Using Topic Models for Twitter Hashtag Recommendation

Fréderic Godin [†]
frederic.godin@ugent.be

Viktor Slavkovikj [†]
viktor.slavkovikj@ugent.be

Wesley De Neve [†*]
wesley.deneve@ugent.be

Benjamin Schrauwen [‡]
benjamin.schrauwen@ugent.be

Rik Van de Walle [†]
rik.vandewalle@ugent.be

[†] ELIS - Multimedia Lab, Ghent University - iMinds, Ghent, Belgium

[‡] ELIS - Reservoir Lab, Ghent University, Ghent, Belgium

[*] Image and Video Systems Lab, KAIST, Daejeon, South Korea

ABSTRACT

Since the introduction of microblogging services, there has been a continuous growth of short-text social networking on the Internet. With the generation of large amounts of microposts, there is a need for effective categorization and search of the data. Twitter, one of the largest microblogging sites, allows users to make use of hashtags to categorize their posts. However, the majority of tweets do not contain tags, which hinders the quality of the search results. In this paper, we propose a novel method for unsupervised and content-based hashtag recommendation for tweets. Our approach relies on Latent Dirichlet Allocation (LDA) to model the underlying topic assignment of language classified tweets. The advantage of our approach is the use of a topic distribution to recommend general hashtags.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*; I.5.4 [**Pattern Recognition**]: Applications

Keywords

Hashtag prediction; microposts; short-text classification; topic models; Twitter.

1. INTRODUCTION

Twitter is one of the biggest and most well-known microblogging sites. With millions of active users generating microposts, there are on average 1.6 billion[1] user queries that have to be served daily. To accommodate easy search of tweets, users can make use of hashtags to categorize tweets. Despite the availability of this feature, only 8% of the tweets contain a hashtag [1]. Assigning tags to tweets requires additional effort from the user. Furthermore, Twitter users are not restricted in the way they apply the tags, since a valid hashtag is any word prepended with the hash "#" character. The free choice of tweet categories on the one hand and the lack of usable hashtags, with which tweets can be categorized on the other hand, make subsequent searches for tweets difficult. To alleviate the problem of tweet categorization, we propose a new method for automatic hashtag recommendation making use of the general topics of the tweets.

The remainder of this paper is organized as follows. In Section 2, we discuss related work and draw a comparison to our approach. The process of creating a dataset of tweets and the pre-processing of the sampled data is described in Section 3. In Section 4, we explain the details of our binary classification algorithm, used to classify between English and non-English language tweets. Section 5 deals with Latent Dirichlet Allocation for modelling the underlying mixture of topics of a set of English language tweets. We give evaluation results of the proposed approach in Section 6. In Section 7, we conclude this paper and discuss future work.

2. RELATED WORK

Despite the fact that a lot of research has been conducted to develop recommendation systems for social networks, only a few authors have addressed the problem of hashtag recommendation for easy categorization and retrieval of tweets.

Related methods for hashtag recommendation exploit the similarity between tweets. Zangerle et al. [2] compare three different approaches to recommend hashtags based on a TF-IDF representation of the tweet. They rank the hashtags based on the overall popularity of the tweet, the popularity within the most similar tweets, and the most similar tweets. The latter approach is reported to perform best on recommending five hashtags. Li and Wu [3] recommend hashtags from similar tweets by using WordNet similarity information and an Euclidean distance metric. Mazzia and Juett [4] chose to use probability distributions to recommend hashtags. In their work, they apply Bayes' rule, under the independence assumption of tweet words, to estimate the maximum a posteriori probability of each hashtag class given the words of the tweet.

As noted by Kywe et al. [1], all previous approaches do not take into account personal preferences when recommending hashtags. Therefore, the authors of [1] combine hashtags of similar users and similar tweets to propose a more personalized set of tags that would suit both user preferences and the tweet content. The TF-IDF approach is again used to construct a feature vector for each tweet. Cosine similarity is used to compare the feature vectors. It is shown that

[1] http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html

WWW, 2013 Rio de Janeiro, Brazil
ACM 978-1-4503-2038-2/13/05.

incorporating information of similar users always performs better than solely using a similar tweet metric [1].

Currently, all previous approaches rely on the similarity between individual tweets and try to recommend existing hashtags. However, the suggested tags are sparse. In our collection of 18 million tweets, 77% of the hastags were used only once and 94% were not used more than five times. Recommending and using such hashtags will not result in improved searchability and indexing. Therefore, in this paper, we propose a method for general hashtag recommendation based on the underlying topics of the tweets. Another advantage of our approach over existing approaches is that it allows for unsupervised categorization of tweets.

3. SAMPLING AND PRE-PROCESSING

In order to build a representative set of data, we made use of the Twitter streaming API[2]. The Twitter streaming API provides 1% random tweets from the total volume of tweets at a particular moment. We have sampled tweets for a total period of 4 days. In this period of time, we have collected approximately 18 million tweets.

Our pre-processing pipeline consists of several parts. We remove URLs, special HTML entities, digits, punctuations and hash characters. We also remove tweets that contain no more than one word and retweets. Because the tweet language identifier provided by Twitter is the language locale selected by the user in his profile, the language of the tweet itself may or may not be the same as the locale. To tackle this problem, we have implemented an unsupervised content-based language classifier, described in detail in Section 4. Next, words in tweets that are common acronyms and slang words are converted to proper English words, using a conversion dictionary[3]. We apply a similar dictionary containing common stop words[4] to remove words from tweets which have a high occurrence frequency, but low content discriminative value.

For the purpose of discovering topics from the content of tweets, we pre-process tweets by applying the Part-of-Speech (PoS) tagger of Gimpel et al. [5] for Twitter data. To this end, for a given tweet, we use the common adjectives and nouns discovered by the tagger to make a set of features that describe the content of the tweet.

4. UNSUPERVISED LANGUAGE CLASSIFICATION

For the purpose of hashtag recommendation, we designed and implemented a binary classifier, based on the Naive Bayes technique, which discriminates between English and non-English language tweets. Many approaches for language identification have already been proposed, but are often dependent on the type of content [6]. For each content type, manual labelling of a training set is needed when applying these supervised approaches. Therefore, we propose an approach which uses the Expectaton-Maximization (EM) algorithm to determine the parameters of Naive Bayes in an unsupervised manner. Given a dataset consisting of n observations $\mathbf{x}$, let d be the number of features in an observation. Then, from Bayes' theorem, we have the following relationship for the posterior class probability:

$$P(y \mid \mathbf{x}) = \frac{P(y)P(\mathbf{x} \mid y)}{\sum_{\forall y} P(y)P(\mathbf{x} \mid y)}, \tag{1}$$

where y denotes the class label. Under the naive assumption, each of the features is considered independent of the other, and the likelihood probability $P(\mathbf{x} \mid y) = P(x_1, \ldots, x_d \mid y)$ decouples into $P(\mathbf{x} \mid y) = \prod_{i=1}^{d} P(x_i \mid y)$.

In order to train a Naive Bayes classifier, the prior probabilities of each class $P(Y = y)$, and the conditional probabilites of the features given the classes $P(X = \mathbf{x} \mid Y = y)$ have to be available. These are usually estimated with the maximum a posteriori (MAP) estimation from the data in the labelled training set. In order to produce good classification results, a large amount of labelled training would be needed. Manually labelling tweets is time consuming and error prone. Instead, we adopt an unsupervised approach based on the Expectation-Maximization (EM) algorithm [7].

Algorithm 1: Unsupervised language classifier based on Naive Bayes and EM.

```
Input: $\mathbf{x}^{(j)} = (x_1^{(j)}, \ldots, x_d^{(j)}), \quad j = 1, \ldots, n.$
Output: model parameters $\theta : \theta_Y = P(Y = y); \quad \theta_X = P(X = x \mid Y = y).$
begin
    1.  Initialize $\theta_Y,\ \theta_X$
    2.  E-step:
        for $\mathbf{x}^{(j)} \in \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ do
            $P(Y = y \mid \mathbf{x}^{(j)}) = \frac{P(y)P(\mathbf{x}^{(j)}|y)}{\sum_{\forall y} P(y)P(\mathbf{x}^{(j)}|y)}$
    3.  M-step:
        $\theta_Y = \frac{\sum_j^n P(Y=y|\mathbf{x}^{(j)})}{n}$
        $\theta_X = \frac{\gamma+\sum_j^n \delta(X,\, x_i^{(j)})P(Y=y|\mathbf{x}^{(j)})}{\gamma\left|V_y\right|+\sum_j^n P(Y=y|\mathbf{x}^{(j)})}$
```

The goal of the EM algorithm is to find the maximum likelihood or the MAP estimate for the model parameters that depend on latent variables. In this context, we treat the class labels for the tweets in the training set as latent variables, and use character n-grams as features for the Naive Bayes model. The model parameters that need to be learned are then the prior distribution of classes, and the conditional probabilities of the character n-grams given the classes. We use a Dirichlet distribution as the prior class probabilities distribution, and the Beta distribution for the conditional probabilities of the character n-grams given the classes. In the expectation step of the algorithm, we calculate the posterior class probabilities given the observations and the current model parameters. In the maximization step, we determine the new model parameters that maximize the expectation. The pseudocode of the algorithm can be seen in Algorithm 1. To eliminate zero probabilities on unseen character n-grams, we use a smoothing parameter γ in the maximization step, where $|V_y|$ is the size of the vocabulary.

[2] https://dev.twitter.com/docs/streaming-apis

[3] http://www.noslang.com/dictionary/

[4] http://www.d.umn.edu/~tpederse/Group01/WordNet/words.txt

5. LATENT DIRICHLET ALLOCATION FOR DISCOVERING HIDDEN TOPICS

Latent Dirichlet Allocation (LDA) is a hidden topic model. It is often used to discover the general topics in large document collections for the purpose of efficient information retrieval. The hidden or latent topics associated with a document form a summary of the document and are sufficient for efficiently retrieving matching documents.

5.1 LDA using Gibbs sampling

Latent Dirichlet Allocation is a generative model which assumes that underlying the data collection, there exists a topic model with T topics. Each document m, containing N_m words, has an associated multinomial topic distribution $\vartheta_{\mathbf{m}}$ over these T topics. From this distribution, a topic $z_{m,n}$ can be determined for each word $w_{m,n}$ of the document. Next, a word $w_{m,n}$ for topic $z_{m,n}$ can be sampled from the topic word distribution $\phi_{z_{m,n}}$. Both ϑ and ϕ are Dirichlet distributions with hyperparameters $\vec{\alpha}$ and $\vec{\beta}$, respectively.

As mentioned before, we will use LDA for hashtag recommendation. In our case, the data collection is a collection of tweets. A document corresponds to one tweet and the words of the document are PoS-tagged common nouns and adjectives of the tweet (see Section 3).

To determine the model parameters, we made use of the collapsed Gibbs sampling algorithm. Gibbs sampling is a technique used to rapidly explore the space around a target distribution using repeated sampling. A topic z_i of word w_i, conditioned on the used words $\vec{w}$ of the model and the topic-word distribution $\vec{z}_{\neg i}$, can be predicted as:

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) \propto \frac{n^{(t)}_{k,\neg i} + \beta_t}{\sum\limits_{t=1}^{V} [n^{(t)}_{k,\neg i} + \beta_t]} * \frac{n^{(k)}_{m,\neg i} + \alpha_k}{\sum\limits_{k=1}^{K} [n^{(k)}_{m,\neg i} + \alpha_k]}, \quad (2)$$

where $n^{(t)}_k$ denotes the topic-word count of topic k and word t, and $n^{(k)}_m$ the tweet-topic distribution. For an elaborate explanation and derivation of the formulas, we refer to [8].

5.2 Calculating the topic distribution of new tweets

To determine the topics of a new tweet, we can again make use of the collapsed Gibbs sampling algorithm. This time we start from a single tweet $\tilde{m}$. The topic-word count distribution was calculated during training and is part of the model $\mathcal{M}$. The conditional distribution $p(z_i = k|\vec{z}_{\neg i}, \vec{w})$ is now equal to:

$$p(\tilde{z}_i = k|\tilde{\vec{z}}_{\neg i}, \tilde{\vec{w}}, \mathcal{M}) \propto \frac{n^{(t)}_{k,i} + \tilde{n}^{(t)}_{k,\neg i} + \beta_t}{\sum\limits_{t=1}^{V} [n^{(t)}_{k,i} + \tilde{n}^{(t)}_{k,\neg i} + \beta_t]} * \frac{n^{(k)}_{\tilde{m},\neg i} + \alpha_k}{\sum\limits_{k=1}^{K} [n^{(k)}_{\tilde{m},\neg i} + \alpha_k]}. \quad (3)$$

The output of the algorithm will now be the topic distribution of the new tweet.

5.3 Selecting keywords for hashtag recommendation

After we have determined the topic distribution of a tweet m, we can select keywords from it. We make use of the topic-term count values $n^{(t)}_k$ to determine the top words of every topic. Based on the number of hashtags we want to recommend, we sample the topic distribution of m for a topic and select a top word from that topic in ranked order. The final result is a set of keywords that resemble the general topic of the tweet.

6. RESULTS

6.1 Unsupervised language classification

In order to train the language classifier described in Section 4, we constructed a training set containing 1.8 million pre-processed tweets. Since the EM algorithm guarantees convergence only to a local maximum, the training result depends on the initial starting conditions. To counter the convergence to local maxima, we trained multiple versions of the classifier with different starting parameters. We used a separate validation set of 1000 randomly selected, labeled tweets, containing 50% positive and 50% negative examples, and picked the classifier that gave the best results on the validation set. We used uninformative priors for the Dirichlet and the Beta distributions. We tested the final effectiveness of the classifier on a different test set, which contains 1000 randomly selected, labeled tweets, with 50% English language and 50% non-English language tweets. From the samples in the test set, 489 were classified as true positives, 485 as true negatives, 15 as false positives, and 11 as false negatives. This results in 97% precision and 97.8% recall. The accuracy of the language classifier is 97.4%. For the test, we used character tri-grams as features since they gave better effectiveness when compared with bi-grams.

6.2 LDA hashtag recommendation model

After filtering out non-English tweets, the remaining tweet collection still contains a lot of different tweets which makes the evaluation challenging. Therefore, we make use of a list of Wikipedia keywords[5] and select the first 4000 words. After filtering on those words, 1.8 million tweets were left. We selected 100 random tweets for testing from the original set of 1.8 million tweets. We trained the LDA model with parameters $\alpha_i = 0.1$, $\beta_i = 0.1$ and $T = 200$. The number of topics is a trade-off between a too general model that has a few topics and a very specific model with many topics that needs a lot of training data. We tried values for T of 50, 100, 200, 300 and 500 topics. Increasing α_i and β_i did not yield better results because the link between co-occurring words within a tweet and the topic-term distribution loosens (see Equation 2). After training, the learned topic model was used to suggest a number of hashtags per tweet that capture the general topic of the tweet. Because evaluating recommendations is a subjective task, 2 persons were asked to independently evaluate whether the suggested word described the topic of the tweet and could possibly be used as a hashtag. Suggestions for which evaluators did not agree on were discussed until a consensus was reached. Evaluation results are depicted in Figure 6.2. A number of examples can be found in Table 1.

For 80% of the tweets, at least one suitable hashtag could be suggested out of five recommended hashtags. The accuracy increases to 91% when ten hashtags could be suggested.

[5] `http://gibbslda.sourceforge.net/wikipedia-topics.txt`

Table 1: Example tweets and corresponding suggested hashtags. Suitable hastags are highlighted.

Tweet Text	Suggested Hashtags
yay , we got sixth period today x	school , business, light, time , period
mixed emotions right now.. #hedeservedthis	head , love , song, positive , negative
please rt!! sign bernie sanders petition for the fiscal cliff! http://..	fiscal , political , traffic, president , policy
left my trunk open for two days..	life, people, car , mount, story
new to the library are 8000 ebooks via oxford scholarship online: http://..	room, business, people, university , hotel
comfort, elegance, prettiness.	little, good, love, relationship, god
red and pink the little mephisto - artisoo	case, blue, little, people, many

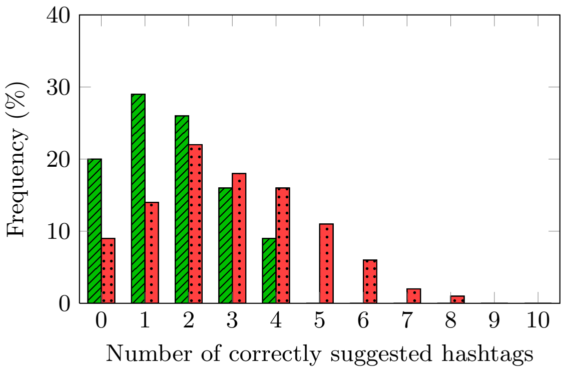

Figure 1: Histogram of the number of correctly suggested hashtags per tweet when five (lines) or ten (dots) hashtags are suggested.

When no hashtag could be suggested, this was often due to the lack of context which yielded an equalized topic distribution. Even for humans, the meaning of the tweet was not clear. Another reason is that some words and topics are sparse in the tweet collection which makes it difficult to deduce a peaked topic distribution and to suggest multiple suitable hashtags. For some cases the tweets simply did not have enough content words. In 51% of the cases, at least two out of five hastags were found suitable.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a new approach to recommending hashtags for tweets. The proposed method provides for easy indexing and search of tweets. The contributions of our work are twofold. We designed and developed a binary language classifier for tweets based on the Naive Bayes method and Expectation-Maximization. Next, we applied a Latent Dirichlet Allocation (LDA) model in the context of tweet hashtag recommendation. The LDA model was trained to cluster English language tweets in a number of topics from which keywords can be suggested for new tweets.

The advantage of our approach is that it can recommend hashtags for tweets in a fully unsupervised manor. This makes the approach easily portable to other sets with different content or much bigger sets. Another advantage over existing work is the suggestion of general hashtags instead of sparse existing hashtags which enables effective categorization and search of tweets.

To be able to suggest more suitable hashtags in the future, several approaches could be taken. A first step would be to disambiguate tweets by using more semantic knowledge. This could be done for example by using the semantic web to enrich the tweet. Another approach would be to incorporate information about the user based on previous tweets to allow for a more personalized approach. Finally, we could make use of the user feedback to improve the topic model.

8. ACKNOWLEDGMENTS

The research activities described in this paper were funded by Ghent University, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

9. REFERENCES

[1] S. M. Kywe, T.-A. Hoang, E.-P. Lim, and F. Zhu, “On Recommending Hashtags in Twitter Networks,” in *The 4th Int. Conference on Social Informatics*, 2012.

[2] E. Zangerle, W. Gassler, and G. Specht, “Recommending#-tags in twitter,” in *Proceedings of the Workshop on Semantic Adaptive Social Web*, 2011.

[3] T. Li, Y. Wu, and Y. Zhang, “Twitter hash tag prediction algorithm,” in *ICOMP’11 - The 2011 International Conference on Internet Computing*, 2011.

[4] A. Mazzia and J. Juett, “Suggesting Hashtags on Twitter,” tech. rep., Computer Science and Engineering, University of Michigan, 2009.

[5] K. Gimpel, N. Schneider, B. O’Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, “Part-of-speech tagging for twitter: annotation, features, and experiments,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT ’11, (Stroudsburg, PA, USA), pp. 42–47, Association for Computational Linguistics, 2011.

[6] T. Baldwin and M. Lui, “Language identification: the long and the short of the matter,” in *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[7] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977.

[8] G. Heinrich, “Parameter estimation for text analysis,” tech. rep., 2004.