A Context-Aware Approach to Detection of Short Irrelevant Texts

Sihong Xie*, Jing Wang*, Mohammad S.Amin†, Baoshi Yan†, Anmol Bhasin†, Clement Yu*, Philip S.Yu*
*Department of Computer Science University of Illinois at Chicago, Chicago, IL, USA
†LinkedIn Corp. Mountain View, CA, USA

Abstract

This paper presents a simple and effective framework that can detect irrelevant short text contents following blogs and news articles, etc. in a context-aware and timely fashion. Nowadays, websites such as Linkedin.com and CNN.com allow their visitors to leave comments after articles, and spammers are exploiting this feature to post irrelevant contents. Visited by millions of readers per day, these websites have extremely high visibility, and irrelevant comments have a detrimental effect on the visiting traffic and revenue of these websites. Therefore, it is critical to eliminate these irrelevant comments as accurately and early as possible. Different from traditional text mining tasks, comments following news and blog articles are characterized by briefness and context-dependent semantics, making it difficult to measure semantic relevance. What's worse, there could be only a handful of comments soon after an article is posted, leading to a severe lack of information for semantics and relevance measurement. We propose to infer "context-aware semantics" to address the above challenges in a unified framework. Specifically, we construct contexts for comments using either blocks of surrounding comments, or comments collected via a principled transfer learning approach. The constructed contexts mitigate the sparseness and sharply define context-dependent semantics of comments, even at the early stage of commenting activities, allowing traditional dimension reduction methods to better capture the semantics of short texts in a context-aware way. We confirm the effectiveness of the proposed method on two real world datasets consisting of news and blog articles and comments, with a maximal improvement of 20% in Area Under Precision-Recall Curve.

1. Introduction

Popular online content providers such as LinkedIn.com and CNN.com are attracting millions of visitors per day. Meanwhile, spammers and irresponsible visitors are leaving irrelevant comments after the major contents, making the websites less attractive to visitors and reducing the websites' traffic and revenue. It is critical to detect these irrelevant contents accurately as soon as possible. However, this is not an easy task due to the following reasons. First, comments are usually very short, and given such limited information, it is difficult to capture the semantics and relevance of the comments. Second, under different contexts, the same word can have quite different meanings. For example, given two news articles on real estate and NASA's mars exploration plan, respectively, the term "space" used in the comments of these articles can refer either to "an area rented or sold as business premises" or "the physical universe beyond the earth's atmosphere", two completely different concepts. The key observation is that the "context" of a comment plays an important role in defining the semantics and relevance of the comment. Third, in real world applications, there are situations where irrelevant comments are posted soon after the release of an article, with only a small number of comments. In Figure 1, we plot the number of articles on LinkedIn's news channel having various percentage of irrelevant comments at early stages. For instance, in Figure 1(a), we count the number of articles having 10%, 20%, etc. of irrelevant comments among the first 10 posted comments. It is obvious that a large number of articles have at least one irrelevant comment among the first 10 comments. The earlier one can remove these irrelevant contents, the less the visitors will be distracted. We call this task "early detection" of irrelevant contents, where irrelevant comments have to be spotted when there are only a handful of comments following the same article. It is much more difficult to measure the context-aware semantics and relevance of a comment at an early stage, since there is less information about the context of the comment.

Previous works failed to address the above challenges in a single framework. Regarding short text mining, there are two traditional ways: topic modeling and transferring of external data sources. [20] proposes to enhance the bag-of-word model using LDA [4]. In [26, 27], the authors propose novel topic models for short texts, and yet they did not address early detection Exploiting external corpus is also proposed to address the short text challenge, such as the works in [29, 21, 15, 11]. However, under the specific setting of the paper, how to define and transfer from external sources have not been investigated. Furthermore, these works focus on handling the sparseness of individual documents, instead of mitigating the sparseness of corpus that arises in early detection. The works [23, 13, 3, 14, 17, 16] try to characterize and catch irrelevant comments via bag-of-word model, sequence mining or information theoretical approach, but they also fail to address all the above challenges. On the one hand, the above methods derive the semantics of comments in a context-agnostic way, leading to more confusing semantics and degraded irrelevant content detection performance. On the other hand, early detection of irrelevant comments, though being critical in real applications, has been overlooked so far, to the best of our knowledge.

We propose to resolve the above three challenges in a unified framework. We want to derive context-dependent (i.e. context-aware) semantics of short texts regardless of the stages of commenting activities, such that it is more accurate in

978-1-4673-8273-1/15/$31.00 ©2015 IEEE

relevance measurement than those derived without considering contexts (context-agnostic). The context-dependent semantics of a comment is determined by the semantic environment (surrounding texts) where the comment sits in (such as the varying meaning of the word "space" in the above example). It is essential to select proper texts that are semantically meaningful and comparable to a comment as its context. We construct the "native context" of a comment as the set of the comments posted for the same article, since these comments are more likely to be similar to each other in terms of language, topics, etc.. The constructed native contexts can be coupled with any topic models to derive context-dependent semantics from short comments. Specifically, one can treat a native context as a corpus and employ any topic models such as LDA or SVD to find the context-dependent latent topics of the comments.

The native context constructed above assumes that there are sufficient comments posted for one article to serve as the context of a comment. However, regarding the early detection of irrelevant comments, one needs to tell irrelevant comments from only a handful of other comments. In other words, there are only a small number of comments in a native context at an early stage, posing difficulties to most topic models, which usually require a moderate number of documents for reliable topic inference. A key observation is that comments posted for articles on similar topics are more likely to have similar usages of language. For example, the comments following *articles* on "real estate" are more likely to use the term "space" in the sense of "residential/commercial space" rather than "space exploration". We propose to transfer similar short texts from other articles of similar topics to construct "transferred contexts", which inherit the strength of native contexts but avoid the sparseness of contextual information. Then similar topic models can derive context-dependent semantics for relevance measurement. The contributions of the paper are as follows:

- We identify the challenge of context-dependent irrelevant text detection, where the semantics of texts has to be considered under certain contexts. We propose "native contexts" (Section 4.1) that sharply define context-dependent semantics and relevance.
- We identify the challenge of early detection of irrelevant contents without sufficient context. We propose "transferred contexts" (Section 4.2) to address the sparseness of contextual information and accurately detect irrelevant comments in a timely fashion.
- We test the proposed methods on two real world datasets from a article sharing platform and a blog service (Section 5). We confirm the effectiveness of the proposed approaches with significant improvements over baselines.

2. Irrelevant content detection

Nowadays, popular websites allow users to post their opinions, mostly in the form of text comments following articles published by the websites. For example, on news websites such as CNN.com, a visitor can express his/her opinions

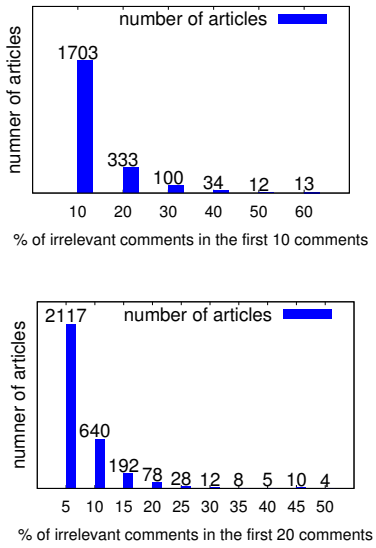


Fig. 1: Early detection as a real world problem

after reading the news about Obama's promotion of a new healthcare plan. Digg.com, wordpress.com and other social networks try to improve user engagements by deploying news and article sharing platforms, where their members can read the shared articles and post their opinions as responses. Due to the high visibility of the news and social network websites, spammers are joining the community to produce junk comments. Also, there are readers who are exploiting the traffic to these websites and distracting other visitors to irrelevant topics. These irrelevant comments can be detrimental to user experience of the websites, whose traffic and revenue will be affected. It is therefore an emergency task for the operators of these popular websites to detect undesirable comments and take appropriate actions. Intuitively, a normal comment should either respond to the contents of the article it follows, or sound similar to other comments following the same article (we called these comments the "surrounding comments"). Therefore, the irrelevant comments can be detected by measuring the similarity between a comment and the article it follows, and also between the comment and its surrounding comments. If either of the similarities is too low, then the comment is likely to be an irrelevant one [16, 23, 5]. Indeed, content similarity is the most natural definition of relevance, as it is the way human interpret contents.

More formally, assume an article $\mathbf{w}_d \in \mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$ is followed by a set of C_d comments $\mathbf{Q}^d = \{\mathbf{q}^d_1, \ldots, \mathbf{q}^d_{C_d}\}$ (see Table 1 for a summary of notations). $\mathbf{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ and $\mathbf{q}^d_k = \{q^d_{kn}\}_{n=1}^{N^d_k}$ are the vectors of words of the d-th article and the k-th comment for the article, respectively. N_d and N^d_k are the lengths of the article and the comment, respectively. Assume $f(\cdot)$ is a language model, which is a transformation from the bag-of-word vector representation of a document to anther vector representation. For example, LDA (Latent Dirichlet Allocation) maps a document to a vector of topic distribution, while an identity transformation is simply

the bag-of-word vector of a document (see Section 3 for more details). Such a transformation might be necessary for text mining since it can potentially capture the high-level meanings of the documents, especially when the documents are short. Given a transformation $f(\cdot)$, the signals for irrelevant comment detection based on *text* can be calculated as the cosine similarity between $f(\mathbf{q}_k^d)$ (the comment) and $f(\mathbf{w}_d)$ (the article the comment follows) and the mean of $\{f(\mathbf{q}_1^d), \ldots, f(\mathbf{q}_{C_d}^d)\}$ [5]:

$$\cos(f(\mathbf{w}_d), f(\mathbf{q}_k^d)) = \frac{\left\langle f(\mathbf{w}_d), f(\mathbf{q}_k^d) \right\rangle}{\|f(\mathbf{w}_d)\| \cdot \|f(\mathbf{q}_k^d)\|} \quad (1)$$

$$\cos(\mathbf{m}_d, f(\mathbf{q}_k^d)) = \frac{\left\langle \mathbf{m}_d, f(\mathbf{q}_k^d) \right\rangle}{\|\mathbf{m}_d\| \cdot \|f(\mathbf{q}_k^d)\|} \quad (2)$$

where $\mathbf{m}_d$ is the center of all transformed vectors of comments following $\mathbf{w}_d$

$$\mathbf{m}_d = \frac{\sum_{\mathbf{q} \in Q^d} f(\mathbf{q})}{C_d} \quad (3)$$

We call Eq.(1) the "comment-to-article" irrelevance signal and Eq.(2) the "comment-to-center" irrelevance signal.

From the above formula, one can see that similarity measurement requires a vector representation of texts, namely the transformation $f(\cdot)$. Ideally, $f(\cdot)$ should capture the meaning of the texts well for the detection signals to make sense. However, this is not an easy task and there are three challenges. First, comments are usually very short, compared to the documents processed in traditional text mining. In general, the articles published by the websites are of medium length such that they are easy for the readers to follow. In contrast, the comments that follow are usually short, since readers are less serious and therefore unable or unwilling to produce long and organized texts. Figure 2 shows the distribution of the length of comments from a social network website, and one can see that most of the comments have less than 150 words. Due to the sparsity of the comment texts, the information provided by individual comment is very limited, and dimension reductions are usually required for this situation [20, 4, 26, 27], though it is unclear from the previous work that how effective these methods are in the irrelevant short text detection task.

Second, the semantics of comments are context-dependent. Specifically, a word in the comments might mean two different things under articles on two different topics, as the above-mentioned example shows. This variety of the semantics of words can not be fully captured by the bag-of-words representation or any other dimension reduction methods such as LDA [4], pLSA, SVD, etc., since these models ignore the contexts where a piece of text is generated. These methods are "context-agnostic". As a result, given a comment, these models will give the same vector representation for the comment, no matter where the comment is posted. This is undesirable since under different contexts, an ideal language model should be able to capture subtle semantic difference.

Third, in real world applications, real time actions to irrelevant contents are of high priority. Spammers or promoters are more likely to post junk comments soon after an article is posted, such that a larger amount of audience can see their comments (as shown in Figure 1). Meanwhile, if too many visitors read the undesirable comments, they can have an unpleasant experience, leading to a lower user engagement. Therefore, it is necessary for website operators to detect irrelevant comments as soon as they show up. However, the lack of surrounding comments makes it difficult to define context for a comment, and one might have to resort to less effective context-agnostic approaches. To sum up, it is an important yet difficult problem to detect irrelevant short texts, with context-dependent semantics and lack of contexts. Before we address the above challenges, we first briefly review some existing context-agnostic methods.

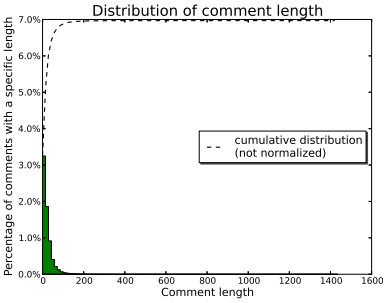


Fig. 2: Distribution of length of comments

TABLE 1: Notations

Symbol	Meaning
$\mathbf{W}$	the collection of major posts
$\mathbf{w}_d$	the d-th post
$\mathbf{Q}^d$	the comments following the d-th post
$\mathbf{q}_k^d$	the k-th comment following the d-th post
C_d	the number of comments following post $\mathbf{w}_d$
N_d	the length of the d-th post
D	the size of the corpus
$\|\cdot\|_F$	Frobenius norm of a matrix
$\langle \cdot, \cdot \rangle$	inner product
$f(\cdot)$	a transformation defining a language model

3. Context-Agnostic Detection Models

3.1. Simple Language Model

The simplest language model is perhaps the bag-of-words representation of documents. Using this model, a document $\mathbf{w}$ is given by a vector describing the number of occurrences of words (or the TF-IDF processed version) in the document. Then the bag-of-word vector transformation function $f_{bow}(\cdot)$ is simply an identity function. [16] adopts this language model and use the comment-to-article similarity (Eq.(1)) to detect irrelevant comments. A drawback of bag-of-words vector representation is that the vectors are usually sparse, given a large vocabulary. Indeed, in [20], it is shown that LDA (introduced next) can greatly improve the classification performance based on cosine similarity on short texts.

3.2. Probabilistic Topic Models

Probabilistic topic models assign a distribution of topics to a document. A popular one is the LDA (Latent Dirichlet Allocation) model. The success of LDA relies on its ability to learn topic distributions of terms and documents simultaneously. LDA assumes that a document is a mixture of topics and each word in the document is generated according to the topic of the document and the distribution of words over topics. More formally, given a document $\mathbf{w}_d = \{w_{dn}\}_{n=1}^{N_d}$,

$$\begin{aligned} \theta_d &\sim \text{Dir}(\boldsymbol{\alpha}) \\ z_{dn} &\sim \text{Multi}(\theta_d) \quad \forall n = 1, \ldots, N_d \\ w_{dn} &\sim \text{Multi}(\boldsymbol{\Phi}_{z_{dn}}) \quad \forall n = 1, \ldots, N_d \end{aligned}$$

where θ_d is the K dimensional topic distribution of document $\mathbf{w}_d$, and z_{dn} is the topic of the word w_{dn}. Dir($\boldsymbol{\alpha}$) is the Dirichlet distribution with parameter $\boldsymbol{\alpha}$ and Multi(θ) is the multinomial distribution with parameter θ. Given a corpus $\mathbf{W}$, LDA infers the quantities θ_d, z_{dn} and $\boldsymbol{\Phi}$. Monte Carlo Markov Chain (MCMC) and variational methods are widely used for model inference and learning. Let $f_{lda}(\mathbf{w}_d) = \theta_d$ be the vector transformation function derived from LDA.

3.3. Matrix Factorization based Models

Besides LDA, matrix factorization based methods are also employed to find topics of documents. Usually, the observed corpus is modeled as a term-document matrix $\mathbf{W}$ (here we abuse the notation), which is further factorized into the product of two or three matrices. For example, in LSI (Latent Semantic Indexing [6]) or SVD [8],

$$\mathbf{W} = U\Sigma V^\top \tag{4}$$

where U (V) is the left (right) matrix of singular vectors and Σ is the diagonal singular value matrix. Here U gives the topic distributions of words and V gives the topic distributions of documents. Therefore, the vector transformation function is given by $f_{svd}(\mathbf{w}_d) = V_d$, where V_d is the d-th row of V. In a similar form, non-negative matrix factorization (NMF) has also been shown to be effective in finding latent topic of documents in information retrieval [25]. Formally, NMF solves the following optimization problem

$$\min_{U,V} \quad \|\mathbf{W} - UV^\top\|_F \tag{5}$$

$$\text{s.t.} \quad U_{ij} \geq 0, \quad V_{ij} \geq 0 \quad \forall i, j \tag{6}$$

Similar to SVD, a row vector in the factor matrix V gives the topic distribution of a document and $f_{nmf}(\mathbf{w}_d) = V_d$.

3.4. Detection Signals based on Context-Agnostic Models

Based on the above models and Eq.(1) and Eq.(2), we define several irrelevant comment detection signals, which are summarized in Table 2. In the table, each row specifies a signal (e.g. σ_1), and the signals in the rows "Native" and "Transferred" will be defined in the next section. A check mark under the column "Mean" ("Article") indicates that Eq.(2) (Eq.(1),respectively) is used to compute the signal. Note that each of $\sigma_i, i = 1, \ldots, 4$ includes two similarities. These models cannot handle context-dependent semantics: none of them takes the contexts of a comment into account when computing the transformations $f(\cdot)$, thus the derived signals $\sigma_1, \ldots, \sigma_4$ fail to capture the context-dependent semantics when used for irrelevant comment detection.

TABLE 2: Irrelevant comment detection signals based on context-agnostic models

Signal	Context	Transformation	Mean	Article
σ_1	Agnostic	f_{bow} [16]	✓	✓
σ_2	Agnostic	f_{lda} [20]	✓	✓
σ_3	Agnostic	f_{svd}	✓	✓
σ_4	Agnostic	f_{nmf} [25]	✓	✓
σ_5	Native	f_{svd}^N	✓	
σ_6	Transferred	f_{svd}^G	✓	

4. Context-Aware Detection Signals

Below we first introduce "native context" to derive context-dependent semantics of short comments. Then we point out a practical situation where this native construction may fail, and propose a "transferred context" to handle the difficulty.

4.1. Native Contexts

The vector transformation function $f(\cdot)$ used in Eq.(1) and Eq.(2) should depend on the contexts of a comment. We observe that an article sets up the topics that are to be discussed by the comments that follow, which should have similar usages of language. Therefore, the articles naturally separate all comments into groups, each of which defines a context for the comments within. If one can learn a language model (a transformation) using such contexts for the comments, then context-dependent semantics of the comments are more likely to be well-captured.

Formally, we define the native context (NC) of a comment, say $\mathbf{q}_k^d$, to be the neighboring comments following the same article as $\mathbf{q}_k^d$, namely, all the comments in $\mathbf{Q}^d$:

$$\mathsf{NC}(\mathbf{q}_k^d) = \mathbf{Q}^d$$

To learn a context-aware language model for $\mathbf{q}_k^d$ using $\mathbf{Q}^d$, matrix factorizations, such as SVD, can be applied to the term-document matrix constructed from $\mathbf{Q}^d$:

$$\mathbf{Q}^d = U^d\Sigma^d(V^d)^\top \tag{7}$$

Here we abuse the notation by using $\mathbf{Q}^d$ for both the set of comments and the term-document matrix constructed from the set. We use superscript d to emphasize that the decomposition depends only on the neighboring comments, instead of *all*

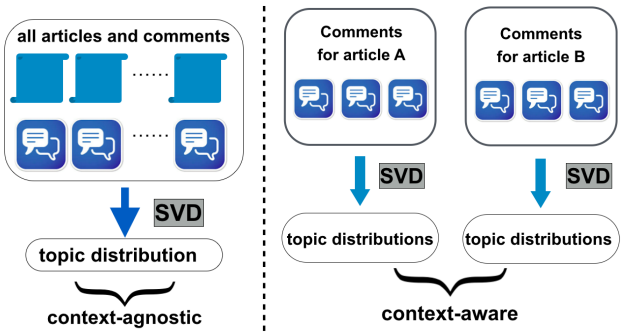


Fig. 3: Context-Agnostic vs. Context-Aware methods

comments in the corpus. The resulting factor matrix V^d gives a context-aware topic distribution of the comments:

$$f^N_{svd}(\mathbf{q}^d_k) = V^d_k \tag{8}$$

where V^d_k is the k-th row of V^d and $f^N_{svd}(\cdot)$ is the vector-to-vector transformation obtained by decomposing the native context using SVD. Lastly, we compute a signal (σ_5 in Table 2) for irrelevant comment detection by plugging $f^N_{svd}(\cdot)$ in Eq.(2) and Eq.(3):

$$\cos(\mathbf{m}_d, f^N_{svd}(\mathbf{q}^d_k)), \qquad \mathbf{m}_d = \frac{\sum_{\mathbf{q}\in Q^d} f^N_{svd}(\mathbf{q})}{C_d} \tag{9}$$

Note that we do not include the corresponding article $\mathbf{w}_d$ in the decomposition in Eq.(7), since the length of an article and a comment can differ dramatically such that the decomposition will be biased to favor the article. Indeed, we observed in the experiments, that including the article in the native context of a comment actually hurts the performance (not reported). As a result, we do not use comment-to-article similarity for detection. Nonetheless, one will soon see that the articles play a critical role in addressing the sparsity issue in early detection. In summary, the difference between context-agnostic and context-aware language models is demonstrated in Figure 3. On the left we pool all articles and comments together and apply SVD to the corresponding term-document matrix, and on the right we perform multiple SVDs on the term-document matrices derived from native contexts.

4.2. Early Detection of Irrelevant Comments

Although the proposed native context can define and measure context-dependent semantics and relevance in normal settings, it is insufficient for the early detection task. In particular, when there are only a small number of comments following one article, the term-document matrix ($\mathbf{Q}^d$ in Eq.(7)) fails to provide enough information for SVD to infer meaningful topic distributions for the comments. Even if one could manage to estimate the topic distributions of the comments, the comment-to-center similarity signal would not make much sense. This is because the center $\mathbf{m}_d$ in Eq.(9) is the mean of a small sample and thus the variance of this estimation can be rather high according to large sample theory [1], making the signal too noisy for reliable detection. However, if one totally ignores contextual information, the context-dependent semantics cannot be sharply defined. As shown in the experiments, the lack of context leads to degenerated performance.

We propose to generalize the native contexts and add more information. The native context for a comment is defined based on the "comment-follows-article" relationship, as shown in the right panel of Figure 3. The essence of native context is to exploit the topical coherence among comments following the same article. We adopt the same idea to include more comments to define a useful context that can mitigate the sparseness of comments in early detection. The intuition is that articles of similar topics are likely to be followed by comments of the same topics, with similar usage of language. For example, the term "space" in the comments following *multiple* articles on "real estate" is likely to unambiguously refer to "a continuous area for human daily activities", instead of "the physical universe around the earth". Therefore, we can transfer the comments from articles with similar topics to define a context for the comments under investigation. Such transfer is possible since popular websites store past articles and the associated comments in their databases. However, there are drifts in concepts and distributions in the comments in different articles, not all historic comments are useful for the current detection tasks. To address this issue, among the comments from similar articles, we only transfer comments that are most similar to the current ones. We define these transferred comments, together with the current comments, as the "transferred context".

Algorithm 1 Constructing Early Detection Signal using Transferred Context

Input: An article $\mathbf{w}$ with its comments $\mathbf{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_C\}$, a collection of past articles $\{\mathbf{w}_d\}_{d=1}^D$ and associated comments $\{\mathbf{Q}^d\}_{d=1}^D$.
Output: Irrelevance detection signal σ_6 for $\mathbf{q}_k \in \mathbf{Q}$.

Derive LDA topics for $\{\mathbf{w}\}$ and $\mathbf{Q}$ using trained LDA model.
Retrieve top ℓ most similar articles to $\mathbf{w}$ from $\{\mathbf{w}_d\}_{d=1}^D$ using LDA topics. The retrieved articles are $R = \{\mathbf{w}'_1, \ldots, \mathbf{w}'_\ell\}$.
for $\mathbf{q}_i \in \mathbf{Q}$ **do**
    Retrieve top 50% most similar comments to $\mathbf{q}_i$ from the comments associated with articles in R.
end for
Define transferred context for $\mathbf{Q}$ as the union of the retrieved comments and $\mathbf{Q}$.
Apply SVD to the transferred context to find context-dependent semantics of $\mathbf{Q}$.
Return σ_6 calculated using Eq. (2) and Eq. (3).

The idea of constructing transferred contexts and the corresponding detection signal is described in Algorithm 1, and is demonstrated in Figure 4. In summary, transferred contexts address the sparsity of neighboring comments that native contexts suffer, and allow topic models to define context-aware semantics that is not available in context-agnostic methods.

Since we are focusing on early detection, efficiency becomes an issue. Here we claim that the run-time of Algorithm 1 allow the algorithm to be practically useful. First, there is no intensive computation involved in deriving topics using

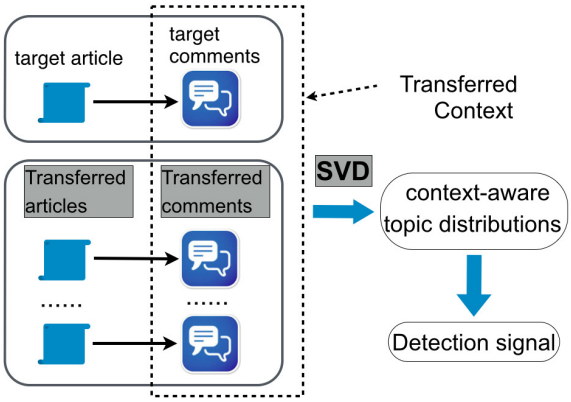


Fig. 4: Transferred Contexts

a trained LDA model. The retrieval of articles (step 4) can be done in parallel frameworks like MapReduce. Similarly, step 5 to 7 can be done in parallel, where each $\mathbf{q}_i$ can be processed independently. Lastly, though in general SVD requires cubic time complexity, the matrix to be decomposed here is small and sparse. There are fast algorithms that can exploit the sparsity of the matrix. If this really becomes an bottleneck, one may resort to parallelized SVD [2].

5. Experiments

5.1. Preparation of Datasets

We obtained two real world datasets from the news channel of LinkedIn.com (News in the sequel) and the blog service Digg.com (Blog in the sequel). For the News data, we obtain a snapshot of the news channel in May, 2013, containing a total of 200,000 comments and 5,000 articles. Since labeling a comment as relevant or irrelevant requires reading and comparing the comment and the followed article, it is very time-consuming and costly to label all comments collected, therefore we randomly sample 20,000 article-comment pairs and send them to the crowdsourcing service crowdflower.com. The crowdsourcing tasks are such designed that one task consists of an article and 10 comments, randomly picked from the pool of all following comments. A worker is instructed to first read the original article and then the comments, if he/she finds a comment is irrelevant to the article, he/she should label the comment as positive, otherwise negative. The workers are required to label all the comments to get the credit. We take several measures to ensure a certain level of label quality. Firstly, we inject an editor-labeled golds in each task, and the crowdflower platform has a mechanism to prevent a worker from further labeling the tasks if his/her competence based on the golds is lower than a pre-defined threshold. Secondly, we require that each comment is labeled by 3 workers in order to derive a confidence level of the majority voting. After harvesting the labels, we discard those comments with the lowest confidence level and keep only 6952 of them. Lastly, human experts in our corporation looked into a small amount of randomly picked labeled comments to check that the crowdsourced labels are consistent with our definition of "irrelevance". The details of the blog dataset can be found in [23]. The characteristics of these two datasets are summarized in Table 3, from which we observe that negative instances significantly outnumber positive ones, presenting an imbalance class distribution (note that this is also true for early detection tasks, see Figure 1).

TABLE 3: Dataset Characteristics

	News	Blog
# articles	363	20
# comment-article pairs	6,952	2,109
% positive instances	4.54%	28.2%

5.2. Experimental Settings and Results

Baselines

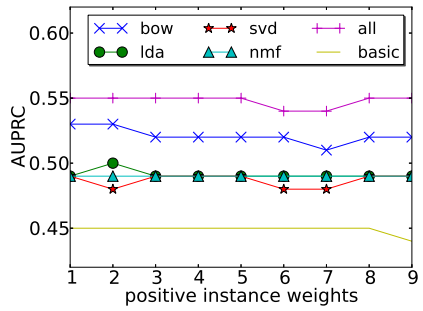
Note that the method proposed in [16] is basically σ_1 without smoothing (which requires a larger corpus retrieved from the web). σ_2 corresponds to the approach in [20] and σ_4 corresponds to that in [25]. We demonstrate the effectiveness of the context-aware signals by comparing them to several enhanced baselines proposed in [20, 16, 25]. Each enhanced baseline consists of two parts of features: the basic features and one of the baseline context-agnostic signals $\sigma_1, \ldots, \sigma_4$. For the News dataset, a comment can be characterized by basic features based on the author's social network connections and certain text features that are not derived from semantic relevance, such as the lengths of the comments, containment of any sensitive keywords, etc..[1] We also include the output of a maximum entropy text classifier as an additional basic feature. For the Blog dataset, we withhold 50% of the comment-article pairs in the Blog dataset as training data and train various classifiers (SVM, kNN, naive Bayes), whose predictions of a comment being irrelevant are treated as basic features. To derive the signals $\sigma_1, \ldots, \sigma_4$: 1) we train an LDA[2] model using all articles, then predict the topics of all comments. 2) we construct a term-document matrix using all articles and comments, then use SVD and NMF to decompose the resulting matrix and obtain topics of articles and comments. We fix the number of topics in SVD, LDA and NMF at 50 without parameter searching.

Effectiveness of Native Contexts

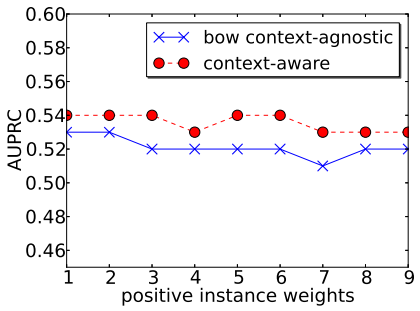
Recall that the information in the constructed native contexts is given by the signal σ_5 in Table 2. To demonstrate that the proposed native context can enhance various context-agnostic methods, we compare the classification performance of the basic features with and without signal σ_5. Without searching the parameter, we set the number of topics in Eq.(7) to 20,

1. Due to corporation privacy, we are unable to discuss the details of these features

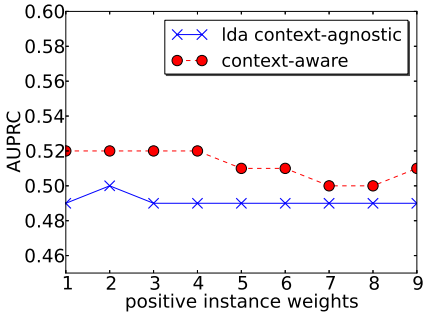
2. use the implementation GibbsLDA++, with default parameters except the number of topics



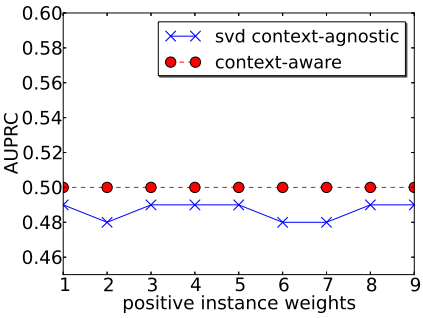
(a) Context-agnostic methods



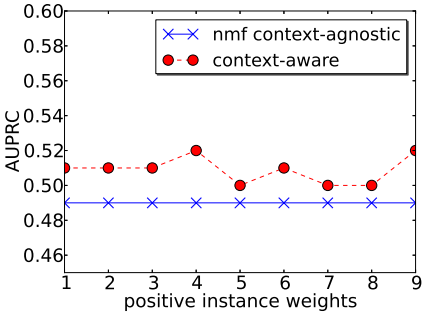
(b) BOW with native context



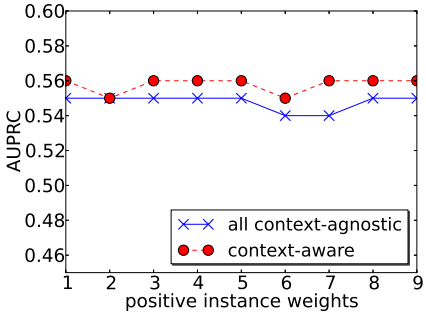
(c) LDA with native context



(d) SVD with native context

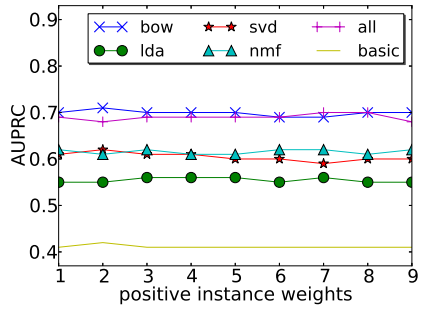


(e) NMF with native context

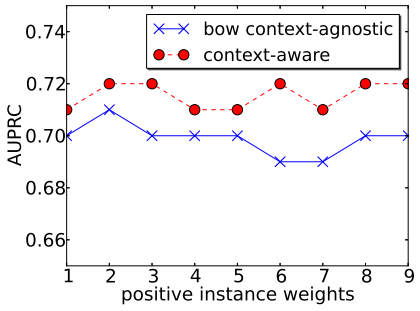


(f) All context-agnostic signals with native context

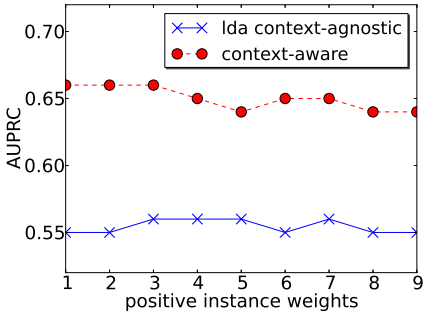
Fig. 5: Effectiveness of Native Context on the News dataset



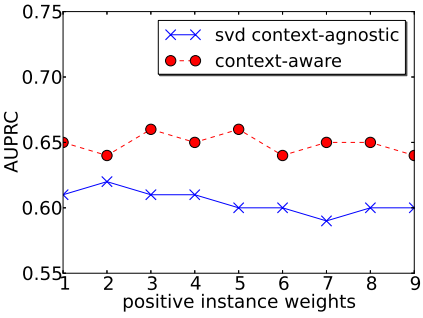
(a) Context-agnostic methods



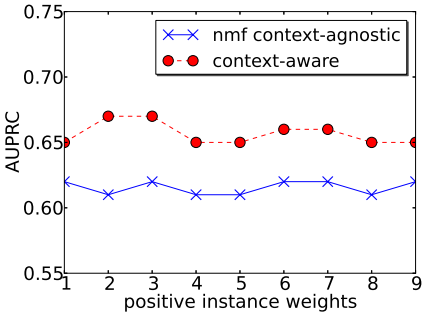
(b) BOW with native context



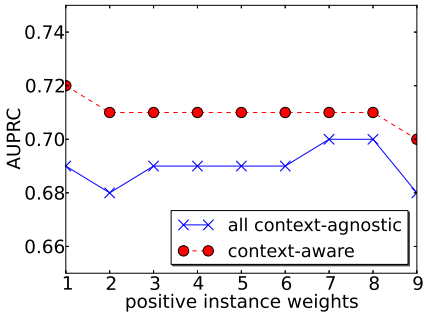
(c) LDA with native context



(d) SVD with native context



(e) NMF with native context



(f) All context-agnostic signals with native context

Fig. 6: Effectiveness of Native Context on the Blog dataset

as there are less documents in native contexts. Since there are several context-agnostic methods (BOW, LDA, SVD and NMF), we add σ_5 to each of the signals in $\{\sigma_1, \ldots, \sigma_4\}$ corresponding to the above methods. For example, σ_5 can be combined with σ_1 and other basic features. We also add σ_5 to all of $\{\sigma_1, \ldots, \sigma_4\}$ and other basic features. In sum, we have 5 different combinations of σ_5 with the other signals. If the combinations of features with σ_5 outperform the same sets of features without σ_5, then it is demonstrated that the native context does capture context-dependent semantics, which would otherwise be unavailable through context-agnostic methods.

We use the random forest implementation in sklearn[3] to evaluate each set of features, since random forest has been proven to be effective for imbalance two-class problems, as it is the case in this paper. Regarding the forests, we use 100 random trees, each of which grows to its full depth. The performance of random forest is evaluated using 10-fold cross validation. We choose AUPRC (Area Under Precision-Recall Curve) as our performance metric, as in real world applications like spam detection, one usually wants to achieve high precisions with low recalls. Note that one can adjust the cost of false negatives in imbalance classification problems. Therefore, with the weight of negative instances fixed at 1, we give different weights to positive instances, ranging from 1 to 9 with stepsize 1. Random decision trees can gracefully take care of the weights.

In Figure 5 (News dataset) and Figure 6 (Blog dataset), we demonstrate the performance of various signal combinations. In Figures 5(a) and 6(a), one can observe that the signals $\sigma_i, i = 1, \ldots, 4$ improve the detection performance based on the rest of the basic features. This shows that the similarity between the usage of words or topics of a comment and the proceeding article or surrounding comments can significantly improve the performance. Surprisingly, on both datasets, f_{bow} outperforms any other single dimension reduction methods (f_{lda}, f_{svd} or f_{nmf}) that try to capture the topics of the comments. This is because comments are usually too short to provide sufficient information for topic modeling. In Figure 5(a), we observe that by combining all context-agnostic signals, one can obtain a significant improvement on the News dataset, though not so on the Blog dataset in Figure 6(a). We improve the performance of context-agnostic signals consistently by including a context-aware signal σ_5, as shown in Figures 5(b)-5(e), and Figures 6(b)-6(e). For example, on the News dataset, the native context maximally improves LDA and NMF by 6.1%. On the Blog dataset, the improvements are even more significant, where the native context improves LDA by 20% (Figure 6(c)). More importantly, the improvements are consistent regardless of the cost of false negatives, eliminating the time-consuming process of tuning the cost parameter in real world applications.

In Figure 5(f) and Figure 6(f), we show the improvements due to native contexts on the combination of all context-agnostic signals. The improvements are 1.8% on the News dataset and 4.3% on the Blog dataset. Note that using all 4 context-agnostic models gives the best performance on the News dataset (Figure 5(a)), and the proposed native context brings the AUPRC even higher. In real world applications, it is more important to locate certain points on the precision-recall-curve where precisions are high. In Figure 7(a) and Figure 7(b), we plot the PRCs when bundling all context-agnostic models with and without σ_5 for both datasets. The areas where precisions are at least 80% are annotated using arrows. It is clear that native contexts consistently improve the performance over the combined context-agnostic models by achieving higher recalls in the critical regions.

3. scikit-learn.org

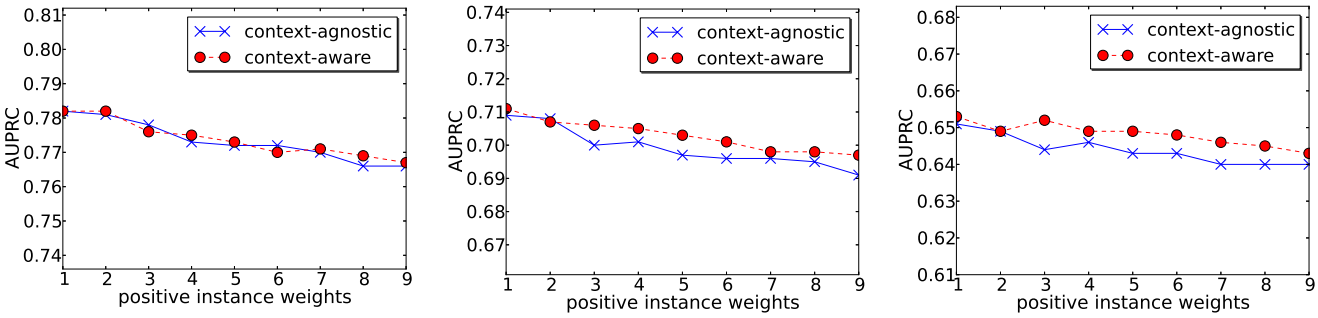
Effectiveness of Transferred Contexts

For each irrelevant comment, we randomly sample a certain number (2, 4, and 6) of relevant comments following the same article, then we treat the irrelevant comment and the sampled relevant comments as the only available comments for the article. We run Algorithm 1 to construct transferred contexts and derive detection signal σ_6 in Table 2. σ_6 is then added to the combination of all context-agnostic signals $\sigma_1, \ldots, \sigma_4$, since the combined signals have the best performance on this dataset (Figure 5(a)). We do not include the comment-to-center similarity for $\sigma_1, \ldots, \sigma_4$, since there are only a very small number of comments at an early stage and the estimated center is inaccurate. The context-agnostic signals are generated as follows: SVD and NMF are used to decompose the term-document matrices derived from articles and the associated positive/sampled negative comments; LDA and BOW are the same as they were in the last experiment. Since there is a source of randomness due to sampling, we repeat the experiment 10 times for each parameter setting and report the mean AUPRC. We perform this experiment only on the News dataset, since there are only 20 articles in the Blog dataset, based on which the results might not be significant.

The mean of AUPRC of the methods with and without σ_6 are compared in Figure 8. Each of the figures (from left to right) is obtained using different number of sampled normal comments. In Figure 8(a), one can see that transferred contexts only slightly change the AUPRC, when the detection task is relatively easy (smaller number of comments to distinguish). However, when there are more negative samples but insufficient contexts, the detection tasks become much more difficult. In such situations, the transferred contexts start to serve as a good source for detection signal σ_6. In Figure 8(b) and 8(c), one can see that σ_6 improves the AUPRC more than it does in Figure 8(a). In particular, in Figure 8(c), the improvements are most obvious.

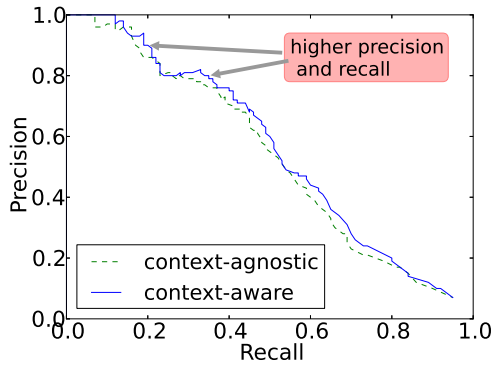
6. Related works

Spam detection in user-generated contents has been studied intensively in previous works, which define anomalies using various of signals based on temporal and spatial user behaviors [18, 7, 9, 10], text contents [23], network connections [22, 28], etc. In [3], they analyze several textual features that might

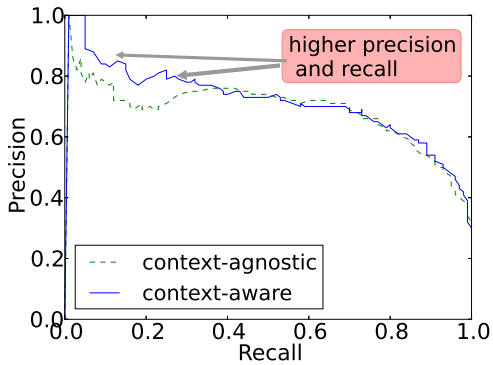


(a) 2 relevant + 1 irrelevant comments (b) 4 relevant + 1 irrelevant comments (c) 6 relevant + 1 irrelevant comments

Fig. 8: Effectiveness of Transferred Contexts on the News dataset



(a) PRC for all context-agnostic signals with and without native context on the News dataset



(b) PRC for all context-agnostic signals with and without native context on the Blog dataset

Fig. 7: Precision-Recall Curves for the context-agnostic and context-aware detections

be predictive in detecting comment spams. [16] is probably the first work to consider irrelevant comments. Their method is simply to employ as the detector the KL-divergence between the smoothed probabilistic word distributions of the posts and comments. In [23], in order to facilitate post-comment similarity measurement, they propose a more complicated comment processing pipeline, including search engine based expansion, co-reference resolution, proper nouns identification, LDA topic modeling. In [13], they propose a frequent subsequence mining based method to catch comment spams. In [14], they define spam as "content that is uninformative in the information-theoretic sense". They propose to use entropy rate as a way to measure informativeness of a comment. These works differ from the proposed method in that they do not consider any context information of texts. Besides, they fail to handle the early detection task, which is critical for operating social network and news websites.

Learning from short text has been a major difficulty in text mining, and there are some active research on short text topic modeling [20, 26, 27]. In [20], they focus on search engine snippets and medical article abstract classification. LDA derived topic distribution of the short texts is employed to enhanced the bag-of-word representation. In [27], they propose to first estimate the topic distributions of words using the term correlation matrix. Then, fixing the topic distributions of words, they estimate the topics of documents. In [26], they propose a novel topic model that can directly model the word co-occurrence (biterm) patterns of terms without the document-level topics. Although the above works can effectively model topic distributions of short texts, they do not consider context-dependent semantics and early detection.

Another direction in handling short text is via document expansion [23, 29, 12, 21, 11] or transfer learning [19]. In [23, 21, 29], they use search engines as a way to augment the short texts. Specifically, they issue a query for a piece of short text and append the returned results to the short text, such that more relevant information is included. Given the number of short texts to be processed is huge, it is inefficient to issue queries on search engine for each piece of short text, and thus their methods are not applicable to our problem settings. In [11], they use a similar idea but propose to include WordNet and Wikipedia as external texts to enrich short texts. There have also been some works on using transfer learning to help find the topics of short texts. In [30], they learn entity types from short text queries by using Word2Vec in a graph-based modeling. In [12], they show that by simultaneously modeling the topics of both long and short texts can mitigate the sparsity of short texts. In [15], they propose to transfer external texts to help find the topics of short text via selective sampling the external texts. These methods are not directly comparable

to the proposed approach in the paper, as one can always plug one of these methods in the dimension reduction step of our method. In [24], a method is proposed to understand short texts in users' queries in search engine, however, their method requires users' feedbacks, making it not suitable for spam detection tasks.

7. Conclusion

In this paper, we propose to use contexts to resolve the challenges in irrelevant comment detection: briefness of comments, variety of semantics of words and lack of information in early detection. Native context is proposed to capturing context-dependent sematics of words, leading to better performance than 4 traditional models without considering contexts. Then transferred context is proposed to handle the more difficult situation when there is insufficient information for native context. Experimental results on two real world datasets confirm the effectiveness of the proposed context-aware approach.

Acknowledgment

This work is supported in part by NSF through grants III-1526499, CNS-1115234, and OISE-1129076, Google Research Award, and the Pinnacle Lab at Singapore Management University.

References

[1] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3rd edition, 2003.

[2] MW Berry, D Mezher, B Philippe, and Sameh A. *Parallel algorithms for the singular value decomposition*. 2006.

[3] Archana Bhattarai, Vasile Rus, and Dipankar Dasgupta. Characterizing comment spam in the blogosphere through content analysis. Computational Intelligence in Cyber Security, 2009.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. 2003.

[5] Yi Chang, Xuanhui Wang, Qiaozhu Mei, and Yan Liu. Towards twitter context summarization with user influence models. WSDM, 2013.

[6] Scott Deerwester. Improving information retrieval with latent semantic indexing. In *Proceedings of the 51st ASIS Annual Meeting*, 1988.

[7] Geli Fei, Arjun Mukherjee, Bing Liu 0001, Meichun Hsu, Mal Castellanos, and Riddhiman Ghosh. Exploiting burstiness in reviews for review spammer detection. In *ICWSM*, 2013.

[8] G.H. Golub and C.F.V. Loan. *Matrix Computions*. John Hopkins University Press, 3rd edition, 1996.

[9] Qingbo Hu, Guan Wang, and Philip S. Yu. Deriving latent social impulses to determine longevous videos. WWW, 2014.

[10] Qingbo Hu, Guan Wang, and P.S. Yu. Assessing the longevity of online videos: A new insight of a video's quality. In *DSAA*, 2014.

[11] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. CIKM, 2009.

[12] Ou Jin, Nathan N Liu, Kai Zhao, Yong Yu, and Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. CIKM, 2011.

[13] Ravi Kant, Srinivasan H Sengamedu, and Krishnan Kumar. Comment spam detection by sequence mining. WSDM, 2012.

[14] Alex Kantchelian, Justin Ma, Ling Huang, Sadia Afroz, Anthony Joseph, and J. D. Tygar. Robust detection of comment spam using entropy rate. AISec, 2012.

[15] Guodong Long, Ling Chen, Xingquan Zhu, and Chengqi Zhang. Tcsst: transfer classification of short & sparse text using external data. CIKM, 2012.

[16] Gilad Mishne, David Carmel, and Ronny Lempel. Blocking blog spam with language model disagreement. In *AIRWeb*, 2005.

[17] Gilad Mishne and Natalie Glance. Leave a reply: An analysis of weblog comments. WWW, 2006.

[18] Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Spotting opinion spammers using behavioral footprints. KDD, 2013.

[19] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10), 2010.

[20] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. WWW, 2008.

[21] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. WWW, 2006.

[22] Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu. Review graph based online store review spammer detection. ICDM, 2011.

[23] Jing Wang, Clement T. Yu, Philip S. Yu, Bing Liu, and Weiyi Meng. Diversionary comments under political blog posts. CIKM, 2012.

[24] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. SIGIR, 1996.

[25] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. SIGIR, 2003.

[26] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. WWW, 2013.

[27] Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. SDM, 2013.

[28] Junting Ye and Leman Akoglu. Discovering opinion spammer groups by network footprints. ECML/PKDD, 2015.

[29] Wen-Tau Yih and Christopher Meek. Improving similarity measures for short segments of text. AAAI, 2007.

[30] Jingyuan Zhang, Luo Jie, Altaf Rahman, Sihong Xie, Yi Chang, and Philip S Yu. Learning entity types from query logs via graph-based modeling. In *CIKM*, 2015.