and the amount of fashion-related words in their tweets as well as their profile description, we are able to collect a list of candidate fashion accounts.

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. In order to capture whether an account is fashion-related or not, we use two

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.
We conduct a crowdsourced data collection on Amazon Mechanical Turk where we ask workers to classify whether a given account is fashion-related or not. Our dataset consists 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 the user graph through the snowballing approach. The Twitter API allows us to search for the 15 most relevant users when provided a topic. 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 and a list of 75 non-fashion users gathered from curated list of top Twitter accounts for various non-fashion topics, such as sports, technology, science, and politics. 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 shown in Figure 1.

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.

![Figure 1: Example non-discriminative words, stopwords, and filtered fashion vocabulary.](image)

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.

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

---

1. https://www.umetrics.com/blog/top-fashion-twitter-accounts
3. goo.gl/aV7WW, goo.gl/LZeFkv, goo.gl/3dor03
5. goo.gl/4DlLav
6. goo.gl/5QrEa
7. goo.gl/eNXDME

---

The Twitter API limits the capability of its backwards history search to a maximum of 3,200 most recent Tweets for each user.
a SVM with an RBF kernel (with a coefficient $\gamma$) 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 $\gamma$. Our results is based on a setting of $C = 4$ and $\gamma = 0.05$.

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.

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image-heavy</td>
<td>@LaurieTrott, @jacvanek</td>
</tr>
<tr>
<td>Link-heavy</td>
<td>@annadellorusslo, @wanderlustandco</td>
</tr>
<tr>
<td>Off-topic fashion workers</td>
<td>@adelefashionmag, @JessC_M</td>
</tr>
<tr>
<td>Fashion+X</td>
<td>@ALTOnmagazine, @ProjectMMNYC</td>
</tr>
</tbody>
</table>

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

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.

**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 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
4 DISCUSSION AND FUTURE WORK

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 personnel 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 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], [17] 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, 16]. 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 Naïve 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