SemTagger: A Novel Approach for Semantic Similarity Based Hashtag Recommendation on Twitter

Kuntal Dey[†], **Ritvik Shrivastava**[⋆], **Saroj Kaushik**[§], **L Venkata Subramaniam**[†]
[†]IBM Research, India; [⋆]NSIT Delhi, India; [§]IIT Delhi, India
[†]{kuntadey, lvsubram}@in.ibm.com; [⋆]ritviks.it@nsit.net.in; [§]saroj@cse.iitd.ac.in

Abstract

This paper proposes a semantic similarity based novel approach, to assign or recommend a hashtag to a given tweet. The work uses a Latent Dirichlet Allocation (LDA) based learning approach. In the training phase, we learn the latent concept space of a given set of training tweets, via topic modeling, and identify a group of tweets that act as representatives of the topic. In the inference phase, we create a probability distribution of a given test tweet belonging to the learned topics, and find the semantic similarity of the test tweet with representative tweets for each topic. We propose two assignment approaches. In one approach, we assign hashtags to a target tweet, by obtaining these from a set of representative training tweets, that have the highest semantic similarities with the target tweet. In the other approach, we combine (a) the semantic similarity of the target tweet with the representative tweets, and (b) the assignment probability of the target tweet to a given topic, and assign hashtags using this joint maximization. The hashtags are assigned to the target tweet, by selecting the top-K values from the combination. Our system yields F-score of 46.59%, improving over the LDA baseline by around 6 times.

1 Introduction

1.1 Background and Motivation

The hashtag recommendation problem for Twitter addresses suggesting appropriate hashtags to a user for assigning to a tweet they would post. Recommendation of hashtags for Twitter messages has emerged as a mainstream area of research. Practically, only around 10-15% Twitter data tends to have hashtags, as observed by (Hong et al., 2011). And yet, as observed in the literature, hashtags play a critical role in solving significant problems, e.g., information diffusion (Starbird and Palen, 2012) (Tsur and Rappoport, 2012), topic modeling (Asur et al., 2011) and many other problems as observed by the literature survey conducted by Dey et al. (2017). All of the above indicate that it is important to solve the problem of hashtag recommendation.

The problem has been received with strong research enthusiasm in recent times. Several research solutions have been proposed. Some early-breaking works include the works by Zangerle et al. (2011), Ding et al. (2012) and Ding et al. (2013), that follow approaches such as *tf-idf* and translational models. Several other approaches emerged over time. Topical models, such as Zhang et al. (2014) and Gong et al. (2015), started finding way into the literature. Deeper and more focused methods started getting proposed, such as recommending hashtags for tweets containing a hyperlink by (Sedhai and Sun, 2014). Subsequently, deep neural network based models emerged. Weston et al. (2014) predicted hashtags using a convolutional neural network (CNN) (Krizhevsky et al., 2012) based approach, and learned semantic embeddings of hashtags. Gong and Zhang (2016) used CNN with attention mapping. They attained an F-score of 39.8%, which is the best in the literature till date.

1.2 Central Idea

We observe that, while Dirichlet and specifically Latent Dirichlet Allocation (LDA) (Blei et al., 2003) based approaches exist in the literature to solve the problem at hand, these works tend to model the topics appearing in a given target tweet as a semantic (topical) alignment with the training tweets, and use the hashtags appearing in those tweets for recommendation. An important aspect

Proc. of ICON-2017, Kolkata, India. December 2017 ©*2017 NLPAI*, pages 178–187

that appears unexplored is the semantic similarity of the target tweet, with the training tweets that are topically aligned. In the current work, we hypothesize that, considering the semantic similarity of the training tweets that are topically (LDA-wise) based aligned to the target tweet, and assigning hashtags to the target tweet using this similarity, is an effective methodology for recommending hashtags to tweets.

In the training phase, we use a LDA-based topic modeling, to learn the semantic concept space covered by the training tweets, and identify topics via topic modeling. We identify a group of tweets that act as representatives of the topic. For inference (assigning hashtags to a given target tweet), we create a probability distribution of the target tweet belonging to the learned topics. We subsequently find the semantic similarity of the target tweet with representative tweets for each topic, using a state-of-the-art model externally learned specifically for Twitter (Dey et al., 2016).

We propose two variants for making the recommendation. In one variant, we recommend hashtags to the target tweet, using the semantic similarity of the target tweet with the representative tweets for each topic derived, and picking from the more similar training tweets. In the other variant, we combine (a) the semantic similarity of a target tweet with the representative tweets for each topic derived, and, (b) the assignment probability of the target tweet to a given topic, to obtain a combined score of each representative tweet (across the different topics) to get selected. We rank the representative tweets based on the score of combination, and recommend hashtags based upon the hashtags observed in the top-K ranked tweets. We empirically determine K as 3, and observe that our methodology produces highly effective results, lifting the F-score by around 6 times from the LDA baseline.

1.3 Our Contributions

The contributions of our work are the following.

- We provide a novel methodology to address the problem of hashtag recommendation on Twitter. Our approach replies upon recommending hashtags to a given target tweet, based on semantic similarity of the target tweet with topically similar training tweets.
- We propose *SemTagger*, a framework where we learn the latent concept space of a given set of training tweets, via topic modeling, and assign hashtags to test tweets using (a) a combination of the semantic similarity of a test tweet with representative training tweets, and the assignment probability of the test tweet to a given topic, and (b) assigning hashtags by selecting the top-K values from the combination thus computed.
- We empirically determine the effectiveness of the proposed approach. In our experiments, we observe that our methodology delivers an F-score of 46.59%, which is around 6 times higher compared to a corresponding LDA baseline of 7.79%.

The rest of the paper is as follows. Section 2 provides an overview of the literature in the space of Twitter hashtag recommendation. This is followed by the details of our methodology in Section 3. Section 4 presents the experiment design and results. Section 5 is used for a brief discussion of a few aspects of interest. The paper is finally concluded in Section 6.

2 Related Work

Hashtag recommendation has been established as a well-accepted research problem for nearly a decade now. Multiple approaches have been proposed by researchers exploring the problem from several aspects. In an early work, while solving a sentiment classification problem, Davidov et al. (2010) had attempted to address hashtags indicative of sentiments. However, the first-ever work that focused completely on hashtag recommendation, was carried out a year later, by Zangerle et al. (2011). In this work, the authors used the *tf-idf* approach to compare tweet-pair similarity, and thus computed the similarity of a target tweet with given training tweets. They subsequently retrieved tweets with the most similar messages, and heuristically ranked and recommended the hashtags that appeared in the extracted tweets. In a body of works that followed, Ding et al. (2012) and Ding et al. (2013) converted the hashtag recommendation to a translation problem. Their model is centered around an unsupervised learning method using a latent variable estimation based topical translation model. They hypothesize that hashtags and trigger words of tweets are two different languages with the same meaning that occur in parallel. They use "topic-specific word trigger to bridge the vocabu-

lary gap between the words in tweets and hashtags, and discovers the topics of tweets by a topic model designed for microblogs".

Subsequently, a large number of research works started emerging in the literature, that attempted to solve the problem. Several novel approaches were proposed, covering different aspects of the problem. One such work, that attempted to recommend hashtags only to the tweets containing a hyperlink in the content, was proposed by Sedhai and Sun (2014). Their approach consisted of two phases. In the first phase, they selected a set of candidate hashtags using the attributes computed from tweet content, such as hyperlinked documents, named entities contained in the referred webpage as well as present in the tweet, and the domain of the content of the webpage that the hyperlink refers to. In the second phase, they formulate as a learning-to-rank problem, and solve with RankSVM to aggregate and rank the candidate hashtags selected in the first phase.

Gong et al. (2015) proposed a Dirichlet based method. They adopted a Dirichlet based mixture model, incorporating types of hashtags as hidden variables. Motivated by Liu et al. (2012) and philosophically akin to Ding et al. (2012) and Ding et al. (2013), they also model assuming that hashtags and tweet content are parallel descriptions of the same content.

A topic-based hashtag recommendation method was proposed by She and Chen (2014). This work treated hashtags as topic labels, and performed supervised topic model learning over these labels, to discover inter-word relationships. They treated the words as one of two types: background words that are prevalent in many of the tweets, and local topic words that are more specific to that topic. They inferred the probability that a hashtag will be contained in a new tweet, and generated hashtags for recommendation using a symmetric Dirichlet distribution of the local and background words. Zhang et al. (2014) proposed another topic-based hashtag recommendation method. Their work used a topical model based method, incorporating both temporal and personal information. They extended over the well-established translational model for hashtag recommendation. They divided the time horizon into T epochs, and analyzed at a per-epoch level to ensure temporal relevance of recommended hashtags. They drew from a multinomial word-topic distribution and recommended the hashtags that have the maximum probabilities in the draw. Among other works, Godin et al. (2013) too proposed another effective topic-based hashtag recommendation method.

The recent advances in deep neural network based learning (deep learning), has motivated researchers to attempt such techniques on the hashtag recommendation problem. In an early application of deep learning on this problem, Weston et al. (2014) predicted hashtags using a convolutional neural network (CNN), and learned semantic embeddings with hashtags. They posed as a supervised learning problem, treating the hashtags as labels assigned to the tweet content. Their model represents the words, as well as the entire textual posts, as embeddings in the intermediate layers of their deep-CNN architecture. The recent work by Gong and Zhang (2016) used CNN with attention mapping. They, too, converted the words into embeddings, and used a local small window based attention map, where each given window surrounds a word around which the attention is provided. They attained an F-score of 39.8%, lifting the performance over a LDA baseline by 6.42 times, making the work the most effective hashtag recommendation system known in the literature.

Our work uses the LDA-based models, but introduces a novel mechanism of augmenting topical similarity with semantic similarity of target and training (known) tweets. This approach is the first of its kind, and it outperforms the systems known in the literature except the work by Gong and Zhang (2016). However, the practicability of deep learning in real-life systems that are often used from mobile phones, remains a question till date. Deep learning on mobile phones has remained a challenge [1] that has not been addressed till date in a satisfactory manner. And yet, 85% of the total usage time on Twitter happens on mobile phones[2]. Our approach is lightweight, making it practical and useful in real life, including being usable from mobile phones. Thus, while in terms of performance (F-score) metrics our model is second to a deep-learning based model (Gong and Zhang, 2016), practically, not counting the deep learning systems that are not fit for use in real-life solutions that often are executed on mobile phones, our work establishes a new real-life benchmark.

[1] https://conferences.oreilly.com/strata/strata-ca-2017/public/schedule/detail/56179

[2] https://twitter.com/wsjtech/status/451886622788055040?lang=en

3 Details of Our Approach

We use a topic modeling and semantic similarity driven approach to model our solution framework. The details of *SemTagger*, our framework, are presented below.

3.1 Data Cleaning

The very first step followed in the training as well as inference phases, is data cleaning. This comprises of the following steps.

- **Removal of tweets without any hashtag:** In order to train our model, we need tweets that necessarily contain hashtags. Further, since the objective of the present work is to perform hashtag recommendation, the target (test) tweets that we shall assign hashtags to, will also need to contain ground-truth hashtags assigned by the user posting the tweet. The testing will be performed by hiding the hashtags from the target tweets and assigning the predicted hashtags to these tweets using our model; however, the performance of our model will be validated by the ground-truth hashtags that were hidden. Thus, all the tweets we use for our process necessarily need to contain at least one hashtag. Driven by this requirement, we retain only those tweets that contain at least one hashtag, and eliminate the remaining tweets.

- **Non-English tweet removal:** Since the focus of our work is around tweets authored in the English language, we eliminate the non-English tweets from our dataset. The language-marker field present in the raw Twitter data indicates the language of each given tweet, which is used to detect whether a given tweet is in English or not. This frees our dataset from extraneous and non-useful tweets, and retains only the English tweets that are of interest.

- **Non-ASCII character removal:** Since the non-ASCII characters do not add value to the work, we eliminate the non-ASCII content present in each given tweet (that has been retained otherwise), and retain the remaining part of the text.

After the data cleaning process, we are left with only English tweets, with at least one hashtag, and containing no non-ASCII character.

3.2 Preprocessing

Both in the training and testing phases, we first preprocess the dataset. This includes performing the following operations on each tweet:

1. **Tweet normalization:** We normalize tweet content, by resolving many colloquial on-the-net expressions appearing as part of user-generated social media text, but do not appear in any traditional dictionary. For instance, what appears as *aaf* on Twitter, is expanded to *as a friend* after the tweet normalization process. We normalize the tweets using a net slang dictionary[3] and Han-Baldwin normalization corpus.

2. **Stopword removal:** Stopword removal is an essential step of our process. This step ensures that the superfluous words with practically no information content for the task under consideration are eliminated (such as prepositions, article *etc.*). We perform stopword removal using an online dictionary[4].

The architecture of the data cleaning and preprocessing phases are given in Figure 1.

3.3 Topic Model-Based Training

We perform topic model-based training from the given tweets, to construct a topic distribution model. We subsequently identify a representative set of tweets for each of the topics detected. The training pipeline has been illustrated in Figure 2.

3.3.1 LDA-Based Topic Modeling

We perform LDA-based topic modeling on the training tweet set. This is performed over two steps.

First, a document is created as a concatenation of all the tweets present in the training dataset, minus the hashtags. That is, for a given set of tweets $T = \{t_1, t_2, ..., t_n\}$, containing hashtags $H = \{h_1, h_2, ..., h_m\}$, a document D is constructed as

$$D = \bigcup_{i=1}^{n} t_i - \bigcup_{j=1}^{m} h_j \quad (1)$$

Next, the document is processed for LDA-based topic modeling, and a set of topics $Z =$

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

[4] https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

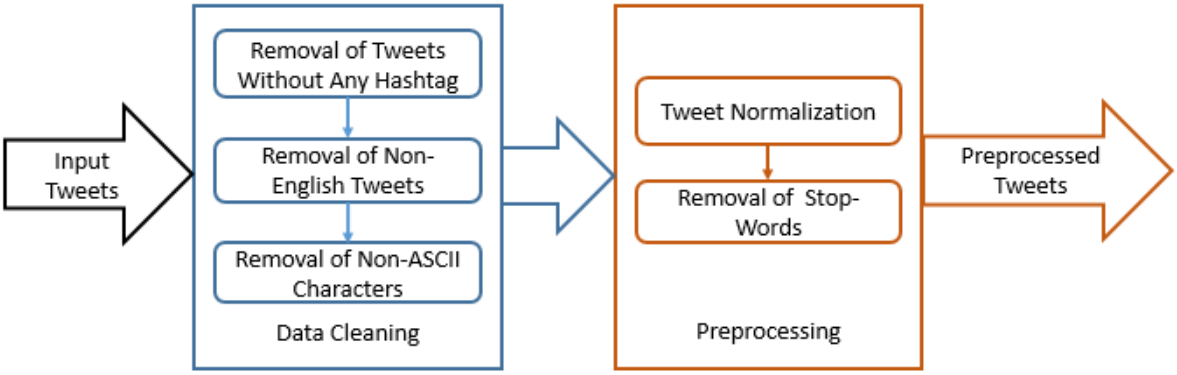

Figure 1: Data Cleaning and Preprocessing

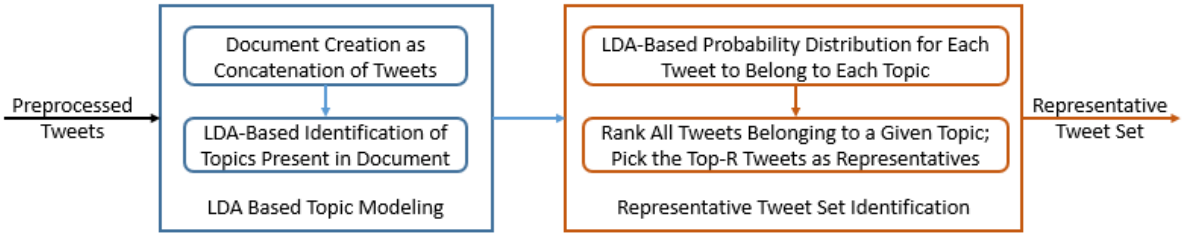

Figure 2: Training and Representative Tweet Set Identification

$\{z_1, z_2, ..., z_l\}$ are learned. Please note that, LDA (Blei et al., 2003) is traditionally modeled as a joint distribution in the following manner:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D} = \prod_{i=1}^{K} p(\beta_i).\prod_{d=1}^{D} p(\theta_d).$$
$$\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K}, z_{d,n}) \quad (2)$$

Here, $\beta_{1:K}$ represent the topics where each β_k is a distribution over the given vocabulary, θ_d are the topic proportions for document d, $\theta_{d,k}$ is the topic proportion for topic k in document d, z_d are the topic assignments for document d, $z_{d,n}$ is the topic assignment for word n in document d, and w_d are the observed words for document d. This process learns the semantic concept space of the training tweets, in form of latent topics.

3.3.2 Representative Tweet-Set Identification

We identify a set of tweets that act as representative tweets for each identified topic. For this, we generate the probability distribution of each tweet to belong to each topic derived, using LDA on the tweet content. For each topic, we rank the tweets by the probability value that a tweet belongs to the topic. We finally pick all the tweets ranked within the top R, to form a representative tweet set of size R for that topic. The output of the training process constitutes of a set of topics $Z = \{z_1, z_2, ..., z_l\}$, a set of representative tweets $T_{z,L} = \forall(l \in L)\{t_{z_l}\} = \forall(l \in L)\{t_{1,l}, t_{2,l}, ..., t_{n,l}\}$ associated with each topic.

3.4 The Hashtag Recommendation Methodology

After topic training and representative tweet set identification, the system becomes capable of assigning hashtags to target (test) tweets provided as input. For this, we first create a probability distribution of a given test tweet belonging to the learned topics. This, again, is performed by generating the LDA-based probability distribution of the tweet content, that quantifies "how much" a tweet belongs to each topic. Using this baseline, we propose a few variants (heuristics) based upon semantic similarity detection, to perform hashtag assignment to each given test tweet. We broadly categorize these approaches in two categories: *semantic similarity based* and *joint probability maximization based* hashtag recommendations.

3.4.1 Semantic Similarity Based

The first method we propose is a semantic similarity rank based hashtag recommendation. Figure 3

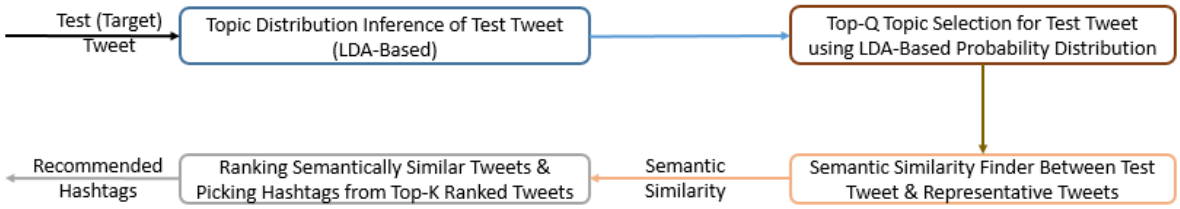

Figure 3: Semantic Similarity Based Hashtag Recommendation

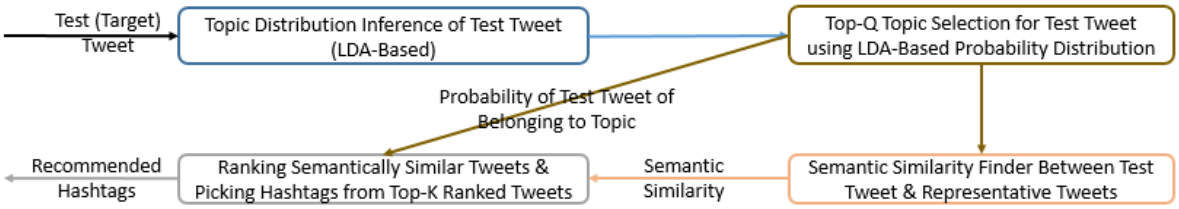

Figure 4: Joint Distribution Maximization Based Hashtag Recommendation

provides a block-level illustration of this method. In this approach, we first select the best (highest probability) Q topics out of Z, using the probability distribution of the given test tweet. We measure the semantic similarity between the test tweet and all the R representative tweets T_{Z_q} across all the top Q topics Z_q. For measuring semantic similarity, we use a transfer learning approach: we use an external semantic similarity learning model given by Dey et al. (2016), which was specifically trained for semantic similarity quantification on Twitter. We rank all the $R*Q$ representative tweets by their semantic similarity scores with the test tweet, and select the hashtags given by the top-K ranked tweets, where K is an externally specified integer.

3.4.2 Joint Distribution Maximization Based

The second method we propose uses the semantic similarity based model as the baseline; however, unlike the earlier approach which was topic distribution-agnostic for ranking the semantically similar tweets, this is topic distribution-aware. Figure 4 provides a block-level illustration of this method. Here, we maximize the combination of (a) the semantic similarity of the test tweet with the representative tweets of a topic, and, (b) the assignment probability of the test tweet to the topic. If the assignment probability of a test (target) tweet t to a topic z_l is $P(t, z_l)$, and the semantic similarity of a test tweet t with one given representative tweet t_j of a topic is $SS(t, t_j)$, then, the combined score for each <test tweet, representative tweet> pair is:

$$CS(t, t_j) = P(t, z_l) \times SS(t, t_j) \quad (3)$$

We rank the $CS(t, t_j)$ values thus obtained, and pick the top-K tweets based upon this rank to select hashtags for the task of recommendation. Thus, in this case, the semantic similarity values of the representative tweets with the test tweet, are not ranked directly; instead, first, the semantic similarity values are combined (multiplied) with the probability of the test tweet belonging to that topic, and then, this combination (product value) is ranked. We assign the hashtags by selecting the top-K tweets in a decreasing (ranked) order of product values, thus inherently selecting the maximal values from the combined distribution.

The overall process that we follow, is given in Algorithm 1.

4 Experiments

We present the details of the experiments conducted and results obtained below.

4.1 Data Description and Tools Used

Using Decahose[5], we collect 10% random sample of all the tweets made on Twitter for 31^{st} January, 2016, and retain all the English tweets that have at least one hashtag associated. We remove the retweets and quoted tweets from both the training and test tweets, as it is trivial to assign hashtags to such tweets, given the actual or recommended hashtags to the corresponding original tweets. We

[5]https://gnip.com/realtime/decahose/

clean the data to remove all hashtags that are simple stopwords [6], and remove the tweets that comprise of only such hashtags (if a tweet has other hashtags too, we retain it). Further, we empirically retain all the tweets that use at least one hashtag which has been used between 200-500 times in the original dataset. This produces a set of $251,649$ English tweets with at least one hashtag. We randomly split into three sets: $175,000$ for training, $25,000$ for validation and the remaining $51,649$ for testing. We evaluate the effectiveness of our system by comparing the recommended hashtags with the actual hashtags present in the test tweets. The dataset details are presented in Table 1.

Tweet Selection Criteria	Count
Total tweets	$34,114,982$
English tweets	$13,410,808$
Tweets with at least one hashtag	$2,417,163$
Hashtag count based retention	$251,649$
Training tweets	$175,000$
Validation tweets	$25,000$
Testing tweets	$51,649$

Table 1: Data description

We use the Stanford NLP Toolkit (Manning et al., 2014) for PoS tagging, Porter stemmer (Porter, 2001) for stemming the tweets, MALLET (McCallum, 2002) for training the LDA based topic models, and Weka (Hall et al., 2009) for running the semantic similarity model.

4.2 Experimental Results

To evaluate the performance of our system, we use precision (Pr), recall (Re) and F-score (F1), computed as $Pr = \frac{N_c}{N_s}$, $Re = \frac{N_c}{N_t}$ and $F1 = \frac{2.Pr.Re}{Pr+Re}$, where N_c and N_s are the correct and total number of hashtags recommended for a given tweet respectively, and N_t is the total number of hashtags present in the semantically similar training tweets under consideration. In an embodiment of our methodology where the number of hashtags to be predicted in the test tweet is provided as an input, we perform experiments by limiting our system to predict the required number of hashtags. We empirically choose the size of the representative tweet set $R = 100$; as well as, we empirically pick the top $Q = 3$ topics that a test tweet is aligned to.

[6] https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

4.2.1 Selecting K

Experiment	F1(%)	Experiment	F1(%)
Top 1	36.67	Top 2	42.93
Top 3	**46.59**	Top 4	36.52

Table 2: Selecting the value of K using F-scores

Next, we select K, the number of representative tweets to consider for computing semantic similarity with the test tweet. In order to select an effective value of K, we vary the value of K from 1 to higher values, and observe the impact of the values on the final F-score that our system produces. Specifically, we use the semantic similarity match based methodology described in Section 3.4.1. As clear from Table 2, the impact of considering a larger number of semantically similar representative tweet for comparison with a test tweet, is the most effective for $K = 3$. Hence, we choose the value of $K = 3$ for the subsequent experiments.

4.2.2 Joint Distribution Maximization

We compute the combination of the semantic similarity of a test tweet with the given representative tweets of the topics, and the probability of the representative tweets to belong to the respective topics, using Equation 3. The scores are ranked, and we pick the tweets that are ranked in the top-K to select hashtags for the task of recommendation. We empirically observe K=3 to deliver the best performance, wherein, the F1-score is **46.28%**, precision 34.33% and recall 70.99%.

4.2.3 At-Least-One vs. Multiple Correct Predictions

One way to evaluate the effectiveness of our approach is to ask the following questions.

– How well does our methodology predict at least one hashtag correctly? This is answered by examining whether there is any overlap between the recommended hashtags for the tweet and the ground truth (actual hashtags seen in the tweet). In the joint distribution maximization based recommendation approach, we observe at least one hashtag to have been recommended (predicted) correctly in 66.74% cases.

– How well does our methodology predict more than one hashtag correctly? This is answered by examining whether at least two (or more) hashtags overlap, between the recommended hashtags for the tweet and the ground truth (actual hashtags

seen in the tweet). In the joint distribution maximization based recommendation approach, we observe two or more hashtags to have been recommended (predicted) correctly in 42.24% cases.

4.2.4 Comparison with Other Works

In absence of benchmark datasets for comparison, we create a LDA-based baseline score. For this, akin to the rest of our approach, we pick the top 3 topics that the test tweet is aligned to. For each topic, we pick the one representative tweet that has the highest likelihood of belonging to that topic (amongst all the tweets that represent the topic). We perform hashtag assignment to the test topic, using the 3 training tweets thus selected across the 3 topics. The LDA baseline gives 7.79% F1-score. Since our system yields a best-case F1 performance of **46.59%** (with the semantic similarity based approach), the lift we obtain over the LDA baseline is $46.59/7.79 \approx$ **6**, which is large.

Method	Lift
Naive Bayes	3.27
IBM1 (Liu et al., 2011)	3.55
TopicWA (Ding et al., 2012)	4.71
TTM (Ding et al., 2013)	5.87
SemTagger (Joint Maximization)	**5.94**
SemTagger (Semantic Similarity)	**6**
CNN+Att.-5 (Gong and Zhang, 2016)	6.42

Table 3: Lifts over the baseline, across methods

Further, we observe that, our model (F-score 46.59) yields an F-score higher than the literature (39.8). However, in absence of benchmark data, we compare our work with the literature using the **lift** over the baseline LDA values. Table 3 captures these values. Clearly, the lift obtained by our work is comparable to Gong and Zhang (2016), and it consistently outperforms the rest of the literature.

5 Discussion

We discuss a few interesting observations below.

5.1 Significance of Using Lift as a Measure

No standard dataset has been made available yet for the task of hashtag recommendation. Further, many of the existing literature have not released codes, and reimplementation of these codes are always prone to errors. We note that, the methodology that has acted as the benchmark of baseline, is the LDA-based approach. Given these observations, we use the list obtained by the model over baseline LDA, as the approach for validation. Here, the well-known LDA baseline is implemented. Subsequently, a ratio of the performance (F-score) obtained by our system, is compared with that obtained by the baseline LDA implementation. Further, comparing the performance of our system with other works in the literature, becomes meaningful and error-free by this comparison mechanism, in spite of absence of benchmark data as well as released codes for the task of Twitter hashtag recommendation.

5.2 General Observations

Our model is highly novel, and the lift we obtain (lift ≈ 6) is comparable to the state-of-the-art (Gong and Zhang, 2016), and it outperforms all other works available in the literature. Further, our model is lightweight and robust, as opposed to the computationally expensive deep-learning approach of the state-of-the-art. This makes our work useful and effective in real-life applications. We also note that the difference of performance between the two models we proposed - the *semantic similarity based* model and the *joint optimization based model* - is not much, though, the former model performs marginally better compared to the later for the current dataset.

6 Conclusion

In this paper, we proposed a novel hashtag recommendation approach for tweets, based on semantic similarity. We used LDA-based topic model training. For assigning hashtags to a target tweet, we proposed two variants. In one variant, hashtags are assigned to a target tweet, such that, the hashtags are obtained from a set of representative training tweets having the highest semantic similarities with the target tweet. In the other variant, we assigned hashtags to target tweets using (a) a maximization function that combines the probability of a given target tweet belonging to a topic, and the semantic similarity of representative training tweets that belong to that topic, and (b) assigning hashtags observed in the top-K ranked tweets in the maximized combination. Empirically, our model produced a major lift of 6 times over the LDA baseline. Our approach is robust and lightweight, and usable in real-life settings. *SemTagger*, our propose model, will be useful in applications that recommend hashtags to users, for

assigning to tweets and other social network posts, while they post text content on social network platforms, and also, can be used in other social network based applications.

References

Sitaram Asur, Bernardo A Huberman, Gabor Szabo, and Chunyan Wang. 2011. Trends in social media: Persistence and decay. In *ICWSM*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 241–249. Association for Computational Linguistics.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2016. A paraphrase and semantic similarity detection system for user generated short-text content on microblogs. In *COLING*, pages 2880–2890.

Kuntal Dey, Saroj Kaushik, and L Venkata Subramaniam. 2017. Literature survey on interplay of topics, information diffusion and connections on social networks. *arXiv preprint arXiv:1706.00921*.

Zhuoye Ding, Zhuoye Zhang, and XuanJing Huang. 2012. Automatic hashtag recommendation for microblogs using topic-specific translation model. In *24th International Conference on Computational Linguistics*, page 265.

Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning topical translation model for microblog hashtag suggestion. In *IJCAI*, pages 2078–2084.

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 593–596. ACM.

Yuyun Gong and Qi Zhang. 2016. Hashtag recommendation using attention-based convolutional neural network. In *IJCAI*, pages 2782–2788.

Yeyun Gong, Qi Zhang, and Xuanjing Huang. 2015. Hashtag recommendation using dirichlet process mixture models incorporating types of hashtags. In *EMNLP*, pages 401–410.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Lichan Hong, Gregorio Convertino, and Ed H Chi. 2011. Language matters in twitter: A large scale study. In *ICWSM*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Zhiyuan Liu, Xinxiong Chen, and Maosong Sun. 2011. A simple word trigger method for social tag suggestion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1577–1588. Association for Computational Linguistics.

Zhi Liu, Chen Liang, and Maosong Sun. 2012. Topical word trigger model for keyphrase extraction. In *In Proceedings of COLING*. Citeseer.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

Surendra Sedhai and Aixin Sun. 2014. Hashtag recommendation for hyperlinked tweets. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 831–834. ACM.

Jieying She and Lei Chen. 2014. Tomoha: Topic model-based hashtag recommendation on twitter. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 371–372. ACM.

Kate Starbird and Leysia Palen. 2012. (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 7–16. ACM.

Oren Tsur and Ari Rappoport. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652. ACM.

Jason Weston, Sumit Chopra, and Keith Adams. 2014. # tagspace: Semantic embeddings from hashtags.

Eva Zangerle, Wolfgang Gassler, and Gunther Specht. 2011. Recommending#-tags in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings*, volume 730, pages 67–78.

Qi Zhang, Yeyun Gong, Xuyang Sun, and Xuanjing Huang. 2014. Time-aware personalized hashtag recommendation on social media. In *COLING*, pages 203–212.

Algorithm 1 THE SEMTAGGER ALGORITHM

function **CleanAndPreprocess ():**
$t'_r \leftarrow$ Raw tweets posted by user on Twitter
$t'_r \leftarrow t'_r - t'_{\tilde{h}}$, *i.e.*, remove all tweets without any hashtag
$t'_r \leftarrow t'_r - t'_{\tilde{en}}$, *i.e.*, remove all tweets not in English
$t'_r \leftarrow t'_r -$ char(non-ascii), *i.e.*, remove all non-ASCII characters
$t'_r \leftarrow$ norm(t'_r): perform tweet normalization using net-slang and Han-Baldwin
$t_r \leftarrow$ stopword_remove(t'_r): remove stopwords
return $T \leftarrow \{t_1, t2, ..., t_r, ..., t_n\}$, the cleaned and preprocessed tweets

function **LDABasedTopicModeling (Tweets T):**
$H \leftarrow \{h_1, h_2, ..., h_m\}$: set of hashtags present in the training set
$D \leftarrow \bigcup_{i=1}^{n}(t_i) - \bigcup_{j=1}^{m}(h_j)$: concatenation of all tweets, minus all the hashtags
$Z \leftarrow \{z_1, z_2, ..., z_l\} \leftarrow LDA(D)$: the set of topics identified to be present in the document D
return Z, a set of topics learned over LDA

function **RepresentativeTweetIdentification (Tweets T, Topics Z, Top-Ranks R as Integer):**
for $z_l \in Z$ **do**
    for $t_i \in T$ **do**
        $p_{t_i,z_l} \leftarrow$ LDA-based probability of tweet t_i to belong to topic z_l
        $t'_{z_l} \leftarrow$ insert t_i in sorted order of the value of p_{t_i,z_l}
    end for
    $t_{z_l} \leftarrow$, retain the highest R values contained in t'_{z_l}, discard the rest
end for
return $T_{z,L} \leftarrow \{t_{z_l}\} \forall (l \in L)$

function **SemanticSimBasedRec (Target Tweet t, Topics Z, R representative tweets T_{Z_q} across all topics Z_q, Integer K):**
Find the probability p_l of target tweet t to belong to each topic $z_l \in Z$
Sort by p_l and retain Z_q, the top Q topics
for all retained topics Z_q **do**
    $SS'(t, t_{Z_q}) \forall (t_{Z_q} \in T_{Z_q}) \leftarrow$ semantic similarity of target tweet t with representative tweet t_{Z_q}
end for
$SS \leftarrow$ Sort($SS'(t, t_{Z_q})$)
return Hashtags present in the top-K ranked tweets in SS

function **JointDistrMaxBasedRec (Target Tweet t, Topics Z, Integer K):**
for all topics z_l **do**
    for all representative tweets t_j in topic z_l **do**
        $SS(t, t_j) \leftarrow$ semantic similarity of target tweet t with representative tweet t_j
        $P(t, z_l) \leftarrow$ the LDA-based probability p_{t,z_l} of target tweet t to belong to topic $z_l \in Z$
        $CS'(t, t_j) \leftarrow SS(t, t_j) \times P(t, z_l)$
    end for
end for
$CS \leftarrow$ Sort($CS'(t, t_j)$)
return Hashtags present in the top-K ranked tweets in CS