The official version is published in the Journal of the Association for Information Science and Technology (JASIST)
http://dx.doi.org/10.1002/asi.23980

Improving Interpretations of Topic Modeling in Microblogs

Sarah A. Alkhodair
CIISE, Concordia University, Montreal, Quebec, Canada, H3G 1M8.
E-mail: sa_alkho@ciise.concordia.ca

Benjamin C. M. Fung
School of Information Studies, McGill University, Montreal, Quebec, Canada, H3A 1X1. E-mail: ben.fung@mcgill.ca

Osmud Rahman
School of Fashion, Ryerson University, Toronto, Ontario, Canada, M5B 2K3. E-mail: orahman@ryerson.ca

Patrick C. K. Hung
Faculty of Business and Information Technology, University of Ontario Institute of Technology, Oshawa, Ontario, Canada, L1H 7K4. E-mail: patrick.hung@uoit.ca

Abstract

Topic models were proposed to detect the underlying semantic structure of large collections of text documents in order to facilitate the process of browsing and accessing documents with similar ideas and topics. Applying topic models to short text documents to extract meaningful topics is challenging. The problem becomes even more complicated when dealing with short and noisy micro-posts in Twitter that are about one general topic. In such a case, the goal of applying topic models is to extract subtopics. This results in topics represented by similar sets of keywords, which in turn makes the process of topic interpretation more confusing. In this paper we propose a new method that incorporates Twitter-LDA, WordNet, and hashtags to enhance the keyword labels that represent each topic. We emphasize the importance of different keywords to different topics based on the semantic relationships and the co-occurrences of keywords in hashtags. We also propose a method to find the best number of topics to represent the text document collection. Experiments on two real-life Twitter data sets on fashion suggest that our method performs better than the original Twitter-LDA in terms of perplexity, topic coherence and the quality of keywords for topic labeling.

Introduction

Statistics from Twitter show that around 500 million tweets were tweeted per day in February 2016.[1] The huge volume of text in microblogs contains valuable real-time information from different regions of the world. Having an effective method to automatically extract knowledge from such a volume of textual data would provide tremendous advantages to trend and topic analysis in marketing. *Latent Dirichlet Allocation*(*LDA*) (Blei et al., 2003) is a widely adopted topic modeling method that can automatically generate a set of topics from a large collection of textual data. In this paper, we study the shortcomings of LDA and the challenges of applying LDA to microblogs for topic analysis. Furthermore, we present a customized version of *Twitter-LDA* (Zhao et al., 2011) that can better represent the generated topics for microblogs by incorporating a lexical database, domain-specific keywords, and hashtags into the generative model. The proposed method is specifically designed for the domains that satisfy the following four properties: (1) huge volume of textual data, (2) each individual piece of text is very short with overlapping vocabularies, (3) concepts and terminologies are domain-specific, and (4) terminologies change rapidly by the community over time. To illustrate the effectiveness of the proposed method, we present objective quantitative results, together with users' evaluations, on real-life fashion tweets. To further ensure the generated results are meaningful and useful to fashion practitioners, we closely collaborate with a domain expert in fashion communication. We choose fashion communication as the domain of case study because it satisfies the aforementioned properties. Our method can be generalized to other domains that share similar properties such as video games, photography and social media applications.

The Challenges

The problem of handling large collections of text documents and the effectiveness in extracting useful information from the available data has drawn the attention of many researchers. The absence of semantic structures in such collections makes the process of browsing and accessing text documents with similar ideas, i.e., topics, very difficult. With such large collections, a

[1] http://www.internetlivestats.com/twitter-statistics/

simple search query may result in millions of text documents that overwhelm the user with textual data. Topic Models were proposed to solve this problem by automatically detecting the underlying semantic structure of large text document collections and providing short descriptions of text documents. Uncovering this structure facilitates browsing and exploring the collection and allows the user to effectively access documents with similar topics.

LDA (Blei et al., 2003) is one of the most well-known topic models in the literature and serves as the foundation of many other models. LDA assumes a fixed number of topics for the entire corpus. Each topic in LDA is defined as a distribution over a vocabulary of terms, and each document is modeled as a mixture distribution of underlying topics. The difficulty of applying LDA on short text documents to generate meaningful results raised the need of proposing new topic models to handle them. Twitter-LDA (Zhao et al., 2011) is a topic model that was proposed to handle the micro-posts, known as tweets, available in Twitter. This topic model takes into consideration the observation that a single tweet has a single author and usually covers a single topic.

Dealing with short text documents like tweets is challenging. In addition to the lack of co-occurrence patterns and high sparseness of the short text documents, tweets are very noisy. They are often written using informal English with a lot of slang, domain-specific vocabularies, acronyms, and grammatical errors. They also contain URLs, emoticons, mentions, and hashtags. Even though Twitter has lifted the limitation that a tweet can contain only up to a maximum of 140 characters, cleaning the text of tweets leads to very few words in each tweet, which further complicates the process of extracting meaningful topics. The problem becomes even more challenging when the corpus actually covers one major topic, for example, fashion. In such case, we use topic models to detect subtopics. A major challenge here is that the same fashion-related terms are used across tweets covering different subtopics, leading to even fewer co-occurrence patterns of the distinctive terms that distinguish one subtopic from another. Thus, topics detected by topic models in this case are very similar, making it a difficult task to recognize which topic is represented by a given set of keywords.

Example 1. Table 1 shows two different sets of keywords representing two different topics detected by a topic model. Keywords shown in the table cannot be easily used to recognize what topic is represented by each set of keywords. ■

We employ WordNet to address this challenge and improve the set of

A	hair, fashion, photo, menstyle, kingjames, tbt, home, interiordesign, designer, women
B	fashion, style, mensfashion, wear, ootd, onlineshopping, stylish, fashionable, love, menstyle

Table 1: Two sets of keywords representing two different topics

A	prada, philiptreacy, hat, saint, laurent, collect, pradacelebs, campaign, spring, women
B	louisvuitton, assist, team, campaign, givenchy, service, love, tiffani, celebratingmonogram, collect

Table 2: Two sets of keywords representing the topic "Brands"

keywords that represent each topic. WordNet is an English lexical database where words are grouped, based on their meanings, into unordered sets of synonyms. The most distinctive terms to a topic tend to be the most probable terms in its distribution over the vocabulary of terms. However, in the case of detecting subtopics of one general topic, some general terms might have higher probabilities than the most distinctive ones. Using the semantic relations in WordNet, we aim at emphasizing the importance of such distinctive terms. We also aim at taking advantage of the set of hashtags that exists in the corpus. Hashtags in Twitter are strong indicators of the topic covered by a tweet. Emphasizing the importance of terms similar to such strong indicators also improves the topic representation and makes it more focused rather than being about a general topic.

Another known challenge in topic modeling is how to determine in advance the number of topics to be detected. Traditional topic models assume a fixed number of topics that should be specified by the user in advance. Providing a larger number of topics might result in different sets of keywords that represent the same topic. Having such results reduces the effectiveness of the topic model in terms of providing meaningful results to the user.

Example 2. Table 2 shows two sets of keywords detected by a topic model. By glancing through these sets of keywords, one can notice that all topics are mainly about "Brands". Providing the user with two sets of keywords about the same topic makes the interpretability process more confusing. ■

To address this challenge, we propose to employ clustering algorithms to merge similar sets of keywords into a single topic and thus adjust the number of topics to be presented to the user. By doing this we provide the user with an estimation of the number of topics to be extracted from the text document collection, and at the same time the user still has the flexibility of choosing the number of topics that best serves his/her needs for obtaining different topic granularities.

Several works proposed to employ WordNet in topic models as a pre-processing step (Lu, 2013) or a post-processing step (Musat et al., 2011) to handle average-length text documents. Our model, on the other hand, focuses on very short text documents in Twitter and employs the semantic relations between terms in WordNet as an intermediate step in the inference scheme of the topic model.

Contributions

To the best of our knowledge, this is one of the first works that combines WordNet, hashtags, and topic models with the goal of improving the sets of keywords used to represent each topic extracted from short texts in Twitter. The main contributions of this paper are:

- *Improved topic representation.* We used an English lexical database, WordNet, along with the set of hashtags that exists in the corpus, to improve the set of keywords representing every topic by emphasizing the importance of distinctive terms in the distribution of every topic over the vocabulary of terms. Experimental results suggest that our method provides the user with better sets of keywords to represent each topic than does Twitter-LDA, and it improves the user's interpretation of the detected topics.

- *Customized taxonomy for a specific domain.* Our proposed approach can be used to dynamically build a customized taxonomy for a specific domain. To illustrate this capability, we chose *fashion* as the domain in this study. Specifically, we use the maximal frequent itemsets extracted from the corpus to dynamically build a customized version of WordNet that contains fashion-related terms. The customized Wordnet can be used in different text mining tasks.

- *Adjustment of the number of detected topics.* We propose a mechanism to automatically adjust the number of topics to be presented to the user by merging topics represented by similar sets of keywords into a single topic. Experimental results suggest that the coherence of the merged topics is better than the coherence of the original topics.

- *Exploring changes of fashion topics over time.* To illustrate the capability of the proposed method, we evaluate the method by exploring the fashion topics and showing how they were covered over time and how specific users covered these topics by their tweets. Finally, we evaluate our method on two data sets collected from Twitter. The results obtained from the experiments show that our method is better than the original Twitter-LDA in terms of perplexity and topic coherence and provides better results in terms of the quality of the detected topics.

The rest of the paper is organized as follows: First, we discuss some related works, followed by the problem description. Then we briefly provide some background information before describing our proposed method. Finally, we cover the experiments and results, followed by the conclusion.

Related Work

Topic Models

Modeling text documents has attracted a lot of attention in the past years. One of the most well-known topic modeling algorithms is the *Latent Dirichlet Allocation* (*LDA*) (Blei et al., 2003). LDA models each document as a probability distribution over topics and every topic as a probability distribution over a fixed vocabulary of terms. Researchers have also proposed to include additional information sources such as *Author Topic Model* (Rosen-Zvi et al., 2010), and *Topic Over Time* model (Wang & McCallum, 2006). While these topic models aim at handling average-length text documents, our proposed method focuses on short text in microblogs.

Many recent works focused on dealing with short text documents. Based on the observation that a single tweet usually covers only one topic, Zhao et al. (2011) proposed *Twitter-LDA* to model topics in the short messages of Twitter. Sasaki et al. (2014) extended Twitter-LDA by enabling the ratio between topical and background words to be different for every user. In this paper, we introduce a customized version of Twitter-LDA by incorporating a lexical database, domain-specific keywords, and hashtags into the generative model to better represent the generated topics for microblogs.

Another category of topic models is the nonparametric models, based on the *Hierarchal Dirichlet Processes (HDP)* (Teh et al., 2004), where the user does not need to specify the number of topics in advance. Researchers also proposed to include additional information such as authorship (Dai & Storkey, 2009), time (Dubey et al., 2013), and word embeddings (Batmanghelich et al., 2016). Our proposed method is a parametric topic model that provides a mechanism to automatically adjust the number of topics to be presented to the user and, at the same time, provides the user with the flexibility of choosing the number of topics that best serves his/her needs.

Non-Textual Data in Topic Models

Several works in topic models utilize non-textual data such as prices (Iwata & Sawada, 2013), Authors' demographic information (Z. Yang et al., 2015), and geographical locations (Kotov et al., 2013, 2015). Other works also proposed the inclusion of hashtags in topic models. She & Chen (2014) , for example, used topic models for hashtag recommendation. Ma et al. (2013) focused on understanding hashtags and their semantic correlations. Other works employed WordNet in topic models for different purposes. Lu (2013), for example, used WordNet for concepts construction as a preprocessing step and then treat these concepts as observed data. Musat et al. (2011) first applied LDA and then employed WordNet as a post-processing step to build a conceptual ontology. Our proposed method also utilizes non-textual data by incorporating WordNet and hashtags into the generative model of Twitter-LDA to improve the extracted topics.

Problem Description

Given a corpus of tweets, our goal is to improve the most representative keywords of the set of topics covering the corpus by utilizing WordNet and the set of hashtags found in the corpus. We also aim at merging topics represented by similar sets of keywords and building a customized version of WordNet.

Definition 1 (Vocabulary)**.** A vocabulary is a set of distinct terms that are used to construct the text documents denoted by $V = \{v_1, \dots, v_{|V|}\}$. A term v_i is an item from a vocabulary V. ■

Definition 2 (Tweet)**.** A tweet is a textual message that consists of a set of words denoted by $d = \{w_1, \dots, w_{|d|}\}$. ■

Definition 3 (Corpus)**.** A corpus is a collection of tweets $D = \{d_1, \dots, d_{|D|}\}$ that is written using V. ■

Definition 4 (Hashtag)**.** A hashtag is a textual word or phrase that has the symbol # as a prefix. Let H be the set of hashtags in the corpus D. A tweet d_j can link to a set of hashtags $H_{d_j} \subseteq H$. ■

Definition 5 (Author)**.** Let $A = \{a_1, \ldots, a_{|A|}\}$ be the set of authors of D. An author a_b contributed to one or more documents in D. ■

Definition 6 (Topics in tweets)**.** Let $K = \{k_1, \ldots, k_{|K|}\}$ be the set of topics covered by D. A topic k_i is modeled as a probability distribution ϕ^{k_i} over V. Each tweet d_j has a single topic k_{d_j}. ■

Definition 7 (Topic representation)**.** A topic k_i is represented using the top s probable terms S^{k_i} in its probability distribution ϕ^{k_i} over V. ■

WordNet is an English lexical database where words are grouped based on their meanings into unordered sets of synonyms called *synsets*. Each group of synsets denotes a distinct concept and is linked to other groups of synsets via conceptual relations such as hypernyms, hyponyms, and entailment.

Formally, given a corpus of tweets D, we want to model the set of topics K covered by D each as a probability distribution ϕ^{k_i} over V and the interests of each author $a_b \in A$ each as a probability distribution θ^{a_b} over K. We also want to tag every tweet d_j with one of the topics in K, enhance the top s probable terms S^{k_i} representing each topic k_i in K, build a customized version of WordNet to contain fashion-related terms, and merge similar sets of keywords to adjust the number of topics detected from D.

Background Information

In this section, we provide a brief description of Twitter-LDA and posterior inference using Gibbs sampling.

Twitter-LDA

To illustrate how Twitter-LDA (Zhao et al., 2011) works, let K be the set of topics, ϕ^k and ϕ^{BG} be the word distributions for topical and background words, respectively. Let θ^{a_b} be the topic distribution of author a_b and π be the Bernoulli distribution, which determines the choice between topical or background words. Then, Twitter-LDA's generative process is described by Algorithm 1 and its plate notation is shown in Figure 1.

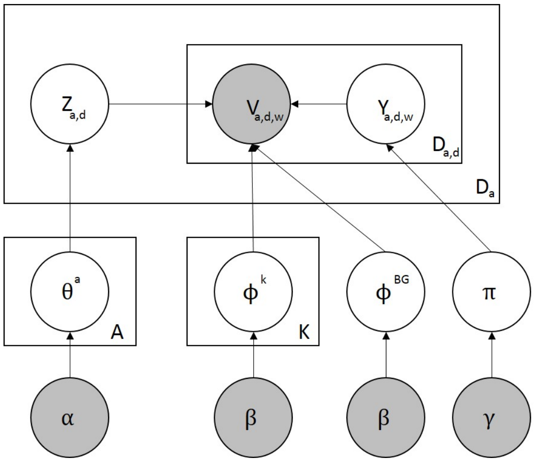

Figure 1: Plate notation of Twitter-LDA

Algorithm 1 The generative process of Twitter-LDA

Input A corpus of tweets D and the set of authors A

Output Distributions of authors over topics Θ and topics over vocabulary Φ, and topics of tweets Z.

```
Draw $\phi^{BG} \sim Dir(\beta), \pi \sim Dir(\gamma)$.
for each topic $k_i$ indexed by $i = 1$ to $|K|$ do
    Draw $\phi^{k_i} \sim Dir(\beta)$.
end for
for each author $a_b$ indexed by $b = 1$ to $|A|$ do
    Draw $\theta^{a_b} \sim Dir(\alpha)$.
    for each tweet $d_j$ indexed by $j = 1$ to $|D_{a_b}|$ do
        Draw $z_{a_b,d_j} \sim Mult(\theta^{a_b})$.
        for each word $w_n$ indexed by $n = 1$ to $|d_j|$ do
            Draw $y_{a_b,d_j,w_n} \sim Mult(\pi)$.
            if $y_{a_b,d_j,w_n} = 0$ then
                Draw $v_{a_b,d_j,w_n} \sim Mult\left(\phi^{BG}\right)$
            else
                Draw $v_{a_b,d_j,w_n} \sim Mult\left(\phi^{z_{a_b,d_j}}\right)$
            end if
        end for
    end for
end for
```

Inference Using Gibbs Sampling

The basic idea of topic modeling is to posit a hidden latent topical structure on the observed data and then use the posterior probabilistic inference to learn this structure. Since it is difficult to obtain the exact value of the posterior distribution, several approximation algorithms were employed such as *variational inference* (Blei et al., 2003) and *Gibbs sampling* (Rosen-Zvi et al., 2010).

Gibbs sampling (Gilks et al., 1996) is a form of *Markov Chain Monte Carlo* that is widely used by topic models to estimate the value of the posterior distribution on random variables. A two-step inference scheme (Rosen-Zvi et al., 2010) is employed. The process starts by running a Gibbs sampler to estimate the value of $P(z, x|D, \alpha, \beta)$, where z and x represent the author and topic assignment of words in D, respectively. And α and β are the hyperparameters of the topic model. Next, the value of the posterior distribution on the random variables Θ and Φ are calculated using the following formulas:

$$\phi^{v_j k_i} = \frac{W(v_j, k_i) + \beta}{\sum_{v'_j} W(v'_j, k_i) + V\beta} \tag{1}$$

$$\theta^{k_i a_b} = \frac{T(k_i, a_b) + \alpha}{\sum_{k'_i} T(k'_i, a_b) + K\alpha} \tag{2}$$

where W is the count matrix that holds the counts for every term-topic pair, and $W(v_j, k_i)$ represents how many times the term v_j was used in topic k_i. Similarly, T is the count matrix that holds the counts for every topic-author pair, and $T(k_i, a_b)$ represents how many terms author a_b used to write about topic k_i.

Methodology

Figure 2 depicts an overview of the core modules of the proposed method. The first module applies some standard text preprocessing steps. The following three modules represent the inference process. First, the process starts by running a Gibbs sampler. Second, WordNet and the set of hashtags are used to adjust the importance of different terms to different topics. Third, the posterior distribution on the random variables is calculated. Finally, the topic clustering module groups similar topics together in order to provide a coherent representation of the topics to the user.

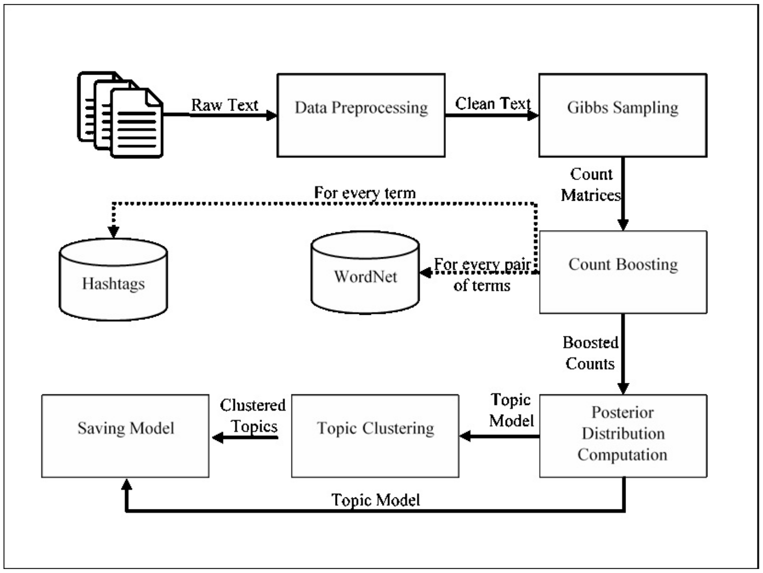

Figure 2: An overview of the proposed model

Gibbs Sampling

This module represents the first step in the inference scheme. A Gibbs sampler is used to estimate the topics and authors assignments, z and x, and record their counts in two count matrices: W and T. The first one contains the counts of every term-topic pair, while the other contains the counts of every topic-author pair. The algorithm of Gibbs sampling has two steps. First, it randomly initializes the topic assignments z and the author assignments x for each word w_i. Second, during each Gibbs sampling iteration, it samples the author assignment x_i and topic assignment z_i for each individual word w_i conditioned on fixed authors and topics assignments for all other words in the corpus. After completing a predefined number of iterations, the assignments x and z and the counts W and T are recorded to be used in the calculation of the posterior distribution on Θ and Φ. We will focus on the topics distributions over terms Φ in the rest of the paper.

Count Boosting

This is the second step in the inference scheme. It takes the term-topic matrix W and uses WordNet and the set of hashtags H to update the counts of different terms according to their importance to different topics.

Our intuition is that among the most probable terms for a topic, those who are semantically similar, are the most distinctive ones to that topic. Therefore, we boost their counts based on their importance to the topic and their semantic similarities. Basically, we take the top l probable terms L^{k_i} for every topic k_i. Then for every pair of terms in L^{k_i}, we boost their counts based on their similarity in WordNet. More specifically, we use WordNet to retrieve the shortest path *dist* between the two terms. For example, let k_i be the current topic of interest. Then for every pair of terms $(v_x, v_j) \in L^{k_i}$, we use WordNet to retrieve the distance between them $dist(v_x, v_j)$ based on their lexical category. We only consider nouns and verbs in our work. Then the counts of both terms are boosted as follows:

$$WN(v_x, k_i) = W(v_x, k_i) + \frac{W(v_j, k_i)}{dist(v_x, v_j)} \tag{3}$$

$$WN(v_j, k_i) = W(v_j, k_i) + \frac{W(v_x, k_i)}{dist(v_x, v_j)} \tag{4}$$

where WN represents the updated term-topic count matrix based on the relationships between terms in WordNet.

We further boost the counts of terms in L^{k_i} for every topic k_i by taking advantage of the set of hashtags H in D. Since hashtags are strong indicators of topics, our intuition is that among the most probable terms of a topic, those who appear in the topics hashtags, are the most representative ones of that topic. Therefore, we boost their counts based on how often these hashtags are used to tag that topic. Let k_i be the current topic of interest. Then for every term $v_j \in L^{k_i}$, we check if it appears in at least one of the hashtags H_{k_i} associated with k_i. Let H_{v_j} be the set of hashtags that contain the term v_j. The count of v_j is boosted as follows:

$$WH(v_j, k_i) = WN(v_j, k_i) + \left[WN(v_j, k_i) * \left(\frac{\sum_{h \in H_{v_j}} hashFreq[k_i][h]}{TotalHashFreq[k_i]} * 100 \right) \right] \tag{5}$$

where WH represents the updated term-topic count matrix based on the set of hashtags, $hashFreq[k_i][h]$ denotes the frequency of hashtag h in tweets about topic k_i, and $TotalHashFreq[k_i]$ denotes the sum of the frequencies of all hashtags in H_{k_i}.

Furthermore, this module builds a customized version of WordNet to include domain-related terms. Adding a new term to customize WordNet for a specific domain should comply with the following criteria: For a term v_j to be added and connected to a set of terms V_j in WordNet, the term

v_j should be related to all terms in V_j in the context of that domain. We use the *maximal frequent itemsets MFI* found in the corpus as a guide to determine where to add these terms and how to connect them to other terms in WordNet.

Definition 8 (Maximal Frequent Itemset (MFI) (Burdick et al., 2001))**.** *Let the vocabulary V be the set of all distinct terms. Let $I \subseteq V$ be an itemset. Let the collection of tweets D be a multiset of subsets of the vocabulary V. The support of an itemset, support(I), is the percentage of tweets in D containing I. An itemset I is a* frequent itemset *if support$(I) \geq$ minSup, where minSup is a user-defined minimum support. A frequent itemset I is a* maximal frequent itemset *(*MFI*) if there is no superset of I that is frequent.*

For the customization purposes, we assume that all terms to be added to WordNet are nouns and all relationships are of type SIMILAR-TO. We then use MAFIA (Burdick et al., 2001) to mine the MFI from the corpus D. Next, for every term v_j in the top probable terms for topic k_i, $v_j \in L^{k_i}$; if it does not exist in WordNet, we find the maximal frequent itemset MFI^{v_j} that contains v_j. If more than one is found, we use the one with the maximum support. Then, if at least one term in MFI^{v_j} exists in WordNet, we add v_j to WordNet. Next, for every item (term) $v_x \in MFI^{v_j}$, we check if it exists in WordNet. If this is the case, we customize WordNet by adding a SIMILAR-TO relationship between v_x and v_j.

Posterior Distribution Calculation

This is the final step in the inference where we actually compute the posterior distribution on the random variables. After boosting the counts, the computation of the posterior distribution on the random variables Φ and Θ is a straightforward step. Given the updated count matrix WH, the distributions of topics over the vocabulary of terms $\phi_{WH}^{v_j k_i}$ is calculated directly from Equation 1 as follows:

$$\phi_{WH}^{v_j k_i} = \frac{WH(v_j, k_i) + \beta}{\sum_{v'_j} WH(v'_j, k_i) + V\beta} \tag{6}$$

Similarly, the author's distributions over topics Θ is calculated from Equation 2 directly.

Improved Topic Clustering

Most parametric topic models assume a fixed number of topics, which is unknown in advance in most cases. We propose a new method that uses the agglomerative hierarchical clustering algorithm (Jain & Dubes, 1988) to adjust the number of topics to be presented to the user based on the Kullback-Leibler (KL) divergence. The KL divergence is used to calculate the similarity between the distributions over vocabulary for every pair of topics. Let $D_{KL}(p||q)$ be the KL distance between the distributions of two topics p and q. Since KL divergence is not symmetric, we calculate the distance $Dst(q,p)$ as follows:

$$Dst(q,p) = \frac{(D_{KL}(p||q) + D_{KL}(q||p))}{2} \tag{7}$$

At every step, the clustering algorithm calculates the KL divergence between every pair of topics and merges the pair with the lowest KL divergence value. To determine the best number of topics to be returned to the user we employed the L method (Salvador & Chan, 2004). The L method builds a two-dimensional evaluation graph where the x-axis represents the number of topics and the y-axis represents the KL-divergence. It then calculates and returns the "knee" of the evaluation graph, which represents the best number of topics that represent the corpus.

Experiments

We evaluated our method in terms of Perplexity, Topics' coherence and their quality. We also analyzed the topics trends and interests of the users over time and show examples of customizing WordNet. In the experiments, we set the number of Gibbs sampling iterations of each topic model at 500 and fixed the hyperparameters at $\alpha = 50/K$ and $\beta = 0.01$.

Data Sets

For evaluation purposes we collected two fashion data sets using Twitter API, namely, **OffAcc** and **FashionKW**. **OffAcc** has 100,099 tweets collected over 20 months, from September 2013 to May 2015. The tweets were retrieved from the official accounts of 51 fashion designers and magazines. **FashionKW** has 122,579 tweets collected over a period of 13 days from March 4, 2015 to March 16, 2015. The tweets were collected by sending search queries that contained 110 fashion related hashtags (keywords) to Twitter API. The resulting corpus was written by 48,643 different users. For repeatability, the tweet IDs are available on http://dmas.lab.mcgill.ca/fung/research/data/AFRH16tweetIDs.txt.

Data Set	Tweets	Authors	Vocabulary
OffAcc data set	83,404	51	29,155
FashionKW data set	38,038	943	35,016

Table 3: Data sets statistics

Data Preprocessing

We cleaned the tweets by removing URLs, emoticons, punctuation marks, mentions, stop words and words that appear in more than 70% of the tweets. The remaining words were then stemmed using a Porter Stemmer (Porter, 1997). The corpus was further processed so that duplicates and tweets with fewer than 3 words were removed. Furthermore, users with fewer than 10 tweets were removed, along with their tweets. Table 3 shows the statistics of the two data sets after the preprocessing.

Perplexity

To compare the predictive performance of our method with Twitter-LDA, we performed a 10-fold cross validation and calculated the perplexity on hold-out testing data for $K = 3, 6, 9, 12, 15$. The perplexity is a widely used measurement to evaluate the ability of a probabilistic topic model to handle unseen documents. Lower perplexity implies better predictive performance of the model. It is defined as a decreasing function of the log-likelihood $\ell(w)$ of unseen documents w, as follows:

$$Perplexity(w) = exp\left[\frac{-\ell(w)}{N}\right] \tag{8}$$

where N denotes the total number of words in the corpus.

We compared the results obtained from applying *Twitter-LDA* and our proposed method *Twitter-LDA with WordNet and Hashtags* (*Twitter-LDA-WNH*) on the OffAcc data set. To evaluate the effect of including the set of hashtags, we also included our method with WordNet only (*Twitter-LDA-WN*) in the comparison. Figure 3 shows that Twitter-LDA-WNH has the lowest perplexity for all values of K, followed by Twitter-LDA-WN, while Twitter-LDA has the higher perplexity values. The results also show that the perplexity values for all three models increased when $k = 15$. Since we have a low number of topics in fashion, usually between 5 to 12, providing the model with a larger K typically results in a complicated model with

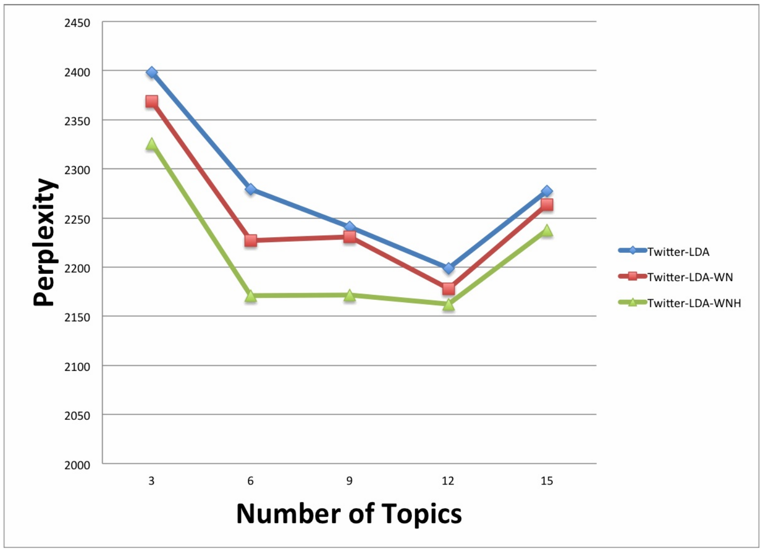

Figure 3: Perplexity on OffAcc data set

many vague topics that are difficult to interpret, and it reduces the model's performance when handling new documents. Similarly, when $K = 3$, we got a complicated model with very general topics, which in turn reduces the ability of the model to handle new documents.

The obtained results suggest that our method, Twitter-LDA-WNH, has the best predictive performance compared to Twitter-LDA and Twitter-LDA-WN for all values of K. The results also reveal that the inclusion of the set of hashtags in our method can further improve the results in terms of handling unseen documents.

Topics Coherence

We further evaluated the quality of our results in terms of topics coherence. Our goal is to show how incorporating WordNet and the set of hashtags in our method helps increase the coherence of the learned topics over topics learned by Twitter-LDA. The employment of the semantic relations in WordNet and the hashtags as an intermediate step helps emphasize the importance of distinctive terms during the inference process itself. Consequently, this helps minimize the effect of the overlapping vocabulary in a specific domain and distinguish the learned topics from each other.

We used the *Normalized Pointwise Mutual Information (NPMI)* (Aletras & Stevenson, 2013) and the *CP*(Röder et al., 2015) in our experiment to measure the coherence of the topics learned by *Twitter-LDA*, our method, *Twitter-LDA-WNH*, and the results of our method after clustering the topics, *Clustered-TLDA-WNH.* In their study, Röder et al. (2015) have evaluated several coherence measures in terms of their correlation to human ratings. Their study shows that NPMI has the strongest correlation to human ratings among all already existing coherence measures while CP outperforms all coherence measures that use direct confirmation, including NPMI. This justifies the reason for using CP and NPMI for our experiment.

CP uses a one-preceding segmentation of the top keywords to calculate the coherence of a topic. For every keyword, the confirmation to its preceding keyword is calculated using Fitelson's coherence (Fitelson, 2003) as follows:

$$\varrho\left(w_i, w_j\right) = \left(\frac{p\left(w_j|w_i\right) - p\left(w_j|\neg w_i\right)}{p\left(w_j|w_i\right) + p\left(w_j|\neg w_i\right)} + \frac{p\left(w_i|w_j\right) - p\left(w_i|\neg w_j\right)}{p\left(w_i|w_j\right) + p\left(w_i|\neg w_j\right)}\right)/2 \tag{9}$$

The arithmetic mean of the Fitelson's coherence results is the CP value of that topic.

NPMI uses a one-one segmentation of the top keywords to calculate the coherence of a topic. For every pair of keywords, the confirmation is calculated as follows:

$$NPMI\left(w_i, w_j\right) = \frac{PMI\left(w_i, w_j\right)}{-log\left(p\left(w_i, w_j\right)\right)} \tag{10}$$

The arithmetic mean of the NMPI results is the overall NPMI value of that topic.

In this experiment we represented each topic as a set of the top 10 most probable keywords in its distribution over terms and used *Palmetto* [2] to calculate the NPMI and CP values for each topic.

A two-samples t-test was conducted to compare the NPMI and CP coherence values of topics learned by Twitter-LDA and Twitter-LDA-WNH. The obtained results show that there is a significant difference in the NPMI coherence for topics learned by Twitter-LDA $(M = -.03, SD = .04)$ and by Twitter-LDA-WNH $(M = .02, SD = .06); t(24) = 2.45, p = .01$. The results also show that there is a significant difference in the CP coherence for topics learned by Twitter-LDA $(M = -.09, SD = .20)$ and by Twitter-LDA-WNH $(M = .26, SD = .21); t(28) = 4.76, p < .001$. These results suggest that topics in our model are more coherent. Specifically, our results

[2] https://github.com/AKSW/Palmetto

suggest that when we incorporate WordNet and hashtags into the generative model of Twitter-LDA, the coherence of the topics increases.

To evaluate the results of topics' clustering, we also compared the coherence of topics learned by Twitter-LDA and our model after clustering, Clustered-TLDA-WNH. We started with an initial value of $K = 15$ and $K = 8$ for the OffAcc and FashionKW data sets, respectively. For the OffAcc data set, the clustering algorithm performed 6 merges and reduced the number of topics to 9. Similarly, for the FashionKW data set, the clustering algorithm performed 2 merges and reduced the number of topics to 6.

We conducted a two-samples t-test to compare the NPMI and CP coherence values of topics learned by Twitter-LDA and Clustered-TLDA-WNH. The obtained results show that there is a significant difference in the NPMI coherence for topics learned by Twitter-LDA $(M = -.03, SD = .04)$ and by Clustered-TLDA-WNH $(M = .08, SD = .09); t(13) = 3.90, p < .001$. Similarly, the results show that there is a significant difference in the CP coherence for topics learned by Twitter-LDA $(M = -.09, SD = .20)$ and by Clustered-TLDA-WNH $(M = .27, SD = .08); t(19) = 6.61, p < .001$. These results suggest that merging topics learned by our model yields more coherent topics than the original topics learned by Twitter-LDA. The results also suggest that the best number of topics for the OffAcc and FashionKW data sets are 9 and 6, respectively.

Users' Evaluation

The objective of the users' evaluation is to compare the quality of the results obtained through the application of our method, Twitter-LDA-WNH, and Twitter-LDA in terms of both interpretation and representation of the topic from the perspective of human users. Due to the fact that judging the quality of a topic is subjective, and to avoid our bias interpretation, we conducted an online survey in March 2017 and asked participants to judge the quality of the top probable keywords generated by the two methods.

Evaluating the Interpretation of Topics

The perplexity results suggest that the best number of topics for the OffAcc data set is in the range of $6 - 12$ topics. The topics coherence results also suggest that the best number of topics for the OffAcc data set is 9 topics and 6 topics for the FashionKW data set. Therefore, in this experiment we set $K = 9$ and $K = 6$ for the OffAcc and FashionKW data sets, respectively, and applied Twitter-LDA and our proposed method Twitter-LDA-WNH.

This result in a total of 30 topics; each method yielding 15 topics. We also carefully prepared a set of 16 labels to cover the most popular topics in the fashion industry as follow: First, we systematically went through the top 10 popular fashion magazines and identified the common topics. Then, we reviewed these topics with a fashion expert, the third co-author of this paper, and merged them into 16 topics with minimal overlapping. We then took the top 10 probable keywords in the distribution of every topic and prepared the test so we had 30 sets of keywords generated by the two methods and a set of 16 labels representing the topics. The sets of keywords were mixed together in random order so that the participants did not know which set was generated by which method. We then asked 105 participants to assign a label to each set of keywords. Our participants were undergraduate students from Ryerson University in Canada with academic backgrounds in fashion.

To evaluate the results, we prepared a standard answer that represents the true topics' labels for each set of keywords based on the judgment of a fashion expert who is the third co-author of this paper. We would like to emphasize that we did not ask the fashion expert to evaluate the performance of our method. To avoid any bias, we provided him with 30 sets of keywords mixed in random order and asked him to assign a label to each set of keywords. The expert did not know the method that generated each set. His interpretations for topics was only used as the gold standard for comparing with responses gathered from other participants to the true answer and recorded the percentage of the correct answers for each set of keywords. For the evaluation purposes, even if a participant selects a label that is different from the golden answer, it does not necessarily mean the answer is wrong in practice. However, we consider the chosen label as incorrect. We acknowledge that this evaluation process is harsh. Given such a harsh evaluation setting, we can still show that our proposed method yields good results.

Table 4 shows that the interpretation of the true topic label of the sets of keywords improved for the OffAcc and FashionKW data sets after applying Twitter-LDA-WNH with an average of 14% and 22%, respectively. A two-samples t-test was conducted to compare the average of users' interpretation of topics learned by Twitter-LDA and Twitter-LDA-WNH. The obtained results show that there is a significant difference in the users' interpretation for topics learned by Twitter-LDA ($M = .38, SD = .05$) and by Twitter-LDA-WNH ($M = .56, SD = .01); t(2) = 4.94, p = .02$. These results suggest that topics in our model become more interpretable by users.

We further analyzed the results and noticed that topics represented by a lot of acronyms, fashion brands, and names did not improve by applying our method. Table 5 shows some examples of such topics after applying

Data set	**Twitter-LDA**	**Twitter-LDA-WNH**
OffAcc	41%	55%
FashionKW	34%	56%

Table 4: Average percentage of the correct answers for both models

	Twitter-LDA		Twitter-LDA-WHN	
Topic #	**Keywords**	**Label**	**Keywords**	**Label**
7	dolce, gabbana, amp, dgwomen, versace, dolcegabbana, dgeditorials, wear, discover, fashion	Brands	dolce, gabbana, amp, versace, wear, fashion, dgwomen, dolcegabbana, summer, dgeditorials	Brands
1	kim, kardashian, tylor, video, beyonce, swift, west, jenner, watch, kany	Celebrities	fashion, time, song, love, thing, kendal, dress, video, west, watch	Media

Table 5: Examples of topics in OffAcc and FashionKW data sets

Twitter-LDA-WNH on OffAcc data set. As shown in the table, the set of keywords representing topic 7 did not improve after applying our method, while topic 1 on the other hand fell under the topic of *Media* instead of *Celebrities*. Another finding was that if some fashion terms do not exist in WordNet, it will diminish the ability of our method to improve some topics. Furthermore, since only 31% of tweets in the OffAcc data set contain hashtags, the influence of emphasizing the importance of terms based on the set of hashtags was limited. As a result, the interpretation of some of these topics did not improve, while the interpretation of others changed completely.

Evaluating the Quality of Topics' Representations

We further evaluated the quality of the keywords used to represent each topic in terms of the number of representative keywords of each topic after

Model	Keywords	Label	# Related words
Twitter-LDA	jewelry, fashion, menstyle, hair, photo, ring, kingsjames, tbt, home, interiordesign	Jewelry	3
Twitter-LDA-WNH	ring, jewelry, fashion, diamond, silver, gold, photo, hair, home, vintage	Jewelry	7

Table 6: Improvements of the sets of keywords resulting from Twitter-LDA-WNH over Twitter-LDA

applying Twitter-LDA and our method, Twitter-LDA-WNH.

Table 6 and Table 7 show some examples of different sets of keywords resulting from applying the two methods and how they were interpreted by users. Table 6 shows how both sets of keywords were interpreted to be about the topic *Jewelry.* The set of keywords resulting from applying Twitter-LDA-WNH was better than the one resulting from applying Twitter-LDA in terms of the number of related keywords to that topic. Table 7 shows how the interpretation of the set of keywords has changed. The label *Celebrities* was assigned for the set of keywords resulting from applying Twitter-LDA. This assignment has shifted to the topic *Events* after applying our method. The obtained results show that the average number of improved representative keywords are 5.3 keywords for our method, in contrast to 3.3 for Twitter-LDA. The results also suggest that although the interpretation of some topics has been completely changed, our method provides meaningful topics with a reasonable number of representative keywords.

Topical Trends Over Time

To illustrate how different topics are covered by tweets over time, we counted the number of tweets written in each time slot for every topic. Figure 4 shows how the 15 topics in the OffAcc data set were covered by the collection of

Model	Keywords	Label	# Related words
Twitter-LDA	red, oscars, carpet, dress, kate, gown, kardashian, kim, wed, celebr	Celebrities	9
Twitter-LDA-WNH	oscars, dress, gown, carpet, photo, red, fashion, night, middleton, jenner	Events	9

Table 7: Different Interpretation of two sets of keywords resulting from Twitter-LDA and Twitter-LDA-WNH

tweets over a time span of 20 months. The result shows that the number of tweets about topics, such as *Shopping*, *Brands*, *Seasons & Collections*, and *Men's Wear*, was mostly stable throughout the year. Tweeting about *Customers & Services*, *Trends & Styles*, and *Jobs* increased slightly in the period of June-July 2014 and November-December 2014, which reflects the heavy shopping periods such as the annual sale season, Christmas, and New Year. On the other hand, topics such as *Media*, *Fashion Week*, *Celebrities*, and *Beauty & Appearance* were heavily covered by tweets in February-March 2014, June 2014, November-December 2014, and February-March 2015, reflecting major fashion events such as the *Fashion Week*, *Oscars*, *Golden Globe*, and *Grammy Awards*. Knowing how different topics are covered by Twitter during the year can be of great importance to marketing and advertising.

Users' Interests Over Time

To illustrate how the interests of users in these topics have changed over time, we chose two fashion magazine accounts, namely, *LuckyMagazine* and *StyleForum*, and two fashion designers' accounts, namely, *Prada* and *YSL*.

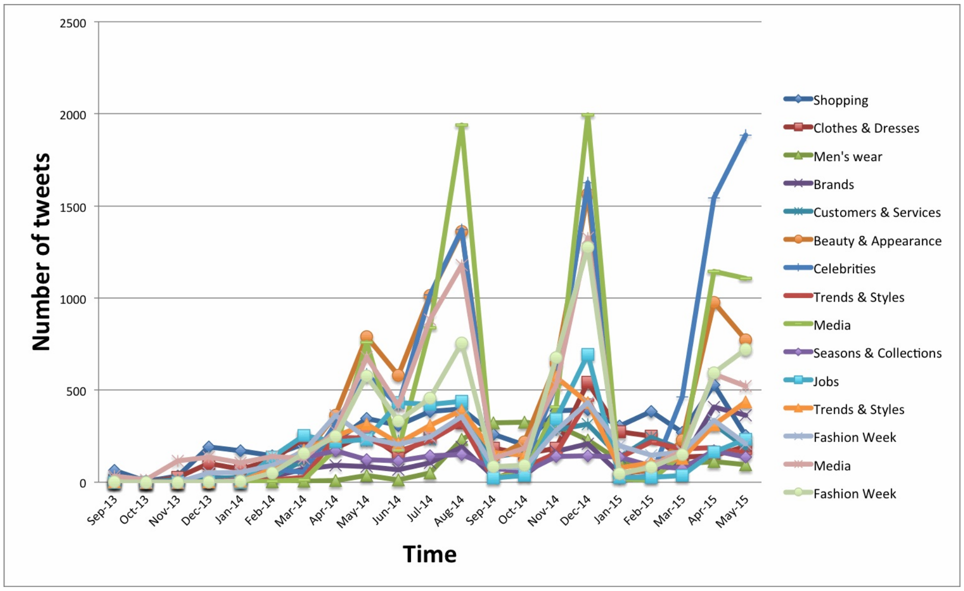

Figure 4: Topical trends over time

Figure 5 shows how the tweets written by StyleForum were mainly about *Trends & Styles*, followed by *Beauty & Appearance*, *Celebrities*, and *Events* such as fashion week. These topics were heavily covered during March-April 2014, June-July 2014, November-December 2014, and March-April 2015, reflecting major fashion events during the year. Figure 6 shows that the LuckyMagazine tweets were mainly about *Beauty & Appearance*, followed by *Celebrities*. Tweeting about such topics noticeably increased during March-April 2014, June-July 2014, November-December 2014, and March-April 2015, which also reflects the major events in fashion. Fashion Designers' interests are shown in Figure 7 and Figure 8. As shown in Figure 7, most of Prada's tweets were about *Customers & Services*. These tweets increased in April 2014, August 2014, February-March 2015, and April 2015. These are usually the times when new seasonal collections are launched by designers. YSL's tweets, on the other hand, were mainly about *Seasons & Collections*, as shown in Figure 8. Similar to Prada, YSL's tweets were heavily about *Seasons & Collections* during April 2014, July 2014, November-December 2014, and January 2015. In general, we noticed that the use of Twitter by fashion designers is somehow limited, while fashion magazines' tweets are more about fashion events, icons, and trends. Knowing the timing and topics of the fashion designers' and magazines' tweets can greatly help marketing

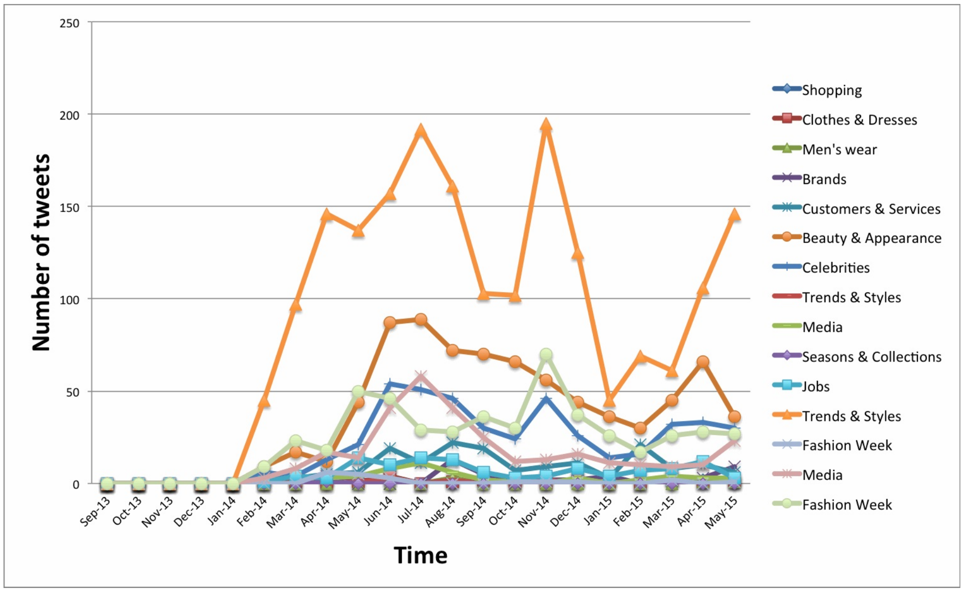

Figure 5: StyleForum's tweets over time

and advertising agencies know when, how, and through which account they can target potential customers.

Customized Taxonomy

In this section we demonstrate the results of WordNet customization to include domain-specific terms. Figure 9 shows some examples of the fashion-related terms that were added to WordNet. Each sub-figure represents one addition. The new term is represented by the node at the top, while the nodes at the bottom represent the terms already existing in WordNet.

In our experiment, WordNet was customized to include fashion acronyms, brands, communities, and other domain-specific terms. Figure 9.a shows how *tbt* [3], a widely used fashion acronym, was connected to *photo, hair*, and *style*. Similarly, figure 9.b shows how *ootd*[4], another acronym, was connected to *style* and *trend*. Figure 9.c and figure 9.d show how the brand,

[3] "Stands for (throwback to) to indicate an old photo, idea, etc." Retrieved December 20, 2016, from http://www.urbandictionary.com/define.php?term=TBT

[4] "Outfit Of the Day." Retrieved December 20, 2016, from http://www.urbandictionary.com/define.php?term=OOTD

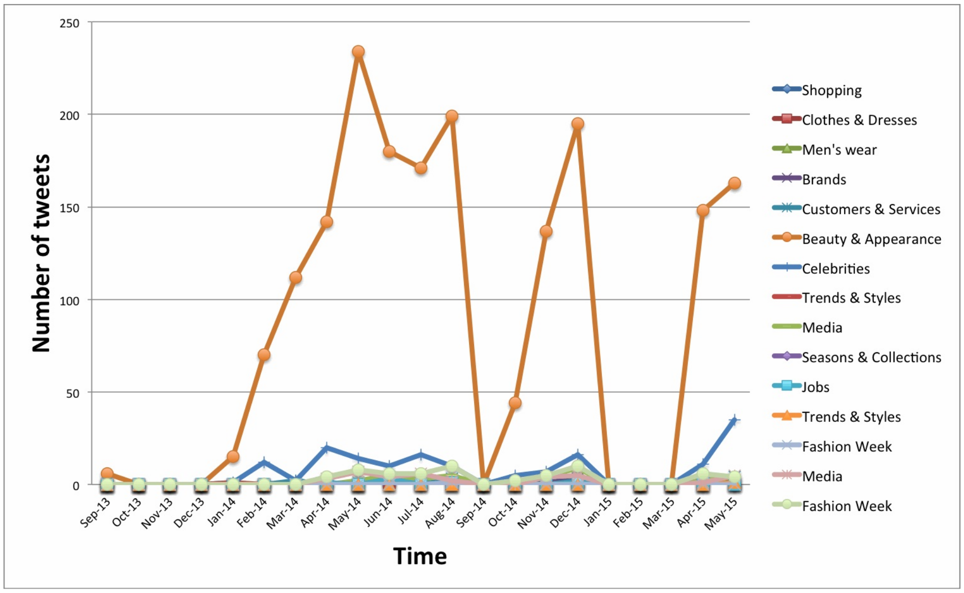

Figure 6: LuckyMagazine's tweets over time

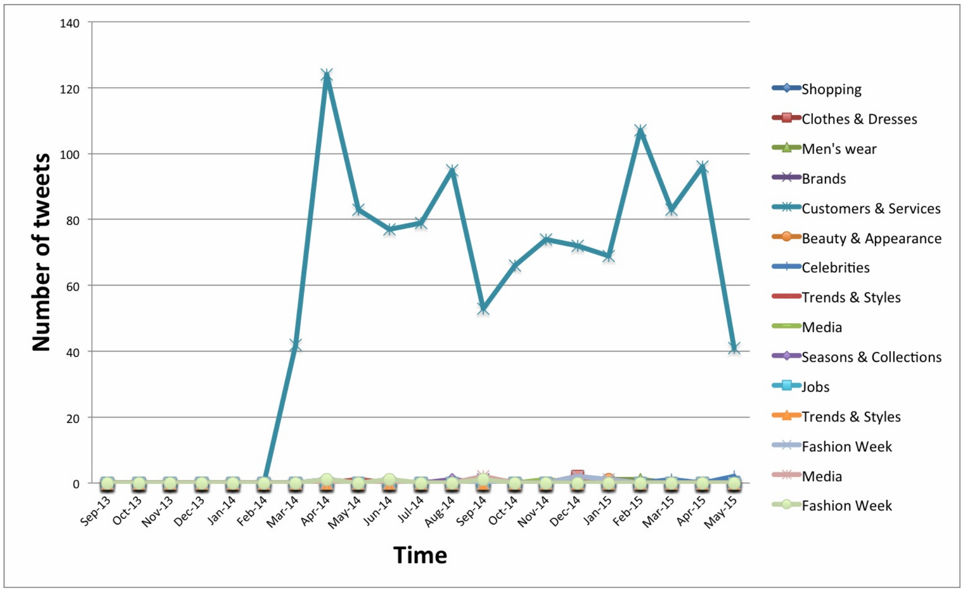

Figure 7: Prada's tweets over time

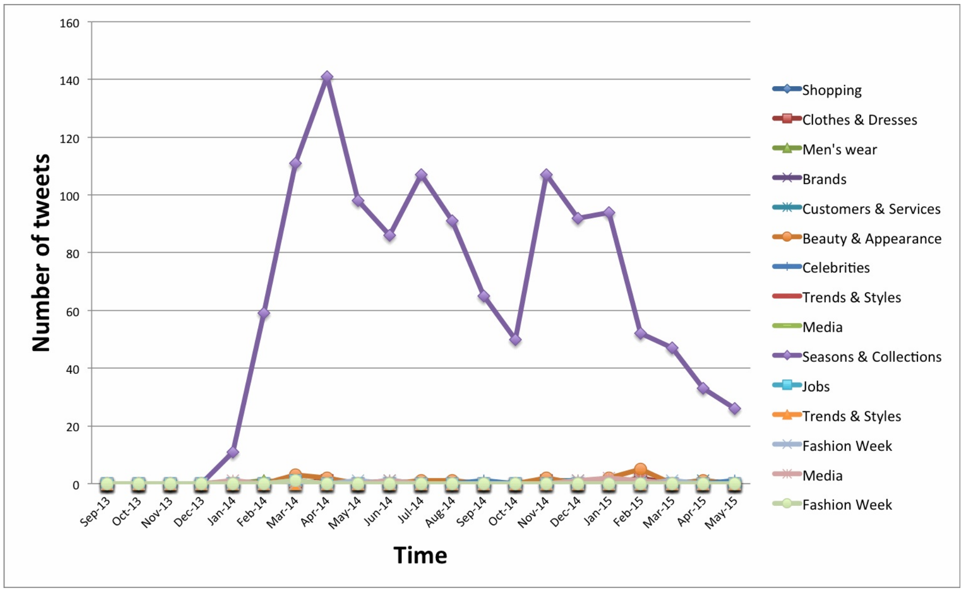

Figure 8: YSL's tweets over time

Gucci,[5] and the fashion community,*Hijabers*,[6] were connected to the term *fashion*. Other domain-specific terms such as *lurex*[7], and *moda*[8] were also added as shown in figures 9.e,and 9.f. These examples show how terms were added to WordNet and connected to related terms in the context of fashion. This can be generalized to dynamically build a domain-specific taxonomy for any domain that shares the same characteristics.

Conclusion

In this paper, we propose a new method that incorporates Twitter-LDA, WordNet, and the set of hashtags available in the corpus with the objective of improving the top probable keywords that represent each topic. Based on the semantic relationships in WordNet and the set of hashtags available in the

[5] "An international fashion company." Retrieved December 20, 2016, from http://www.urbandictionary.com/define.php?term=Gucci

[6] "A fashion community for trends in hijab." Retrieved December 20, 2016, from http://erpub.org/siteadmin/upload/8991ER815006.pdf

[7] "A type of fabric." Retrieved December 20, 2016, from https://en.oxforddictionaries.com/definition/lurex

[8] "Fashion, trend, style." Retrieved December 20, 2016, from https://en.wiktionary.org/wiki/moda

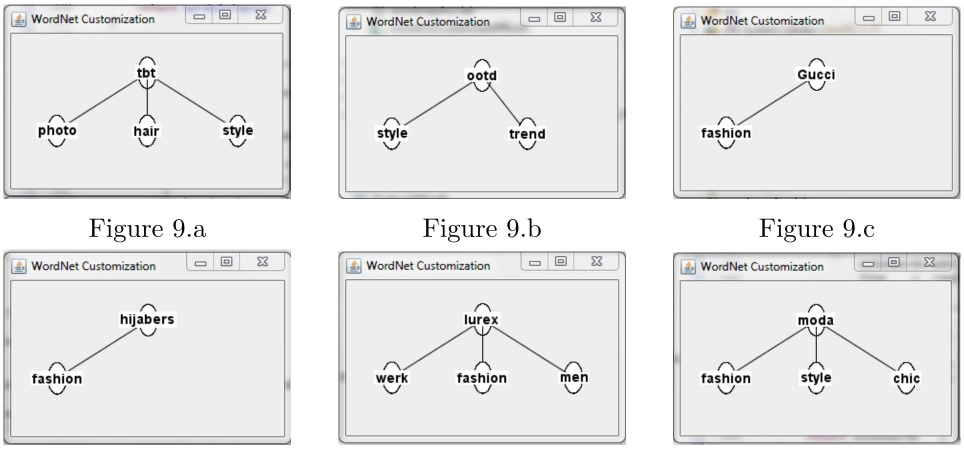

Figure 9.a Figure 9.b Figure 9.c

Figure 9.d Figure 9.e Figure 9.f

Figure 9: Examples of WordNet customization

corpus, the importance of different keywords to different topics is emphasized in the effort of providing the user with a higher quality representation of each topic. A customized version of WordNet is also built to include domain-related terms based on the maximal frequent itemsets found in the corpus. Furthermore, we propose to find the best number of topics covered by the corpus by employing a clustering algorithm to cluster topics based on their similarities in order to get more coherent topics. We further analyze how topics' coverage and users' interests change over time. The proposed method is applied on two real-life fashion data sets collected from Twitter. The obtained results suggest that our method is better than Twitter-LDA in terms of the perplexity, topics' coherence and their quality.

Acknowledgment

This research is supported in part by SSHRC (430-2014-01100), NSERC (356065-2013), and CRC Program (950-230623). The first author is fully sponsored by King Saud University in Riyadh, Saudi Arabia, and the Saudi Arabian Cultural Bureau in Canada.

References

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS'13).*

Batmanghelich, K. N., Saeedi, A., Narasimhan, K., & Gershman, S. (2016). Nonparametric spherical topic modeling with word embeddings. *Computing Research Repository*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, March). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Burdick, D., Calimlim, M., & Gehrke, J. (2001). Mafia: A maximal frequent itemset algorithm for transactional databases. In *Proceedings of the 17th International Conference on Data Engineering (ICDE)* (pp. 443–452).

Dai, A. M., & Storkey, A. J. (2009). Author disambiguation: a nonparametric topic and co-authorship model. In *NIPS Workshop on Applications for Topic Models Text and Beyond* (pp. 1–4).

Dubey, A., Hefny, A., Williamson, S., & Xing, E. P. (2013). A nonparametric mixture model for topic modeling over time. In *Proceedings of the 2013 SIAM International Conference on Data Mining (SDM)* (pp. 530–538).

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, *63*(279), 194–199.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain monte carlo in practice.*

Iwata, T., & Sawada, H. (2013, May). Topic model for analyzing purchase data with price information. *Data Mining and Knowledge Discovery (DMKD)*, *26*(3), 559–573.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data.*

Kotov, A., Rakesh, V., Agichtein, E., & Reddy, C. K. (2015). Advances in information retrieval. In A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.), (chap. Geographical Latent Variable Models for Microblog Retrieval).

Kotov, A., Wang, Y., & Agichtein, E. (2013). Leveraging geographical metadata to improve search over social media. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 151–152).

Lu, H.-M. (2013). Wordnet-enhanced topic models. In *Mining Data Semantics in Heterogeneous Information Networks Workshop (MDS 2013).*

Ma, Z., Dou, W., Wang, X., & Akella, S. (2013). Tag-latent dirichlet allocation: Understanding hashtags and their relationships. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT).*

Miller, G. A. (1995, November). Wordnet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Musat, C., Velcin, J., Rizoiu, M.-A., & Trausan-Matu, S. (2011). Emerging intelligent technologies in industry. In D. Ryko, H. Rybiski, P. Gawrysiak, & M. Kryszkiewicz (Eds.), (chap. Concept-Based Topic Model Improvement).

Porter, M. F. (1997). Readings in Information Retrieval. In K. Sparck Jones & P. Willett (Eds.), (pp. 313–316).

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining* (pp. 399–408).

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010, January). Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)*, *28*(1), 4:1–4:38.

Salvador, S., & Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence* (pp. 576–584).

Sasaki, K., Yoshikawa, T., & Furuhashi, T. (2014, October). Online topic model for twitter considering dynamics of user interests and topic trends. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1977–1985).

She, J., & Chen, L. (2014). Tomoha: Topic model-based hashtag recommendation on twitter. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)* (pp. 371–372).

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*.

Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining (p. 424-433).

Yang, Z., Kotov, A., Mohan, A., & Lu, S. (2015). Parametric and non-parametric user-aware sentiment topic models. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 413–422).

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval* (p. 338-349).