Topic Modeling of Short Texts Using Anchor Words

Florian Steuber
Research Institute CODE
Bundeswehr University Munich
Munich, Germany
florian.steuber@unibw.de

Mirco Schoenfeld
Bavarian School of Public Policy
Technical University Munich
Munich, Germany
mirco.schoenfeld@tum.de

Gabi Dreo Rodosek
Research Institute CODE
Bundeswehr University Munich
Munich, Germany
gabi.dreo@unibw.de

ABSTRACT

We present *Archetypal LDA* or short *A-LDA*, a topic model tailored to short texts containing "semantic anchors" which convey a certain meaning or implicitly build on discussions beyond their mere presence. A-LDA is an extension to Latent Dirichlet Allocation in that we guide the process of topic inference by these semantic anchors as seed words to the LDA. We identify these seed words unsupervised from the documents and evaluate their co-occurrences using archetypal analysis, a geometric approximation problem that aims for finding k points that best approximate the data set's convex hull. These so called archetypes are considered as latent topics and used to guide the LDA. We demonstrate the effectiveness of our approach using Twitter, where semantic anchor words are the hashtags assigned to tweets by users. In direct comparison to LDA, A-LDA achieves 10-13% better results. We find that representing topics in terms of hashtags corresponding to calculated archetypes alone already results in interpretable topics and the model's performance peaks for seed confidence values ranging from 0.7 to 0.9.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**.

KEYWORDS

topic modeling, short text, archetypal analysis, text mining, data mining

ACM Reference Format:
Florian Steuber, Mirco Schoenfeld, and Gabi Dreo Rodosek. 2020. Topic Modeling of Short Texts Using Anchor Words. In *Proceedings of International Conference on Web Intelligence, Mining and Semantics (WIMS'20).* ACM, Biarritz, France.

1 INTRODUCTION

To get a sense of what people are talking about on Social Media platforms one is often faced with the challenge of summarizing a large amount of text data. Hence, Social Media companies and related industries are looking for better tools to summarize more efficiently and more meaningful Social Media posts based on their content. For such tasks, a well-known technique to create summaries which clusters documents by obtaining latent topics is called topic modeling [5]. While, generally, topic modeling is a well-established way to access large collections of text it has considerable difficulties in dealing with short texts where overlapping words across documents are a seldom phenomenon: Topic modeling procedures struggle in capturing the semantics of topics in collections of short documents. This is mainly due to the fact that Topic Modeling is only insufficiently able to handle sparse term document matrices (TDM). Such sparse matrices are therefore often compressed before a topic model is trained on them. A popular method for this is singular value decomposition (SVD). Although this procedure increases the quality of the topic modeling from a statistical point of view, the semantic quality of the topics often suffers.

In this paper, we present a method that supports topic modeling of short documents by evaluating characteristic hints on semantic relationships to pre-configure the topic modeling procedure. These characteristic clues can be, for example, so-called hashtags, which are used in short messages on Twitter, also known as tweets. Hashtags are used by users in their posts to the platform to create associations with one or more topics or discussions. They thus contain valuable information about the semantics and context of a short message, which cannot easily be obtained by analyzing the words alone. Hashtags and especially co-occurrences of different hashtags span a semantic space that is useful for modeling topics. The hashtags are therefore analyzed using a special clustering procedure that describes the feature space by a set of extreme configurations which is resistant against sparse TDMs. This means that the semantic space implicitly contained in the hashtags is encoded into the descriptions of the clusters. This description is then used as a pre-configuration for the topic modeling.

From an information theoretical point of view, we show that the pre-configuration sharpens the topics while their statistical quality remains comparable and even slightly increases. Furthermore, our method is able to support the quantitative evaluation of topics by indicating which topics have emerged from which semantic hashtag clusters. Investigation of hashtag clusters shows that a topic clustering solely based on hashtags and no additional tokens yields equally interpretable topics for topic counts starting at $n = 200$. Finally, we evaluate the model over a range of hyper parameters giving the reader a recommendation of parameter ranges in which the model's performance peaks.

The approach presented here could also prove useful for multilingual corpora. By paying special attention to hashtags, it should be possible to separate tweets by expressed languages sufficiently. However, this will not be evaluated in the present study.

Other texts containing comparable semantic anchor points are political speeches or contributions to debates. They often mention political decisions or laws, which are linked to the topic. This implicitly provides valuable information on the classification of the current contribution and often establishes meaningful references.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WIMS'20, June 30th - July 3rd, 2020, Biarritz, France
© 2020 Copyright held by the owner/author(s).
978-1-4503-7542-9/20/06...$15.00

The remainder of this paper is structured as follows. Section 2 summarizes related work in the field of topic modelling with special attention to the domain of Twitter. Furthermore, the statistical basis of archetypal analysis is explained. Section 3 first gives a rough overview of our model's work flow and then explains each working step in detail. Thereafter, section 4 evaluates the model on an exemplary corpus of tweets and investigates its performance on several intrinsic and extrinsic metrics. Following up, section 5 compares A-LDA to the well known LDA model and discusses suitable values for hyper parameters. Finally, section 6 concludes presented methodology and shows areas of future work.

2 RELATED WORK

This section discusses related work in the fields topic modeling in general and in the domain of Twitter as a social network in particular and shortly summarizes archetypal analysis.

2.1 General Topic Models

Topic models in their early stages extend the concept of Natural Language Processing techniques, especially of vector retrieval models [30]. One early topic model is Latent Semantic Analysis (LSA) [9] which aims to identify common co-occurrences of words in a document corpus using analytical methods. More precisely, a term-document matrix is generated whose elements represent a weighted relationship between two dimensions of the underlying vocabulary calculated as the corresponding TF-IDF score [2]. Oftentimes this matrix is undergone some sort of dimensionality reduction in order to reduce subsequent computation and storage effort. Additionally, compressing term-document matrices partly helps with the sparse population of Bag of Words (BoW) based matrix representations. In LSA, correlations between words are identified by comparing pairs of term vectors with regard to their cosine similarity. Building on that, Probabilistic Latent Semantic Indexing [12] replaces the original assignment function with a mixture decomposition which results in weighting word-document pairs by drawing from conditionally independent multinomial distributions. Latent Dirichlet Allocation (LDA) [5] further improves this model by incorporating a prior distribution for topics and replacing multinomial distributions with dirichlets. Parameter estimation is modelled via variational inference [24].

Multiple extensions have been proposed to the original LDA model. This includes the Relational Topic Model [6] which models a binary random variable between two topics, Correlated Topic Model [4] which accounts for correlating topics, Structural Topic Model [28] which allows to align topics to metadata, or Dirichlet-Multinomial Regression [19] which can incorporate arbitrary features into the dirichlet prior. Another extension called SeededLDA [14] allows incorporating lexical priors into each topic by providing a list of seed words. Inference is conducted using Gibbs Sampling [10]. In this paper, we will utilize SeededLDA by providing a set of provisional topics calculated from a small but expressive subset of the corpus' vocabulary. Further approaches model topics with markov time models [34] or by extracting latent topic from ground truth data [23].

Highly related to the field of topic modeling is the concept of word embeddings which aims for leveraging semantic meanings between tokens and their surroundings. Such word embeddings generally result in probabilistic language models and may be computed by joining context windows of sentences [8] or by training feed-forward neural networks [3]. Input to the network is a 1-in-V encoded BoW vector which then is projected into a lower dimensional representation. Recent work [18] finds that the low dimensional vector representations of multiple trained word embeddings allow for algebraic operations which do express precise syntactic and semantic relations between the operands.

2.2 Topic Models on Twitter

Oftentimes topic modelling is refined to have improved performance on pertinent corpora including social networks. This subsection considers literature for the social network Twitter in particular. A multitude of approaches in this domain utilize the unmodified LDA model itself [22, 32] with the goal of trend detection [16] or topic detection of hashtags [15]. These elaborations often are modelled for a single language, but multilingual models do exist [25]. Many proposed extensions aim to improve LDA's performance on short texts by either incorporating syntactic filtering [21] choosing between multiple BoW distributions while inferring topics [38] or adding components to process temporal changes in inner topic structures [31, 35]. On corpora with known authors of documents, corresponding authorship information can be incorporated in the generative process by modeling a document's topic distribution as a mixture of topics assigned to associated authors [29]. Our approach differs from the Author-Topic-Model in that we do not require authors to exist in the corpus making our approach generalize on different domains.

A predominant technique to significantly increase topic qualities for short text corpora, implicitly including Twitter, is called *pooling*. Pooling postulates that suitable document sizes for topic inference can be created if short texts in general, or tweets in particular, are aggregated with respect to some underlying common feature. In [13] multiple aggregation schemes are discussed. Over time, most pooling approaches specialize on either pooling by conversation [1] or pooling tweets with respect to their author [17].

A somewhat similar approach is the introduction of some hierarchical categories that add a further level of abstraction to LDA's inference process [37]. This can be implemented by choosing a set of categories or document labels defined in either a supervised or unsupervised fashion. A single topic then may only be associated to a restricted subset of categories. An example is the Labelled LDA approach [27] where each word is assigned to a weighted mix of document labels that correspond to characteristics of the underlying post. Labelled LDA is successfully applied to compare similarities of user profiles [26]. Similar to Labelled LDA, our approach also creates implicit labels. However, contrary to representing a classification of a tweet's characteristics our labels correspond to desired topics directly. We find semantically expressive tokens, namely hashtags, to be a good topic identifier as long as they can be clustered in a semantically coherent way rendering them a solid choice for topic labels. The algorithm leading to hashtag clusters will be presented in following section.

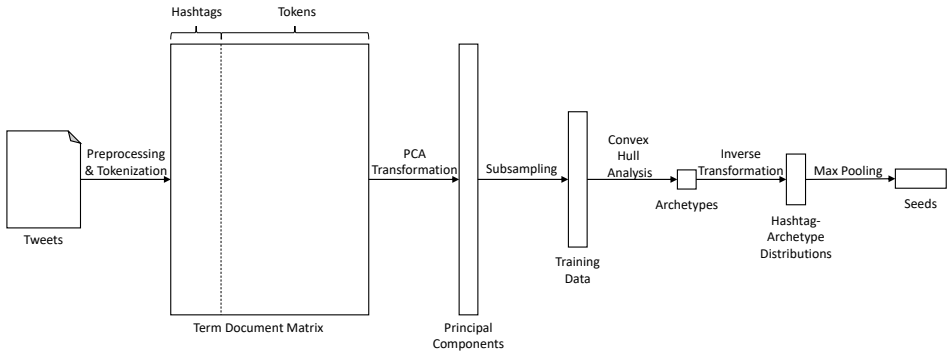

Figure 1: Workflow of Seed Generation

3 ARCHETYPAL LDA

In this section, we propose *Archetypal LDA*, or short *A-LDA*, a topic model and extension of Latent Dirichlet Allocation tailored to the social network Twitter. We first summarize the algorithm's work flow to give the reader an intuition of the general work flow and then explain each individual step in detail.

3.1 General model

We now give a brief overview of the model's work flow. The algorithm takes a collection of tweets together with a desired number of topics k as input and outputs k topics fitting to the provided corpus. From a general perspective, the algorithm can be subdivided into two major parts. The first part, namely *Seed Generation*, is illustrated in figure 1 and starts by performing some data preprocessing which is described in depth in section 4. It includes the elimination of noisy tokens in each document, i.e. tokens that contain non-latin characters, and ranges over the identification of reasonable Bag of Words categories to the computation of corresponding term document matrices. Afterwards, all training data is undergone dimensionality reduction with Principal Component Analysis and subsequent subsampling to significantly reduce computation times for following steps. Based on this, A-LDA computes preliminary topics by expressing the entirety of a corpus' meaningful tokens as a mixture of data points in feature space. Each data point can be interpreted as a latent topic of the represented tokens. We make use of the semantically expressive character of Twitter's hashtags in that they are a solid instantiation of meaningful tokens, i.e. we consider each hashtag to convey a sense of meaning or let the reader connect to a known topic in a way that can't be grasped from the words of the tweet alone. In other scenarios, meaningful tokens can be established terms comparable to Twitter's hashtags. As a next step, however, each hashtag is assigned to its most representative latent topic in terms of cosine similarity.

Seeds are then provided to the second part of the algorithm, in which we utilize SeededLDA [14] to infer topics for the entirety of non-hashtag tokens in the corpus. Provided seeds are used to govern topic inference in a semi-supervised manner. The procedure is capable of adjusting major flaws of the archetypal topic representation. Therefore, this step can be seen as both an assignment process of non-hashtag tokens and a fine-tuning of provided hashtag clusterings.

3.2 Seed Generation

As mentioned earlier, this algorithm aims for creating a topic model tailored to Twitter by assigning large weights to meaningful tokens, i.e. hashtags. The latter can be seen as a summarization of a tweet in as few as one or two phrases. Intuitively, a topical clustering solely based on hashtags should yield semantically comparable results. There are two problems with this intuition, however. First, topic models such as LDA already perform poorly on short texts with sparse term-document matrices if no guidance or pooling technique is involved rendering an even more sparse hashtag-TDM a bad approach to infer suitable topics. Second, only a small subset of 10-15% of tweets are annotated with hashtags. Therefore, the algorithm would not be capable of predicting topics on the vast majority of tweets. The concept therefore needs to be extended in that we first need to approximate a topic clustering for hashtags only and then refine this clustering by incorporating any non-hashtag token into each topic.

Direct clustering of topics based on sparsely co-occurring hashtags is avoided by instead solving a complementary task: We compute an approximation of the data set's convex hull with k points as vectors in n-dimensional principal component feature space. Requiring the data points to be located on the convex hull allows us to represent them as substitutes for any other data point via weighted assignment. Such a vector therefore forms a mixture of multiple data points. As we choose data points to be entries of the hashtag-BoW, the resulting hull points can be seen as a rough approximation of an underlying latent topic. The problem described can be solved by Archetypal Analysis in which case the calculated points may be referred to as *archetypes*.

Archetypal analysis [7] refers to a statistical approach to represent a set X of multivariate data points $\mathbf{x}_i \in \mathbf{R}^m$ as a mixture of p individual vectors $\mathbf{z}_1, ..., \mathbf{z}_p$. These vectors are also called *pure types* or *archetypes* of the data set as they aim to represent the set's most pure or extreme configurations.

Given archetypes $\mathbf{z}_1, ..., \mathbf{z}_p$, each data point $\mathbf{x}_i \in X$ can be expressed as a mixture or linear combination of archetypes. The best approximation of coefficients α_{ik} for a single point is given by the minimizer of the L_2-norm

$$\left\| \mathbf{x}_i - \sum_{k=1}^{p} \alpha_{ik} \mathbf{z}_k \right\|_2^2 \tag{1}$$

under the constraints $\alpha_{ik} \geq 0$ and $\sum_k \alpha_{ik} = 1$. Hence, the best approximation for archetypes $\mathbf{z}_1, ..., \mathbf{z}_p$ minimize the summed squared error over all data points. Equivalently, this problem can be expressed as finding the maximizers of function

$$\sum_{k=1}^{p} \mathbf{z}_k^T S \mathbf{z}_k \tag{2}$$

with $S = X^T X$. The solution to this problem are the eigenvectors of S corresponding to the largest eigenvalues. Vice versa, the archetypes $\mathbf{z}_k$ themselves are considered to be mixtures of data points $\mathbf{x}_i$, such that

$$\mathbf{z}_k = \sum_{j=1}^{n} \beta_{kj} \mathbf{x}_j \tag{3}$$

subject to constraints $\beta_{kj} \geq 0$ and $\sum_j \beta_{kj} = 1$.

Archetypes are computed using an alternating minimization approach. Over the course of multiple iterations, equation (1) is first minimized to express all $\mathbf{x}_i$ in terms of $\mathbf{z}_k$. In the second step, archetypes $\mathbf{z}_k$ are estimated using equation (3) slightly improving previous approximations for archetypes. Since the constraints for coefficients α_{ik} and β_{kj} require non-negativity the archetypes need to enclose X dragging them to the dataset's edge in geometric space. This implies that archetypes in fact represent a rough approximation of the dataset's convex hull, hence, the algorithm itself sometimes is also called *Principal Component Hull Analysis*. Additional work has been conducted that aims to improve the algorithm's time and space complexity [36].

As archetypes indicate a mixture of data points, they can be interpreted as latent topics in a bag of words feature space. In our implementation we chose k to be the number of desired topics in the resulting topic model. However, this can be relaxed in other settings, e.g. where the share of documents containing semantic anchors is significantly smaller.

After assigning a vector representation to k preliminary topics, every hashtag afterwards is associated with exactly one of each topics. Because of the previously conducted PCA there is no direct notion of hashtags in the current feature space. However, both the principal components as well as the calculated archetypes can be transformed back into the original hashtag feature space using PCA's transformation matrix. After the inverse transformation each unique hashtag may now be assigned to a topic by selecting the topic's corresponding archetype whose vector representation is closest to the hashtag in terms of cosine similarity. We refer to hashtags corresponding to archetypes as *Archetypal Hashtags*. Even though hashtags belong to a mixture of topics the preliminary clustering based on minimum distance effectively acts as max pooling and forces the model to decide for a single topic. Note that the following refinement process may still adjust assignment weights and distribute shares of a hashtag to other suitable topics. The assignments of archetypal hashtags are treated as a preliminary topic clustering and are passed as seeds to LDA as seen in the following section.

3.3 Semi-Supervised Learning

The archetypal hashtags identified in the previous sections contain valuable hints on the latent topics of the short texts since these hashtags are assigned by users. We therefore regard the hashtags as "semantic anchors" in the short texts. These semantic anchors need to be incorporated into the training of the LDA which is done by inputting the hashtags as *seed words* for the so-called SeededLDA [14]. In their work, the authors use seed words to influence both the topic-word distributions and the document-topic distributions. More precisely, the generative process of estimating these distributions is guided by seed information on the word level, i.e. a set of user-defined words that are characteristic for the topics in the corpus. Thereby, the procedure is comparable to bootstrapping [33] or prototype-based learning [11].

To recall the original LDA procedure, the topic-word distributions and the document-topic distributions are given by ϕ_k and θ_d:

- For each topic $k = 1 \dots T$ choose $\phi_k \sim Dir(\beta)$
- For each document d, choose $\theta_d \sim Dir(\alpha)$.
- For each token $i = 1 \dots N_d$:
  - Select a topic $z_i \sim Multinomial(\theta_d)$
  - Select a word $w_i \sim Multinomial(\phi_{z_i})$

where T is the number of topics, α and β are Dirichlet priors on the per-document topic distributions and the per-topic word distribution, i.e. hyperparameters of the model, ϕ_k is the Multinomial word distribution for topic k, and θ_d is the topic distribution for document d.

Regarding the topic-word distributions in SeededLDA, each of the T topics is a mixture of *two* topic-word distributions $\phi^r_{z_i}$ and $\phi^s_{z_i}$ which are *r*egular topic distributions and the associated *s*eed topic distributions. By specifying a parameter π_k, the probability is fixed with which words are drawn either from $\phi^r_{z_i}$ or $\phi^s_{z_i}$ effectively allowing the user to configure the influence of the seed topic distributions on the generation of topic-word distributions. To estimate the document-topic distribution, first, the seed information is transferred to the documents containing the relevant seed words and a group of the identified document is sampled. Then, from this group of documents an intermediate group-topic distribution is drawn as a prior to the estimation of the document-topic distribution θ_d for the whole corpus. This relaxes the relation between the number of seed and regular topics and ties the topic distributions of all documents within a group.

In A-LDA, the set of seed words is given by the archetypal hashtags identified in the previous sections. These seed words are then used to guide the LDA following the above-mentioned approach by the SeededLDA. To ease the following evaluation, we let the number of archetypes k decide upon the number of topics T such that $k = T$.

Note that the step of semi-supervising the LDA is tied to the unsupervised identification of semantic anchors by using archetypal analysis, i.e. the procedure of A-LDA remains an unsupervised variant of LDA.

A crucial aspect of supervising the LDA, however, is selecting a suitable seed confidence parameter π_k. This parameter controls to what extent seed words are allowed to become part of other topics. How we fixed this parameter will be part of the following Section.

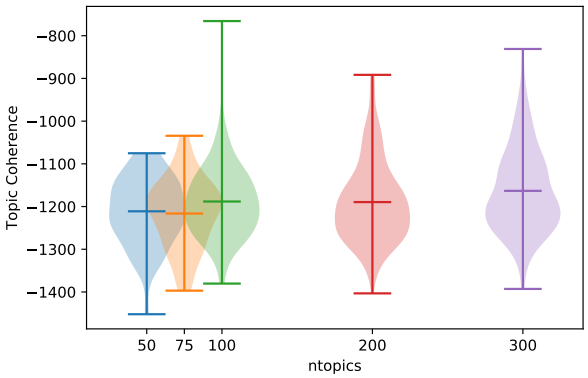

Figure 2: Topic Coherences in A-LDA as function of topic counts

4 EXPERIMENTS

The following section evaluates the effectiveness of the proposed algorithm A-LDA in terms of performance on a list of topic model metrics. We discuss suitable value ranges for underlying hyper parameters and finally present resulting words for selected topics to give the reader an intuitive idea of model's quality.

Subsequent results were obtained on a training data set consisting of roughly 100 million tweets. The data covers four weeks starting mid October 2019 and is collected via various Twitter API's, above all Twitter's Streaming API. To ensure unbiased and heterogeneous results all tweets were evenly split into buckets of one million entries and evaluated via cross validation. Multiple passes of our A-LDA approach were conducted with varying values $k \in \{50, 75, 100, 200, 300\}$ and a fixed number of 500 iterations per pass.

4.1 Data preprocessing

Data preprocessing was performed on the textual content of each tweet in order to later define suitable Bag of Words (BoW) categories on the cleaned up corpus. It included tokenization as well as discarding any tokens with a length of fewer than three characters. Punctuation and special characters were removed with exception of the #-character as this tag was used for hashtag identification. Words were normalized by mapping them to their lower case representations and non-latin characters were omitted. We refrained from utilizing word lemmatization because appropriate converting algorithms are fully developed for few languages only and we wanted to maintain the ability of process language independent texts with equal precision. Likewise, word stemming was avoided because of the general loss of expressive power. For the sake of interpretability, we restrict presented evaluation results to the unilingual execution.

All cleaned up tweets were used to generate two different bag of words categories with corresponding Term-Document matrices (TDM). The first one consisted of hashtags only extracted either by a preceeding #-tag or by retrieving the *entities.hashtags* key from a tweet's JSON representation. We will refer to this subset as *hashtag-BoW* and *hashtag-TDM*, respectively. The second category included the entirety of tokens, i.e. all non-hashtag tokens plus the hashtag-BoW. Hence, the hashtag-BoW was a strict subset of the common BoW. Both categories were filtered such that the most frequent 10% of words, corresponding to corpus specific stop words, were discarded. Likewise, the least frequent 20% of tokens were omitted as well as they would primarily add noise to the model. Note that both categories were utilized for distinct tasks: the hashtag-TDM was used for computing initial seeds only whilst the complete TDM was passed to SeededLDA for fine tuning and assigning non-hashtag tokens to topics.

As precautionary measure for upcoming approximation tasks, we followed up with an intermediary step of dimensionality reduction. This substantially reduced the hashtag-TDM in size while maintaining a suitable amount of expressive power. More precisely, dimensionality reduction was conducted by using Principal Component Analysis (PCA). We determined the number of dimensions to project onto empirically by performing an exploratory run of PCA on the data. We found that when transforming the data set down from 60-80k to as few as 2000 dimensions the model was still capable to express over 85% explained variance of the original model. PCA's transformation matrix was saved so that an inverse transformation later could be conducted at any point reversing the principal components back into the original feature space. For further decrease in size we subsampled rows of the transformed TDM with a sampling rate of 0.7. Subsampling needs to be done after dimensionality reduction in order to not lose expressive power of principal components.

4.2 Quantitative Analysis

We report results for various intrinsic evaluation metrics namely topic coherence, model perplexity, and number of words per topic. All experiments conducted in this section are based on passes of A-LDA with varying numbers of topics selected out of the set $n \in \{50, 75, 100, 200, 300\}$. The goal is to empirically determine a reasonable number of topics on the given corpus. Building on a limited range of topic numbers we then compare our presented approach A-LDA to the well known Latent Dirichlet Allocation.

Topic Coherence, as defined by Mimno et al. [20], describes a measure of the co-occurrence frequency between all disjoint pairs of a topic's $M = 5, \ldots, 20$ most representative words. The term *most representative* refers to the probability that a specific word is generated by its corresponding topic as seen in the model's posterior topic-word distribution ϕ_k. The authors postulate that topic coherence is a measure very much corresponding to a qualitative classification of inner topic consistency that a human judge would give. Mathematically, topic coherence is defined as

$$C(t, V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{4}$$

where $D(v)$ denotes the document frequency of a word v, $D(u, v)$ denotes the co-document frequency of two words and $V(t)$ are the most probable words in topic t. According to this definition topic coherence increases as co-occurrence counts of most frequent words increase as well.

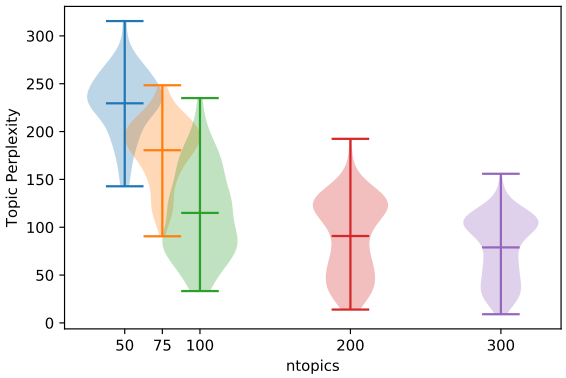

Figure 3: Topic Perplexity in A-LDA as function of topic counts

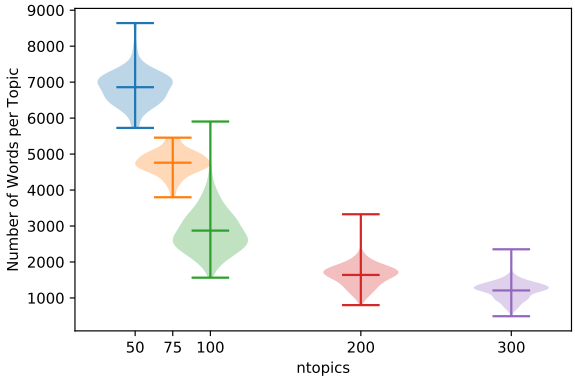

Figure 4: Number of Words per Topic in A-LDA as function of topic counts

We report results in figure 2. Each violin plot represents the distribution of topic coherence values for a pass of A-LDA with topic counts given on the x-axis. Average coherence values slightly increase with topic numbers and, as topic numbers increase to 100 or higher, we notice the emergence of a few topics that relatively score up to 30% better than the distribution's mean. The figure therefore indicates to use at least 100 topics to obtain reasonable results.

Model Perplexity measures how well the topic-word distributions ϕ_k predicts samples of data. It can be calculated as $2^{H(x)}$ where $H(x)$ is the model's posterior distribution entropy. Note that entropy maximizes if all words are assigned with an equal probability over multiple topics. However, such a model would be least expressive because it effectively fails to assign topics all together. Instead, it is preferable to maximize a word's assignment probability for a single topic inducing a notion of lightweight determinism. This on the other hand results in lower values of entropy rendering models with lower perplexity to be favorable.

Figure 3 visualizes the changes in each topic's perplexity as the number of topics change. Here, we can see a strong improvement in perplexity with increasing topic numbers. After 100 topics this increase slowly stagnates, however. As perplexity values halve for three-digit topic counts this is a strong indication to refrain from lower topic counts. We will later see that this indication mirrors in a topic's interpretability as seen by a human judge.

Number of Words per Topic Partly in accordance to a model's perplexity we instead might ask about the number of words that are assigned to each topic. This gives information about how topic sizes are distributed and whether few topics correspond to small and isolated group of words indicating highly coherent topics which indeed may be desirable. Another view on the metric is to identify topics that function as large collection basins, i.e. topics whose included words could not be assigned to more semantically coherent topics but have low inner topic consistency. We define the number of words within a topic in LDA's probabilistic environment as the cardinality $card(M^{(\theta)})$ where $M^{(\theta)}$ is a subset of topic θ's posterior topic-word distribution. More formally, we define M as

$$M^{(\theta)} = \{x | P(x|\theta) > \min_{y} P(y|\theta), y \in \phi_\theta\}, \tag{5}$$

the set of words in a topic that have assigned a higher than the minimum appearance probability.

Results for word counts in accordance to topics are shown in figure 4. Unsurprisingly, the number of words drastically diminish as topic counts increase. As underlying statistical property we also notice a small decrease in the variances of word counts with the exception of $n = 100$ topics. Even though variances decrease the difference between maximum and minimum for 300 topics still is 2000 words which manifests the quality difference between extreme configurations of topics.

Summarizing the findings previous quality metrics we find that topic counts starting from 100 upwards result in a reasonable model qualities whereby larger topic counts are always preferable. Based on this insight we omit lower counts and restrict further evaluations to be performed on 100 to 300 topics only.

4.3 Topic Interpretability

We now look at the resulting topics and investigate topic interpretability based on human judgement by running A-LDA multiple times with varying topic counts $n = 100, 200$ and 300. Selected examples of resulting topics are reported depending on the topic count. For each topic, all archetypal hashtags are listed that were provided as seeds to SeededLDA. In addition, the most probable tokens of the same topic after training are shown as well.

At first, we focus on examining archetypal hashtags. A coherent clustering of hashtags would confirm our premise of archetypal analysis being a reasonable approach for preliminary topic groupings. In table 1 results are shown for one of 100 topics. Due to the relatively low topic count any topic is assigned a vast number of hashtags. This circumstance reflects in each topic's purity. Archetypal hashtags from multiple domains are mixed up making it hard to find a consistent and common topic label even for humans. Interestingly, seeds do not negatively affect the interpretability of tokens after training which mainly can be subsumed under the

context *Religion*. Looking back at archetypal hashtags we indeed find dispersed tokens corresponding to the same category as well.

The quality of results significantly improves when increasing topic numbers as seen in tables 2 and 3 for $n = 200$ and $n = 300$ topics, respectively. This indicates that less than 200 topics simply is an inadequate choice of topic counts. To obtain comparable results, we chose presented topics such that there are some category overlaps over different topic counts. For example, the first topic presented in table 2 corresponds to the same category as the first topic in table 3. Consistency of both archetypal hashtags and tokens are positively affected by increasing topic sizes. Quite noticeable, both categories yield satisfactory topic clusters which are pairwise consistent. Unsurprisingly, in contrast to most relevant words, archetypal hashtags often correspond to some real word event and hence bear more information per token. Archetypal hashtags within a single topic belong to the same context with few intruder words. Hence, forcing an assignment of hashtags towards points on the enveloping convex hull indeed provides a reasonable topic grouping on its own already. Incorporating additional tokens into each topic further refines the model.

5 DISCUSSION

This section discusses differences in model quality between LDA and A-LDA by investigating the impact an incorporation of archetypal hashtags has on the final results. We also discuss how well A-LDA does on preserving provided seeds through inferring non-hashtag tokens into topics.

5.1 Effects of Seed Confidence

Seed Confidence is a hyper parameter in SeededLDA's configuration and refers to the likelihood of the computed seeds being an accurate clustering of topics. The higher its value the more it restricts LDA's capabilities to infer other suitable topics for given tokens. In other words, the closer seed confidence converges to one the fewer changes will be made to the initial clustering. While provided seed assignments are progressively getting more fixed with higher values of seed confidence this exact statement is not true for any other tokens in the corpus. Latter ones will still be freely distributed to each topic with respect to carried out inference.

A-LDA transfers the task of clustering tokens from a problem of finding co-occurrences into a geometric problem of approximating the data set's convex hull. This on the other hand alters the optimization objective from identifying the most similar existing data points in feature space to calculating nearly arbitrarily created points in the topic hypercube, the so called archetypes, that describe the data set's individual components. They do not correspond to a token directly but rather can be seen as a latent topic. Every existing data point then can be represented as a linear combination of all archetypes effectively creating a notion of weighted assignments. Hence, an archetype can be expressed as a weighted combination of tokens. Note that this intuition coincides with LDA's idea of representing topics as a mixture of words.

We now investigate the impact of providing pre-clustered hashtags conducted on multiple runs with ever increasing seed confidence values $\{0, 0.5, 0.7, 0.9, 1\}$. A seed confidence level of 0 ignores all provided seeds rendering the outcome identical to a non-seeded

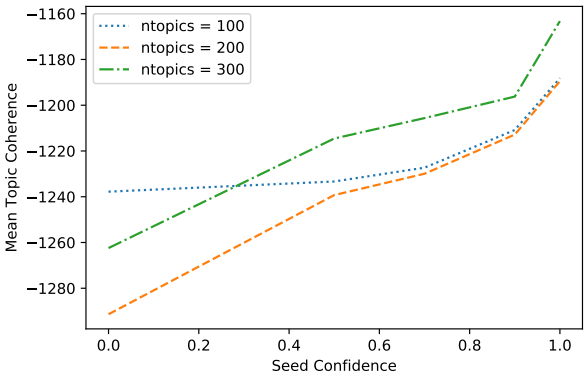

Figure 5: Effects of Seed Confidence in terms of mean Topic Coherence

pass through LDA. Each pass of LDA and SeededLDA is executed on multiple topic numbers creating a total of 15 different scenarios. Figure 5 depicts resulting model qualities in terms of average topic coherence scores when clustering the corpus into 100, 200, and 300 topics, respectively. Each line corresponds to a pass with a fixed number of topics. For all passes, increasing the seed confidence results in increasingly better topic coherence values. The increase gradually becomes stronger when raising the number of topics and reaches up to 10% better results compared to the original LDA model. In terms of resulting model coherence it is always favorable to include preliminary hashtag clusters and setting seed confidence values near the maximum of one provides the most optimal results.

Improvements as well are noticeable in figure 6 where we consider average topic perplexity as subject of optimization for identical runs. This time, the steepest decrease in costs occurs at 100 topics. Interestingly, 200 and 300 topics show immediate improvements when regarding the binary decision whether a seed should be used or not but hardly yield better results when varying seed confidence levels of greater or equal 0.5. In all cases, A-LDA achieves improved quality values up to an increase of 13% over the original model.

5.2 Hashtag Groupings

A major goal of this approach was to emphasize a corpus' meaningful tokens in the learning process by assigning large weights to it. Depending on the corpus, these meaningful tokens can be thought of as semantically expressive topic summarizations or categorical descriptions. In the case of Twitter, they manifest as hashtags as these are supposed to subsume the content of a tweet in as few as one or two phrases. Hence, we can interpret hashtags to be a rough summarization of what topic the actual text content is about.

It is therefore natural to ask how much differently hashtags are assigned in the presented approach in contrast to the well known LDA itself. To compare the quality difference of individual passes with respect to this intuition, we formulate the problem of calculating hashtag takeover rate as follows. Let the calculated archetypal hashtags for each individual topic be given. Now we ask to what extent this exact hashtag grouping will be carried over

Table 1: Examples of archetypes and most probable tokens word-topic-distribution with 100 topics

Type	100 Topics
Archetypal Hashtags	Action, WeIndiansAreBrothers, ASAPNatinTo, MichaelJackson, romance, Kaviliya, Jesus, WeIndiansAreUnited, AngelinaJolie, Peace, BreakingNews, Sanditon, hkPoliceBrutalitiy, lol, ThisIsTema, luxury, ResistanceUnited, SuperSoulSunday, worship, TWGRP, ModernLove, LeadRight, Kurdistan, DarkBlueKiss, BorisLetter, VoteBlueToEndThisNightmare, tribute, superfan, cancer, bear, GalaxyNote, plantbased, colors, Fashion, SciFi, Scary, kavinlosliya, kavinians, earth, Borderlands, ImpeachDonaldTrumpNOW, BoomerSooner, supreme, Life, Happiness, true, asianboy, Sneziey, RTB, Plastic, Giveaways, CDNmedia, 2ndAmendment, misfit, Christianity, AliDeserveToStay, cnn
Relevant Tokens	true, worship, Jesus, earth, cancer, real, Life, Christ, power, time, like, Peace, bear, mother, death, Allah, dream, people, Cancer, peace, faith, world, sports, given, Action, supreme, colors, Happiness, Mina, prayer

Table 2: Examples of archetypes and most probable tokens word-topic-distribution with 200 topics

Type	200 Topics
Archetypal Hashtags	freedom, WorldLargestFlag, StandwithHK, StandwithCatalonia, SOSHK, Freedom, Facts, SpainIsAFascitEstate, MukamisMashujaaMessage, religion, memesdaily, humanrights, ISPR, GodBlessAmerica, ResistBeijing, Antisemitism, OurMartyrsOurPride, OurForcesOurPride, FreeIran
Relevant Tokens	freedom, religion, people, Freedom, community, history, Facts, Hindus, stay, American, Stedman, democracy, queen, important, hypocrite, trumps, Americans, sports, Like, Religion, tweet
Archetypal Hashtags	grateful, wizkid, Future, sale, unique, work, insurance, trading, BusinessMgmt, commission, traveling, resistance, inspiring, holidays, parking, craft, progress
Relevant Tokens	work, morning, people, grateful, sale, progress, home, running, unique, weekend, parking, Future, like, today, together, kids, incomplete, Vande, trading, Woman, scrubs, Mataram, racist, relax, resistance, holidays, tired, stay
Archetypal Hashtags	vegan, indiedev, newmusic, indiegame, RavensFlock, Marvel, TuneIn, Overwatch, screenshotsaturday, SpiderMan, peace, indie, DC, feedly, puppy, AvaTweet, indiegames, asian, MashujaaSunday, mindset, sharing, pop, avengers, KIMMINJU, AvengersEndgame, comicbooks, minju, instrumental, musicvideo, dccomics, indiemusic, SkilledTrade, RPG, costume, unsignedartist, ios, designer
Relevant Tokens	like, costume, peace, sharing, real, important, cool, Ajith, designer, movies, Marvel, Bigil, Overwatch, Valimai, Ford, puppy, development, vegan, Godfather, Supreme, film, tinamad, umisip, BlackMagic2019, movie

Table 3: Examples of archetypes and most probable tokens word-topic-distribution with 300 topics

Type	300 Topics
Archetypal Hashtags	freedom, WorldLargestFlag, StandwithHK, StandwithCatalonia, SOSHK, campaign, Freedom,Facts, SpainIsAFascitEstate, MukamisMashujaaMessage, religion, memesdaily, humanrights, ISPR, GodBlessAmerica, ResistBeijing, Antisemitism, OurMartyrsOurPride, OurForcesOurPride, FreeIran
Relevant Tokens	campaign, freedom, religion, HongKong, election, Freedom, people, Warren, respect, HongKongers, democracy, facebook, stories, first, ordering, trumps, Rosie, HongKongers, government, running, bargaining, beacon, bring, Warren's, Girls, fake, culture
Archetypal Hashtags	Action, Jesus, Bible, Prayer, Christian, Nature, Catholic, church, King, Family, YourBattles, speaklife, earth, Do, Culture, death, soul, Shalom, Christ, mother, Transformers, SundaysAtICGC, SadhguruInNYC, MSG, Church, rubber, prayer, meat
Relevant Tokens	Jesus, mother, King, Christian, Church, Christ, Action, church, soul, earth, rubber, Family, Bible, death, time, Wendy, Nature, prayer, crackdowns, girl, like, death, CNN, lot, Sierra, Christ, kins, like
Archetypal Hashtags	FridayThoughts, ocean, photooftheday, usa, oil, ThePhotoHour, Anthropocene, beautifulday, sea, plasticpollution
Relevant Tokens	ocean, heart, enjoy, shoes, restaurant, home, green, horses, hooping, Ocean, Unfollowed, soul, Followed, monitored, plastic, vacation, event, pressure, uncertainty, surprised, Mina

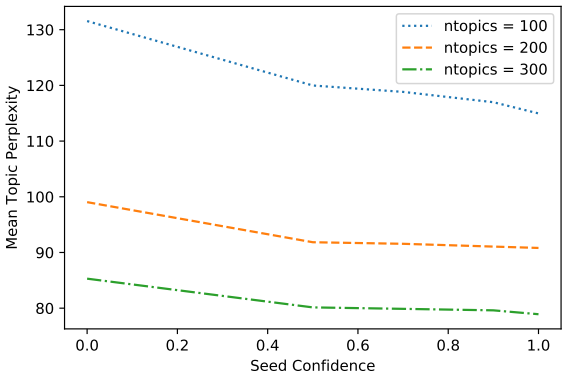

Figure 6: Effects of Seed Confidence in terms of mean Topic Perplexity

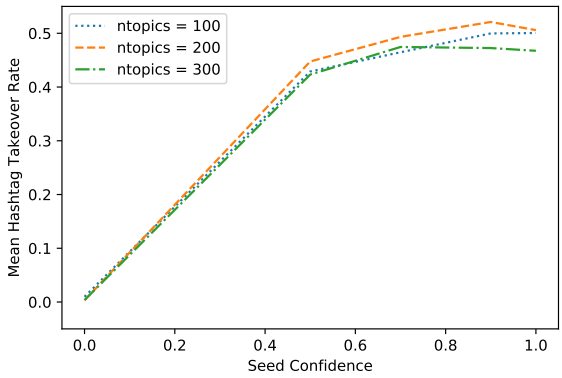

Figure 7: Hashtag Takeover Rate as function of Seed Confidence and Number of Topics

after computing both seeded and non seeded LDA on a data set. In other words, we examine to which of the topics each archetypal hashtag is assigned with the highest probability as seen in LDA's posterior topic-word distributions. It is desirable that assignments are retained after training.

We report results in figure 7 for steadily increasing numbers of topics and seed confidence levels. Again, a seed confidence value of zero corresponds to using LDA with no seeds at all and results are reported for seed confidence values $\{0, 0.5, 0.7, 0.9, 1\}$. In any scenario, A-LDA has a much higher retaining rate of archetypal hashtags rendering it a good approach to preserve previous hashtag groupings over the course of the training. With no seeds at all, hashtags are distributed arbitrarily over all topics to some extent. Furthermore, the average retaining rate of A-LDA increases with higher seed confidence levels up to the point of 0.7 to 0.9. Somewhat surprisingly, the takeover rate slightly falls for the highest possible seed confidence values. This might be an artefact arising at implementation level for extreme probability values. The average takeover rate slightly drops for 300 topics compared to the other scenarios. This is due to the fact that there are many more options to which topic a hashtag can be assigned.

Summarizing we find that the drastically increased retaining rates of previously clustered hashtags strongly indicate that seeding the model with semantic anchor words yields better results that the conventional model. The premise that both unique hashtags and groups of hashtags are insightful summarizations for a latent topic has already been discussed in the last section. Summarizing all aspects considered in this chapter, we recommend to select a seed confidence value of roughly 0.7 to 0.9. This value arises as a trade-off between model quality in terms of topic coherence, topic perplexity, and hashtag takeover rate.

6 CONCLUSION

In this paper we presented *Archetypal LDA*, or short *A-LDA*, a topic model specifically tailored to the social network Twitter. A-LDA is an extension to Latent Dirichlet Allocation in that we guide the conventionally unsupervised process of topic inference by providing a set of seed words that topics should be built around. We find that Twitter's hashtags are suitable instantiations for this task as they are often semantically expressive tokens. This is because they effectively summarize the message of a tweet and also can correspond to real world events which oftentimes have a rich semantical context. We determine groups of related hashtags by utilizing a geometric approximation problem called archetypal analysis. We find that when approximating the convex hull of the corpus' TDM with a set of k archetypes, where k is chosen to be the final topic count, each pure type is an inherent representation of a potential topic. Each hashtag is assigned to its closest archetype in terms of cosine similarity. Groups of hashtags belonging to the same archetype are passed to SeededLDA a variant of LDA allowing for incorporation of preliminary clusterings of words to each topic.

Experiments show that topic counts varying around 200 to 300 topics yield solid results in the final model's quality as shown in multiple intrinsic and extrinsic metrics including topic coherence, topic perplexity, and topic interpretability. In direct comparison to LDA, A-LDA achieves 10-13% better results in all statistical model metrics than its predecessor. We also find that representing topics in terms of their assigned hashtags alone already results in interpretable topics. After careful empirical evaluation of different model hyper parameters we advice readers to set SeededLDA's seed confidence value around 0.7 to 0.9 in order to maximize the model's output performance.

In the future, we will evaluate our approach on multilingual Twitter corpora to see if archetypal hashtags are able to separate tweets based on their language. Further, we would like to transfer the concept of archetypes to work on more general corpora than solely on Twitter. In order to translate the concept an equivalent to hashtags as meaningful tokens has to be found. With regard to this, related literature show promising results in the areas of unsupervised category identification and tag labelling.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their constructive comments, the Chair for Communication Systems

and Network Security, and the research institute CODE for their helpful feedback. Research supported, in parts, by EC H2020 Project CONCORDIA GA 830927.

REFERENCES

[1] David Alvarez-Melis and Martin Saveski. 2016. Topic Modeling in Twitter: Aggregating Tweets by Conversations. In *Tenth International AAAI Conference on Web and Social Media*. The AAAI Press, Palo Alto, CA, USA, 519–522.

[2] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern Information Retrieval*. Vol. 463. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3, Feb (2003), 1137–1155.

[4] David M. Blei and John D. Lafferty. 2005. Correlated Topic Models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'05)*. MIT Press, Cambridge, MA, USA, 147–154.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022.

[6] Jonathan Chang and David M. Blei. 2009. Relational Topic Models for Document Networks. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009 (JMLR Proceedings)*, David A. Van Dyk and Max Welling (Eds.), Vol. 5. JMLR.org, Clearwater Beach, Florida, USA, 81–88. http://proceedings.mlr.press/v5/chang09a.html

[7] Adele Cutler and Leo Breiman. 1994. Archetypal analysis. *Technometrics* 36, 4 (1994), 338–347.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Minneapolis, MN, USA, 1–16.

[9] Susan T Dumais. 2004. Latent Semantic Analysis. *Annual Review of Information Science and Technology* 38, 1 (2004), 188–230.

[10] Thomas L Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences* 101, suppl 1 (2004), 5228–5235.

[11] Aria Haghighi and Dan Klein. 2006. Prototype-Driven Learning for Sequence Models. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (New York, New York) *(HLT-NAACL '06)*. Association for Computational Linguistics, USA, 320–327. https://doi.org/10.3115/1220835.1220876

[12] Thomas Hofmann. 2013. Probabilistic Latent Semantic Analysis. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, United States, 50–57.

[13] Liangjie Hong and Brian D Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*. Association for Computing Machinery, New York, NY, United States, 80–88.

[14] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, 204–213.

[15] Roland Kahlert, Matthias Liebeck, and Joseph Cornelius. 2017. Understanding Trending Topics in Twitter. In *Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband*. Gesellschaft für Informatik eV, Bonn, Germany, 10.

[16] Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line Trend Analysis with Topic Models:# twitter trends detection topic model online. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, 1519–1534.

[17] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, United States, 889–892.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, 3111–3119.

[19] David Mimno and Andrew McCallum. 2008. Topic Models Conditioned on Arbitrary Features with Dirichlet-Multinomial Regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence* (Helsinki, Finland) *(UAI'08)*. AUAI Press, Arlington, Virginia, USA, 411–418.

[20] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 262–272.

[21] Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory Search and Topic Summarization for Twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*. The AAAI Press, Menlo Park, CA, USA, 384–285.

[22] David Alfred Ostrowski. 2015. Using Latent Dirichlet Allocation for Topic Modelling in Twitter. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. IEEE, Anaheim, CA, USA, 493–497.

[23] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections. In *Proceedings of the 17th International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, United States, 91–100.

[24] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. 2000. Inference of Population Structure using Multilocus Genotype Data. *Genetics* 155, 2 (2000), 945–959.

[25] Dasha Pruss, Yoshinari Fujinuma, Ashlynn R Daughton, Michael J Paul, Brad Arnot, Danielle Albers Szafir, and Jordan Boyd-Graber. 2019. Zika discourse in the Americas: A multilingual topic analysis of Twitter. *PloS one* 14, 5 (2019), 23.

[26] Daniele Quercia, Harry Askham, and Jon Crowcroft. 2012. TweetLDA: Supervised Topic Classification and Link Prediction in Twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*. Association for Computing Machinery, New York, NY, United States, 247–250.

[27] Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing Microblogs with Topic Models. In *Fourth International AAAI Conference on Weblogs and Social Media*. The AAAI Press, Menlo Park, CA, USA, 130–137.

[28] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58, 4 (2014), 1064–1082. https://doi.org/10.1111/ajps.12103 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12103

[29] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. 2012. The Author-Topic Model for Authors and Documents. *CoRR* abs/1207.4169 (2012), 8. arXiv:1207.4169 http://arxiv.org/abs/1207.4169

[30] Gerard Salton and Michael J McGill. 1983. *Introduction to Modern Information Retrieval*. mcgraw-hill, New York, NY, USA.

[31] Kentaro Sasaki, Tomohiro Yoshikawa, and Takeshi Furuhashi. 2014. 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)*. Association for Computational Linguistics, Doha, Qatar, 1977–1985.

[32] Marina Sokolova, Kanyi Huang, Stan Matwin, Joshua Ramisch, Vera Sazonova, Renee Black, Chris Orwa, Sidney Ochieng, and Nanjira Sambuli. 2016. Topic Modelling and Event Identification from Twitter Textual Data. *arXiv preprint arXiv:1608.02519* abs/1608.02519 (2016), 1–17.

[33] Michael Thelen and Ellen Riloff. 2002. A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 (EMNLP '02)*. Association for Computational Linguistics, USA, 214–221. https://doi.org/10.3115/1118693.1118721

[34] Xuerui Wang and Andrew McCallum. 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*. Association for Computing Machinery, New York, NY, United States, 424–433.

[35] Yu Wang, Eugene Agichtein, and Michele Benzi. 2012. TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, United States, 123–131.

[36] Zong-Ben Xu, Jiang-She Zhang, and Yiu-Wing Leung. 1998. An approximate algorithm for computing multidimensional convex hulls. *Applied mathematics and computation* 94, 2-3 (1998), 193–226.

[37] Dongjin Yu, Dengwei Xu, Dongjing Wang, and Zhiyong Ni. 2019. Hierarchical Topic Modeling of Twitter Data for Online Analytical Processing. *IEEE Access* 7 (2019), 12373–12385.

[38] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and Traditional Media using Topic Models. In *European Conference on Information Retrieval*. Springer, Dublin, Ireland, 338–349.