# Leveraging Knowledge across Media for Spammer Detection in Microblogging

Xia Hu, Jiliang Tang, and Huan Liu
Computer Science and Engineering
Arizona State University
Tempe, AZ 85287, USA
{xia.hu, jiliang.tang, huan.liu}@asu.edu

## ABSTRACT

While microblogging has emerged as an important information sharing and communication platform, it has also become a convenient venue for spammers to overwhelm other users with unwanted content. Currently, spammer detection in microblogging focuses on using social networking information, but little on content analysis due to the distinct nature of microblogging messages. First, label information is hard to obtain. Second, the texts in microblogging are short and noisy. As we know, spammer detection has been extensively studied for years in various media, e.g., emails, SMS and the web. Motivated by abundant resources available in the other media, we investigate whether we can take advantage of the existing resources for spammer detection in microblogging. While people accept that texts in microblogging are different from those in other media, there is no quantitative analysis to show how different they are. In this paper, we first perform a comprehensive linguistic study to compare spam across different media. Inspired by the findings, we present an optimization formulation that enables the design of spammer detection in microblogging using knowledge from external media. We conduct experiments on real-world Twitter datasets to verify (1) whether email, SMS and web spam resources help and (2) how different media help for spammer detection in microblogging.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Classification*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

## General Terms

Algorithm, Performance, Experimentation

## Keywords

Spammer Detection, Twitter, Emails, SMS, Web, Cross-Media Mining, Social Media

*SIGIR'14,* July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609632 .

## 1. INTRODUCTION

Microblogging – a style of communicating through short-form content – has emerged as a popular social networking platform. Microblogging systems have been increasingly used for large-scale information dissemination and sharing in various fields such as marketing, journalism or public relations. With microblogging's growing popularity, activities of spamming have become rampant in launching various attacks in the medium. For example, the spammers spread ads to generate sales, disseminate pornography, viruses, phishing, or simply to compromise a system's reputation [3]. To improve user experience and the overall value of a system, it is essential to detect spammers in microblogging.

Existing methods for spammer detection in social media [26] focus on using social networking information. These network-based methods characterize the spammers by analyzing the network features, e.g., social status. The assumption behind this strategy is that it is difficult for the spammers to establish a large number of social relations with legitimate users. Different from other social media sites, in microblogging, users can follow anyone without prior consent from the followee. Many users just follow back when they are followed by someone for the sake of courtesy [20]. So, the spammers can easily enhance their influence score to fool the system. In this case, content analysis could complement network-based methods in spammer detection; thus, we explore the use of content information in this work.

A straightforward way to perform content-based spammer detection [22] is to model this task as a supervised learning problem. These methods extract effective textual features from the messages and build a classifier or a regressor based on the features. Given a new user, the built model can output a class label or score to determine whether it is a spammer based on microblogging messages the user posted. Content-based methods become difficult to be directly applied due to the distinct features of microblogging data. First, in microblogging, it is time-consuming and labor intensive to obtain labeled data, which is essential in building an effective supervised spammer detection model. Given the size and dynamic nature of microblogging, a manual labeling process is neither scalable nor sensible. Second, the texts in microblogging are short and noisy; thus, we lack sufficient aggregated information to evaluate the given messages. These present great challenges to directly making use of existing content-based methods for effective spammer detection in microblogging.

While the problem of spamming in microblogging is relatively new, it has been extensively studied for years in other platforms, e.g., email communication [4], SMS [14] and the web [31]. Similarly, the spammers in these platforms unfairly overwhelm other users by spreading unwanted information, which leads to phishing, malware, and scams [20]. Also, it has been reported in Natural Language Processing (NLP) literature that microblogging is not as noisy as was expected [2]. Although microblogging is an informal communication medium, it has been shown to be similar to other platforms [21] and it is seemingly possible to employ NLP tools to "clean" it [11]. Motivated by the previous findings, we explore the possibility of using knowledge learned from other platforms to facilitate spammer detection in the context of microblogging.

In this paper, we explore the use of resources available in other media to help spammer detection in microblogging. To study this problem, we need to answer the following questions: Are the resources from other media potentially helpful for spammer detection in microblogging? How do we explicitly model and make use of the resources from other media for spammer detection? Is the knowledge learned from other media helpful for microblogging spammer detection? By answering the above questions, this paper presents the following contributions:

- Conducting a quantitative analysis of linguistic variation of spam resources from different media;
- Formally defining the problem of leveraging knowledge across media for spammer detection in microblogging;
- Presenting a novel framework of leveraging knowledge from existing corpora to help spammer detection in microblogging; and
- Systematically evaluating the proposed method on real-world Twitter, email, SMS and web datasets and elaborating the effects of the knowledge learned from different media on spammer detection in Twitter.

The remainder of this paper is organized as follows. In Section 2, we conduct a quantitative study to examine the differences between spam corpora in different media from a linguistic perspective. In Section 3, we formally define the problem of leveraging knowledge across media for spammer detection in microblogging. In Section 4, we propose a novel framework for the problem we study. In Section 5, we report empirical results on real-world datasets. In Section 6, we review existing literature related to our work. In Section 7, we conclude this work and present some future work.

# 2. LINGUISTIC VARIATION ANALYSIS

This work is motivated by numerous spam resources available in other well-studied media, e.g., email, SMS and web. A natural question could be, given the short and noisy form of microblogging messages, how different are the texts in microblogging when compared to those in other media? Before proceeding further, we also examine whether the textual information from other media is potentially useful in the problem we study.

## 2.1 Datasets

Two Twitter datasets are used in our study for experiment purposes, i.e., TAMU Social Honeypots and Twitter Suspended Spammers. In addition, three representative datasets from different types of media, including Enron Email Dataset, SMS Dataset and Web Dataset, are used in the analysis. The statistics of the datasets are presented in Table 1. Now we introduce the datasets in detail.

**TAMU Social Honeypots Dataset (TweetH)**: Lee *et al.* [22] created a collection of 41,499 Twitter users with identity labels: spammers and legitimate users. The dataset was collected from December 30, 2009 to August 2, 2010 on Twitter. It consists of users, their number of followers and tweets. We filtered the non-English tweets and users with less than two tweets.

**Twitter Suspended Spammers Dataset (TweetS)**: We employed a data crawling process, which is similar to [32, 34], to construct this dataset. We first crawled a Twitter dataset from July to September 2012 via the Twitter Search API. The users that were suspended by Twitter during this period are considered as the gold standard [32] of spammers in the experiment. We then randomly sampled the legitimate users from a publicly available Twitter dataset provided by TREC 2011.[1] We filtered the non-English tweets and users with less than two tweets.

The first dataset TweetH has balanced number of spammers and legitimate users. To avoid effects brought by different class distribution, according to the literature of spammer detection [22], we made the two classes in TweetS imbalanced, i.e., the number of legitimate users is much greater than that of spammers in the dataset.

**Enron Email Dataset (Email)**: We used a subset of a widely used Enron email dataset,[2] which is collected during the investigation of Enron corporation and contains more than 200,000 emails between its employees. The emails in this dataset are preprocessed and used as a testbed in [25] for experiments. Each email in the dataset is labeled as either "spam" or "ham".

**SMS Dataset(SMS)**: We used the SMS spam collection provided by Almeida *et al.* [1] for analysis. This dataset is constructed based on two sources, Grumbletext web site[3] and NUS SMS Corpus.[4] The spam messages were manually labeled, and the ham messages were randomly sampled from the NUS SMS Corpus. To the best of our knowledge, this is the largest public SMS spam dataset.

**Web Dataset (Web)**: Web spam is a key challenge for internet users. Web pages which are created to deceive other users by manipulating search engine. Webb *et al.* [31] constructed the Web Dataset. This is the largest publicly available dataset to the best of our knowledge. We removed the web pages that have no textual content or only contain http request error information.

## 2.2 Lexical Analysis

To evaluate the style of a language, many metrics have been proposed in literature of linguistics and communication [2, 30]. In this subsection, we first introduce the metrics used in our study and then discuss lexical analysis results on the datasets from different media.

**Basic Statistics**: average Word Length (**WL**, in characters) and average Sentence Length (**SL**, in words) are used to evaluate the basic style of different datasets. In addition to those, we further employ other widely used lexical metrics in the analysis. We list the metrics below.

[1] http://trec.nist.gov/data/tweets/

[2] http://www.isi.edu/~adibi/Enron/Enron.htm

[3] http://www.grumbletext.co.uk/

[4] http://wing.comp.nus.edu.sg/SMSCorpus/

Table 1: Statistics of the Datasets

| | *TweetH* | *TweetS* | *Email* | *SMS* | *Web* |
|---|---|---|---|---|---|
| **# of Spam Messages** | 1,310,318 | 71,842 | 10,582 | 747 | 22,386 |
| **# of Legitimate Messages** | 1,220,198 | 308,957 | 13,990 | 4827 | N.A. |
| **# of Messages** | 2,530,516 | 380,799 | 24,572 | 5574 | 82,386 |
| **Avg. # of Words per Document** | 18.64 | 17.88 | 168.87 | 14.59 | 57.67 |

Table 2: Lexical Analysis Results

| | *Basics* | | *Lexical Analysis* | | |
|---|---|---|---|---|---|
| | WL | SL | TTR | LD | OOV |
| ***TweetH*** | 4.12 | 12.95 | 5.42 | 0.48 | 0.32 |
| ***TweetS*** | 3.95 | 12.38 | 5.65 | 0.50 | 0.31 |
| ***Email*** | 4.52 | 17.88 | 5.46 | 0.53 | 0.29 |
| ***SMS*** | 3.99 | 12.60 | 6.54 | 0.45 | 0.34 |
| ***Web*** | 4.81 | 18.66 | 6.13 | 0.48 | 0.32 |

Table 3: Hypothesis Testing Results (P-Values)

| | *TweetH* | | | *TweetS* | | |
|---|---|---|---|---|---|---|
| | TTR | LD | OOV | TTR | LD | OOV |
| ***Email*** | 0.318 | 0.108 | 0.442 | 0.234 | 0.267 | 0.308 |
| ***SMS*** | <0.01 | 0.205 | 0.350 | <0.01 | 0.082 | 0.163 |
| ***Web*** | <0.01 | 0.623 | 0.398 | 0.108 | 0.551 | 0.462 |

**Type-Token Ratio (TTR)**: This is a widely used metric to evaluate the difficulty (or readability) of words, sentences and documents by measuring their lexical variety [7, 33]. The basic assumption of using TTR is that difficult words are those that appear least often in a document. Given a corpus $D$, TTR is calculated as $TTR(D) = \sum_{w \in D} \frac{Freq(w)}{Size(D)}$, where $w$ means a word (token) in the corpus, $Freq(w)$ means word frequency of $w$ in $D$, and $Size(D)$ means the number of distinct words (types) in $D$. In practice, a higher TTR indicates a larger amount of lexical variation and a lower score indicates relatively less lexical variation [33].

**Lexical Density (LD)**: We employ lexical density to further analyze the stylistic difference between different corpora. Lexical words [15], also known as content or information carrying words, refer to verbs, nouns, adjectives and adverbs. Similarly, given a document $D$, LD is defined as $LD(D) = \sum_{w \in Lex} \frac{Freq(w)}{Size(D)}$, where $Lex$ means the whole lexical words dictionary. In general, a higher lexical density indicates that it is a more formal document, and a lower lexical density represents a more conversational one.

**Out-of-Vocabulary (OOV)**: This metric is to measure the ratio of out-of-vocabulary words in the corpora. We use a list of top 10,000 words with highest frequency provided by the Project Gutenberg [16] in our study. In general, a higher OOV rate indicates that the language is more informal. Many NLP and IR models suffer from high OOV rates.

Experimental results of the lexical analysis are presented in Table 2. By comparing the results of different metrics, we observe the following: (1) The word lengths of different corpora are very similar, and the sentence lengths of TweetH, TweetS and SMS are smaller than those of more formal media Email and Web. This indicates that the textual form of microblogging data is similar to SMS, and relatively different from email and web. (2) In most of the tests, microblogging data is similar to the datasets from the other media. It demonstrates that, although microblogging is considered an informal media, the language use is similar to that in other media, especially in email and SMS. We observe that the type-token ratios of microblogging are smaller than those of SMS and web. It suggests that the language used in microblogging is easier than that in the other two platforms.

We further employ hypothesis testing to examine the lexical differences between microblogging datasets and other datasets. For each lexical metric, we form a null hypothesis for a microblogging dataset and a dataset from the other media. The null hypothesis is: in terms of the specific lexical metric, there is no difference between microblogging data and data from the other media. We test the hypotheses on all pairs of the datasets for all the three lexical metrics.

In particular, to verify the difference between TweetH and Email datasets on the TTR, we construct two vectors $\mathbf{ttr}_{th}$ and $\mathbf{ttr}_{em}$. Each element of the first vector $\mathbf{ttr}_{th}$ is obtained by calculating the TTR score of a subset sampled with bootstrapping from TweetH dataset. Similarly, each element in the second vector corresponds to the TTR score of a subset sampled with bootstrapping from Email dataset. In the experiment, the two vectors contain equal number of elements.[5] Each element in the vectors corresponds to 100 data instances. We formulate a two-sample two-tail t-test on the two constructed vectors $\mathbf{ttr}_{th}$ and $\mathbf{ttr}_{em}$. We examine whether there is sufficient statistical evidence to support the hypothesis that the two datasets have the same sample mean, and it is defined as follows:

$$\begin{aligned} H_0 &: \mu_{th} - \mu_{em} = 0 \\ H_1 &: \mu_{th} - \mu_{em} \neq 0 \end{aligned} \tag{1}$$

where $H_0$ is the null hypothesis, $H_1$ is the alternative hypothesis, and $\mu_c$ and $\mu_r$ represent the sample means of the two vectors, respectively. Similarly, we form the hypothesis testings for other pairs of datasets with other lexical metrics.

The t-test results, p-values, are summarized in Table 3. From the table, we can observe the following: (1) With few exceptions, the results are much greater than the significance level $\alpha = 0.05$. It demonstrates that there is no statistical evidence to reject the null hypothesis in the tests on the two datasets. In other words, the results suggest that microblogging data is not significantly different from the datasets in other media. (2) In some tests, microblogging data appears more similar to Email than the other datasets.

In conclusion, while characteristics of different datasets appear different, there are no statistically significant lexical differences between them. The resources from other media are potentially useful in the task we study. Next, we formally define the problem we study and introduce the proposed learning framework for spammer detection.

[5] Note this is the setting used for experiment purposes, and it is not a mandatory setting for a two-sample t-test.

# 3. PROBLEM STATEMENT

In this section, we first present the notations and then formally define the problem we study.

**Notation:** lower-case bold Roman letters (e.g., $\mathbf{a}$) denote column vectors, upper-case letters (e.g., $\mathbf{A}$) denote matrices, and lowercase letters (e.g., a) denote scalars. $\mathbf{A}(i, j)$ denotes the entry at the $i^{th}$ row and $j^{th}$ column of a matrix $\mathbf{A}$. Let $\|\mathbf{A}\|$ denote the Euclidean norm, and $\|\mathbf{A}\|_F$ the Frobenius norm of the matrix $\mathbf{A}$. Specifically, $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{A}(i,j)^2}$. Let $\mathbf{A}^T$ and $Tr(\mathbf{A})$ denote the transpose and trace of $\mathbf{A}$, respectively.

Let $\mathbf{S} = [\mathbf{X}, \mathbf{Y}]$ be available resources from other media, with the content information $\mathbf{X}$ and identity label matrix $\mathbf{Y}$. We use term-user matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ to denote content information, i.e., posts written by the users, where $m$ is the number of textual features, and $d$ is the number of users in the other media. $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_r\}$ means the combination of content information from multiple media, and $\mathbf{Y} \in \mathbb{R}^{d \times c} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_r\}$ means the combination of label information from the media. For each user $(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^{m+c}$ consists of message content and identity label, where $\mathbf{x}_i \in \mathbb{R}^m$ is the message feature vector and $\mathbf{y}_i \in \mathbb{R}^c$ is the spammer label vector. In this paper, we consider the task we study as a two-class classification problem, i.e., $c = 2$. For example, $\mathbf{y}_i = (1, 0)$ means this user is a spammer. $\mathbf{y}_i^T \mathbf{y}_i = 1$ constrains that $\mathbf{y}_i$ has to have one label and cannot be $(0, 0)$ or $(1, 1)$. It is practical to extend this setting to a multi-class or regression problem. We use $\mathbf{T} \in \mathbb{R}^{m \times n}$ to denote the content information of microblogging users, where $m$ is the number of textual features, and $n$ is the number of users in microblogging. The texts from microblogging and other media share the same feature space.

We now formally define the problem as follows:

*We have a set of resources* $\mathbf{S}$ *from different media, with the content information* $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_r\}$ *and identity label information* $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_r\}$. *Given the content information* $\mathbf{T}$ *from microblogging, our goal is to automatically infer the identity labels for unknown users in* $\mathbf{T}$ *as spammers or legitimate users.*

# 4. LEVERAGING KNOWLEDGE ACROSS MEDIA FOR SPAMMER DETECTION

We plot the work flow of our proposed framework in Figure 1. From the figure, we see that there are two constraints on the learned model for spammer detection. As shown in the upper right part of the figure, the first constraint is from the lexicon information $\mathbf{U}$, which is learned from the other media sources $\mathbf{S}$. As shown in the lower right part of the figure, the second constraint is a Laplacian regularization $\mathbf{M}$ learned from microblogging content information. We now introduce each part of the proposed framework in detail.

## 4.1 Modeling Knowledge across Media

As we discussed in the last section, from a linguistic perspective, it does not show significant difference between microblogging data and other types of data. A straightforward method to make use of external information is to learn a supervised model based on data from the other media, and apply the learned classifier on microblogging data for spammer detection. However, this method yields two problems to be directly applied to our task. First, text representation models, like n-gram model, often lead to a high-dimensional feature space because of the large size of data and vocabulary. Second, texts in the media are short, thus making the data representation very sparse [21].

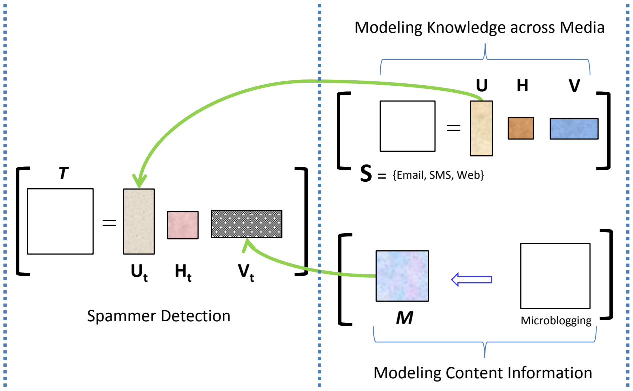

**Figure 1: Illustration of the Proposed Spammer Detection Framework**

To tackle the problems, instead of learning knowledge at word-level, we propose to capture the external knowledge from topic-level. In particular, the proposed method is built on the orthogonal nonnegative matrix tri-factorization model (ONMTF) [9]. The basic idea of the ONMTF model is to cluster data instances based on distribution of features, and cluster features according to the distribution of data instances. The principle of ONMTF is consistent with PLSI [17], in which each document is a mixture of latent topics that each word can be generated from. The ONMTF can be formulated by optimizing:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{H},\mathbf{V} \geq 0} \quad & \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2, \\ s.t. \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I}, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{2}$$

where $\mathbf{X}$ is the content matrix, and $\mathbf{U} \in \mathbb{R}_+^{m \times c}$ and $\mathbf{V} \in \mathbb{R}_+^{d \times c}$ are nonnegative matrices indicating low-dimensional representations of words and users, respectively. $m$ is the size of vocabulary, $c$ is the number of classes, $d$ is the number of users. $\mathbf{H} \in \mathbb{R}_+^{c \times c}$ provides a condensed view of $\mathbf{X}$. The orthogonal and nonnegative conditions of $\mathbf{U}$ and $\mathbf{V}$ provide a hard assignment of class label to the words and users.

With the ONMTF model, we project the original content information from the other media into a latent topic space. By adding a topic-level least squares penalty to the ONMTF, our proposed framework can be mathematically formulated as solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{H},\mathbf{V},\mathbf{W} \geq 0} \quad & \mathcal{J} = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2 + \lambda\|\mathbf{V}\mathbf{W} - \mathbf{Y}\|_F^2, \\ s.t. \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I}, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{3}$$

where $\mathbf{W}$ represents the weights and $\mathbf{Y}$ is the label matrix. In the formulation, the first term is the basic factorization model, and the second introduces label information from the other media by using a linear penalty. $\lambda$ is to control the effect of external information to the learned lexicon $\mathbf{U}$, in which each row represents the predicted label of a word.

As the problem in Eq. (3) is not convex with respect to the four variables together, there is no closed-form solution for the problem. Next, we introduce an alternative scheme to solve the optimization problem.

### 4.1.1 Optimization Algorithm

Following [9], we propose to optimize the objective with respect to one variable, while fixing others. The algorithm will keep updating the variables until convergence.

**Computation of H**: Optimizing the objective function in Eq. (3) with respect to $\mathbf{H}$ is equivalent to solving

$$\min_{\mathbf{H}\geq 0} \quad \mathcal{J}_H = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2. \tag{4}$$

Let $\Lambda_H$ be the Lagrange multiplier for constraint $\mathbf{H} \geq 0$; the Lagrange function $L(\mathbf{H})$ is defined as follows:

$$L(\mathbf{H}) = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2 - Tr(\Lambda_H \mathbf{H}^T). \tag{5}$$

By setting the derivative $\nabla_{\mathbf{H}} L(\mathbf{H}) = 0$, we get

$$\Lambda_H = -2\mathbf{U}^T\mathbf{X}\mathbf{V} + 2\mathbf{U}^T\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}. \tag{6}$$

The Karush-Kuhn-Tucker complementary condition [6] for the nonnegativity constraint of $\mathbf{H}$ gives

$$\Lambda_H(i,j)\mathbf{H}(i,j) = 0 \ ; \tag{7}$$

thus, we obtain

$$[-\mathbf{U}^T\mathbf{X}\mathbf{V} + \mathbf{U}^T\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}](i,j)\mathbf{H}(i,j) = 0. \tag{8}$$

Similar to [9], it leads to the updating rule of $\mathbf{H}$,

$$\mathbf{H}(i,j) \leftarrow \mathbf{H}(i,j)\sqrt{\frac{[\mathbf{U}^T\mathbf{X}\mathbf{V}](i,j)}{[\mathbf{U}^T\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}](i,j)}}. \tag{9}$$

**Computation of U**: Optimizing the objective function in Eq. (3) with respect to $\mathbf{U}$ is equivalent to solving

$$\begin{aligned} \min_{\mathbf{U}\geq 0} \quad & \mathcal{J}_U = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2 \\ s.t. \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{10}$$

Let $\Lambda_U$ and $\Gamma_U$ be the Lagrange multipliers for constraints $\mathbf{U} \geq 0$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, respectively; the Lagrange function $L(\mathbf{U})$ is defined as follows:

$$L(\mathbf{U}) = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2 - Tr(\Lambda_U \mathbf{U}^T) + Tr(\Gamma_U(\mathbf{U}^T\mathbf{U} - \mathbf{I})) \tag{11}$$

By setting the derivative $\nabla_{\mathbf{U}} L(\mathbf{U}) = 0$, we get

$$\Lambda_U = -2\mathbf{X}\mathbf{V}\mathbf{H}^T + 2\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + 2\mathbf{U}\Gamma_U. \tag{12}$$

With the KKT complementary condition for the nonnegativity constraint of $\mathbf{U}$, we have

$$\Lambda_U(i,j)\mathbf{U}(i,j) = 0; \tag{13}$$

thus, we obtain

$$[-\mathbf{X}\mathbf{V}\mathbf{H}^T + \mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + \mathbf{U}\Gamma_U](i,j)\mathbf{U}(i,j) = 0, \tag{14}$$

where

$$\Gamma_U = \mathbf{U}^T\mathbf{X}\mathbf{V}\mathbf{H}^T - \mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T. \tag{15}$$

Let $\Gamma_U = \Gamma_U^+ - \Gamma_U^-$, where $\Gamma_U^+(i,j) = (|\Gamma_U(i,j)| + \Gamma_U(i,j))/2$ and $\Gamma_U^-(i,j) = (|\Gamma_U(i,j)| - \Gamma_U(i,j))/2$ [9]; we get

$$[-(\mathbf{X}\mathbf{V}\mathbf{H}^T + \mathbf{U}\Gamma_U^-) + (\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + \mathbf{U}\Gamma_U^+)](i,j)\mathbf{U}(i,j) = 0, \tag{16}$$

which leads to the updating rule of $\mathbf{U}$,

$$\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j)\sqrt{\frac{[\mathbf{X}\mathbf{V}\mathbf{H}^T + \mathbf{U}\Gamma_U^-](i,j)}{[\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + \mathbf{U}\Gamma_U^+](i,j)}}. \tag{17}$$

**Algorithm 1:** *Modeling Knowledge across Media*

```
Input: $\{\mathbf{X}, \mathbf{Y}, \lambda, I\}$
Output: $\mathbf{V}$
1: Initialize $\mathbf{U}, \mathbf{V}, \mathbf{H}, \mathbf{W} \geq 0$
2: while Not convergent and iter $\leq$ I do
3:     Update $\mathbf{H}(i,j) \leftarrow \mathbf{H}(i,j)\sqrt{\frac{[\mathbf{U}^T\mathbf{X}\mathbf{V}](i,j)}{[\mathbf{U}^T\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}](i,j)}}$
4:     Update $\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j)\sqrt{\frac{[\mathbf{X}\mathbf{V}\mathbf{H}^T+\mathbf{U}\Gamma_U^-](i,j)}{[\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T+\mathbf{U}\Gamma_U^+](i,j)}}$
5:     Update $\mathbf{V}(i,j) \leftarrow \mathbf{V}(i,j)\sqrt{\frac{[\mathbf{X}^T\mathbf{U}\mathbf{H}+\lambda\mathbf{Y}\mathbf{W}^T+\mathbf{V}\Gamma_V^-](i,j)}{[\mathbf{V}\mathbf{H}^T\mathbf{U}^T\mathbf{U}\mathbf{H}+\lambda\mathbf{V}\mathbf{W}\mathbf{W}^T+\mathbf{V}\Gamma_V^+](i,j)}}$
6:     Update $\mathbf{W}(i,j) \leftarrow \mathbf{W}(i,j)\sqrt{\frac{[\mathbf{V}^T\mathbf{Y}](i,j)}{[\mathbf{V}^T\mathbf{V}\mathbf{W}](i,j)}}$
7:     $iter = iter + 1$
8: end while
```

**Computation of V**: Optimizing the objective function in Eq. (3) with respect to $\mathbf{V}$ is equivalent to solving

$$\begin{aligned} \min_{\mathbf{V}\geq 0} \quad & \mathcal{J} = \|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\|_F^2 + \lambda\|\mathbf{V}\mathbf{W} - \mathbf{Y}\|_F^2 \\ & s.t. \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}. \end{aligned} \tag{18}$$

Similar to the computation of $\mathbf{U}$, by introducing two Lagrange multipliers $\Lambda_V$ and $\Gamma_V$ for the constraints, we get

$$\begin{aligned} &[-(\mathbf{X}^T\mathbf{U}\mathbf{H} + \lambda\mathbf{Y}\mathbf{W}^T + \mathbf{V}\Gamma_V^-) \\ &+ (\mathbf{V}\mathbf{H}^T\mathbf{U}^T\mathbf{U}\mathbf{H} + \lambda\mathbf{V}\mathbf{W}\mathbf{W}^T + \mathbf{V}\Gamma_V^+)](i,j)V(i,j) = 0, \end{aligned} \tag{19}$$

which leads to the updating rule of $\mathbf{V}$,

$$\mathbf{V}(i,j) \leftarrow \mathbf{V}(i,j)\sqrt{\frac{[\mathbf{X}^T\mathbf{U}\mathbf{H} + \lambda\mathbf{Y}\mathbf{W}^T + \mathbf{V}\Gamma_V^-](i,j)}{[\mathbf{V}\mathbf{H}^T\mathbf{U}^T\mathbf{U}\mathbf{H} + \lambda\mathbf{V}\mathbf{W}\mathbf{W}^T + \mathbf{V}\Gamma_V^+](i,j)}} \tag{20}$$

**Computation of W**: Optimizing the objective function in Eq. (3) with respect to $\mathbf{W}$ is equivalent to solving

$$\min_{\mathbf{W}\geq 0} \quad \mathcal{J} = \|\mathbf{V}\mathbf{W} - \mathbf{Y}\|_F^2. \tag{21}$$

Similar to the computation of $\mathbf{U}$, by introducing a Lagrange multiplier and satisfying KKT condition, we obtain

$$[\mathbf{V}^T\mathbf{V}\mathbf{W} - \mathbf{V}^T\mathbf{Y}](i,j)\mathbf{W}(i,j) = 0, \tag{22}$$

which leads to the updating rule of $\mathbf{W}$,

$$\mathbf{W}(i,j) \leftarrow \mathbf{W}(i,j)\sqrt{\frac{[\mathbf{V}^T\mathbf{Y}](i,j)}{[\mathbf{V}^T\mathbf{V}\mathbf{W}](i,j)}}. \tag{23}$$

We summarize the algorithm of optimizing Eq. (3) in Algorithm 1, where $I$ is the number of maximum iterations. In line 1, we conduct initialization for the variables. From lines 2 to 8, the four variables are updated with the updating rules until convergence or until they reach the number of maximum iterations. The correctness and convergence of the updating rules can be proven with the standard auxiliary function approach [28].

## 4.2 Modeling Content Information

In this subsection, as shown in the lower right part of Figure 1, we introduce how to model content information of microblogging data in the proposed model.

To make use of the content information of microblogging messages, we introduce a graph Laplacian [8] in the proposed model. We construct a graph based on content information of the users. In the graph, each node represents a user and each edge represents the affinity between two users. The adjacency matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ of the graph is defined as

$$\mathbf{M}(u, v) = \begin{cases} 1 & \text{if } \; u \in \mathcal{N}(v) \; or \; v \in \mathcal{N}(u) \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

where $u$ and $v$ are nodes, and $\mathcal{N}(u)$ represents the k-nearest neighbor of the user. Content similarity is adopted to obtain the k-nearest neighbor in this work. Since we aim to model the mutual content similarity between two users, the adjacency matrix is symmetric.

The basic idea of of using the graph Laplacian to model the content information is that if two nodes are close in the graph, i.e., they posted similar messages, their identity labels should be close to each other. It can be mathematically formulated as minimizing the following loss function:

$$\mathcal{R} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\mathbf{V}_t(i, *) - \mathbf{V}_t(j, *)\|_2^2 \mathbf{M}(i, j). \tag{25}$$

This loss function will incur a penalty if two users have different predicted labels when they are close to each other in the graph. Let $\mathbf{D} \in \mathbb{R}^{n \times n}$ denote a diagonal matrix, and its diagonal element is the degree of a user in the adjacency matrix $\mathbf{M}$, i.e., $\mathbf{D}(i, i) = \sum_{j=1}^{n} \mathbf{M}(i, j)$.

Theorem 1. *The formulation in Eq. (25) is equivalent to the following objective function:*

$$\mathcal{R} = Tr(\mathbf{V}_t^T \mathcal{L} \mathbf{V}_t), \tag{26}$$

*where the Laplacian matrix [8] $\mathcal{L}$ is defined as $\mathcal{L} = \mathbf{D} - \mathbf{M}$.*

*Proof.* It is easy to verify that Eq. (25) can be rewritten as

$$\begin{aligned} \mathcal{R} = & \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{c} \mathbf{V}_t(i, k) \mathbf{M}(i, j) \mathbf{V}_t^T(i, k) \\ & - \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{c} \mathbf{V}_t(i, k) \mathbf{M}(i, j) \mathbf{V}_t^T(j, k) \\ = & \; Tr(\mathbf{V}_t^T (\mathbf{D} - \mathbf{M}) \mathbf{V}_t) \\ = & \; Tr(\mathbf{V}_t^T \mathcal{L} \mathbf{V}_t), \end{aligned} \tag{27}$$

which completes the proof. □

### 4.3 Spammer Detection Framework

As illustrated in Figure 1, we employ two types of information to formulate two kinds of constraints on the learned model. By integrating knowledge learned from other media and content information from microblogging, we can perform spammer detection by optimizing

$$\begin{aligned} \min_{\mathbf{U}_t, \mathbf{H}_t, \mathbf{V}_t \geq 0} \quad & \mathcal{J} = \|\mathbf{T} - \mathbf{U}_t \mathbf{H}_t \mathbf{V}_t^T\|_F^2 + \alpha Tr(\mathbf{V}_t^T \mathcal{L} \mathbf{V}_t) \\ & + \beta \|\mathbf{G}_U (\mathbf{U}_t - \mathbf{U})\|_F^2), \\ & s.t. \;\; \mathbf{U}_t^T \mathbf{U}_t = \mathbf{I}, \;\; \mathbf{V}_t^T \mathbf{V}_t = \mathbf{I}, \end{aligned} \tag{28}$$

where the first term is to factorize the microblogging data into three variables, which are similar to the idea discussed in Section 4.1. The second term is to introduce content information and the third is to introduce knowledge learned from the other media. $\mathbf{U}$ is the lexicon learned from the other media by solving the problem in Eq. (3). $\mathbf{G}_U \in \{0, 1\}^{m \times m}$ is a diagonal indicator matrix to control the impact of the learned lexicon, i.e., $\mathbf{G}_U(i, i) = 1$ represents that the $i$-th word contains identity information, $\mathbf{G}_U(i, i) = 0$ otherwise.

**Algorithm 2:** *Spammer Detection in Microblogging*

```
Input: $\{\mathbf{T}, \mathbf{U}, \alpha, \beta, I\}$
Output: $\mathbf{V}_t$
1: Construct matrices $\mathcal{L}$ in Eq. (26)
2: Initialize $\mathbf{U}_t = \mathbf{U}, \mathbf{V}, \mathbf{H} \geq 0$
3: while Not convergent and iter $\leq$ I do
4:     Update $\mathbf{H}_t(i,j) \leftarrow \mathbf{H}_t(i,j)\sqrt{\frac{[\mathbf{U}_t^T\mathbf{X}\mathbf{V}_t](i,j)}{[\mathbf{U}_t^T\mathbf{U}_t\mathbf{H}_t\mathbf{V}_t^T\mathbf{V}_t](i,j)}}$
5:     Update
       $\mathbf{U}_t(i,j) \leftarrow \mathbf{U}_t(i,j)\sqrt{\frac{[\mathbf{X}\mathbf{V}_t\mathbf{H}_t^T+\beta\mathbf{G}_U\mathbf{U}+\mathbf{U}_t\Gamma_U^-](i,j)}{[\mathbf{U}_t\mathbf{H}_t\mathbf{V}_t^T\mathbf{V}_t\mathbf{H}_t^T+\beta\mathbf{G}_U\mathbf{U}_t+\mathbf{U}_t\Gamma_U^+](i,j)}}$
6:     Update
7:     $\mathbf{V}_t(i,j) \leftarrow \mathbf{V}_t(i,j)\sqrt{\frac{[\mathbf{X}^T\mathbf{U}_t\mathbf{H}_t+\alpha\mathbf{M}\mathbf{V}_t+\mathbf{V}_t\Gamma_V^-](i,j)}{[\mathbf{V}_t\mathbf{H}_t^T\mathbf{U}_t^T\mathbf{U}_t\mathbf{H}_t+\alpha\mathbf{D}\mathbf{V}_t+\mathbf{V}_t\Gamma_V^+](i,j)}}$
8:     $iter = iter + 1$
9: end while
```

This optimization problem is not convex with respect to the three parameters together. Following the optimization procedure to solve Eq. (3), we propose an algorithm to solve the problem in Eq. (28) and summarize it in Algorithm 2. In line 1, we construct the Laplacian matrix $\mathcal{L}$. In line 2, we initialize the variables. From lines 3 to 9, we keep updating the variables with the updating rules until convergence or until the number of maximum iterations is reached.

## 5. EXPERIMENTS

In this section, we empirically evaluate the proposed learning framework and the factors that could bring in effects to the framework. Through the experiments, we aim to answer the following two questions:

- How effective is the proposed framework compared with other possible solutions of using external information across media in real-world spammer detection tasks?
- What impact do the other resources have on the performance of spammer detection in microblogging?

### 5.1 Experimental Setup

We follow a standard experiment setup used in spammer detection literature [34] to evaluate the effectiveness of our proposed framework for leveraging knowledge a<u>C</u>ross media for <u>S</u>pammer <u>D</u>etection (*CSD*). In particular, we compare the proposed framework *CSD* with different baseline methods for spammer detection. To avoid bias, both TweetH and TweetS, introduced in Section 2.1, are used in the experiments. For email data, we consider each sender a user; For SMS and web data, we do not have user information and consider each message as sent from a distinct user. In the experiment, precision, recall and $F_1$-measure are used as the performance metrics.

To evaluate the general performance of the proposed framework, we use all of the three datasets from different media, i.e., Email, SMS and Web datasets. In the first set of experiments, to be discussed in Section 5.2, we simply combine them together and consider them as homogeneous data sources. In the second set of experiments, to be discussed in

**Table 4: Spammer Detection Results on TweetH Dataset**

| | *External Data I (50%)* | | | *External Data II (100%)* | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$-measure (gain) | Precision | Recall | $F_1$-measure (gain) |
| *Least_Squares* | 0.823 | 0.834 | 0.828 (N.A.) | 0.839 | 0.852 | 0.845 (N.A.) |
| *Lasso* | 0.865 | 0.891 | 0.878 (+5.96%) | 0.873 | 0.905 | 0.889 (+5.12%) |
| *MFTr* | 0.866 | 0.899 | 0.882 (+6.49%) | 0.887 | 0.918 | 0.902 (+6.72%) |
| *MFSD* | 0.644 | 0.703 | 0.672 (-18.7%) | 0.650 | 0.715 | 0.681 (-19.5%) |
| *CSD* | 0.906 | 0.939 | 0.922 (+11.3%) | 0.913 | 0.944 | 0.928 (+9.79%) |

**Table 5: Spammer Detection Results on TweetS Dataset**

| | *External Data I (50%)* | | | *External Data II (100%)* | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$-measure (gain) | Precision | Recall | $F_1$-measure (gain) |
| *Least_Squares* | 0.766 | 0.813 | 0.789 (N.A.) | 0.793 | 0.820 | 0.806 (N.A.) |
| *Lasso* | 0.801 | 0.849 | 0.824 (+4.50%) | 0.814 | 0.848 | 0.831 (+3.02%) |
| *MFTr* | 0.810 | 0.857 | 0.833 (+5.58%) | 0.833 | 0.878 | 0.855 (+6.03%) |
| *MFSD* | 0.621 | 0.69 | 0.654 (-17.1%) | 0.642 | 0.681 | 0.661 (-18.0%) |
| *CSD* | 0.832 | 0.875 | 0.853 (+8.13%) | 0.848 | 0.919 | 0.882 (+9.40%) |

Section 5.3, we consider their individual impact on the performance of spammer detection. A standard procedure for data preprocessing is used in our experiments. The unigram model is employed to construct the feature space, tf-idf is used as the feature weight.

As we discussed in Sections 4.1 and 4.3, three positive parameters are involved in the experiments, including $\lambda$ in Eq. (3), and $\alpha$ and $\beta$ in Eq. (28). $\lambda$ is to control the effect of knowledge from other media to the learned lexicon, $\alpha$ is to control the contribution of Laplacian regularization, and $\beta$ is to control the contribution of lexicon to the spammer detection model. Since all the parameters can be tuned via cross-validation with a set of validation data, in the experiment, we empirically set $\lambda = 0.1$, $\alpha = 0.1$ and $\beta = 0.1$ for general experiment purposes. The effects of the parameters on the learning model will be further discussed in Section 5.4.

## 5.2 Performance Evaluation

We compare the proposed method *CSD* with other methods for spammer detection, accordingly answer the first question asked above. The baseline methods are listed below.

- *Least_Squares*: One possible solution for our task is to consider it as a supervised learning problem. We simply train a classification model with the available external data and apply the learned model on microblogging data for spammer detection. The widely used classifier, *Least_Squares* [12], is used for comparison.

- *Lasso*: Sparse learning methods are effective for high-dimensional data in social media. We further include *Lasso* [29] as the baseline method, which performs continuous shrinkage and automatic feature selection by adding $l_1$ norm regularization to the Least Squares.

- *MFTr*: Although we first present a quantitative linguistic variation analysis and provide a unified model for spammer detection across different media, domain adaption and transfer learning have received great attention in various applications [27]. We apply a widely used transfer learning method [23], which transfers the knowledge directly from labeled data in the source domain to the target domain for classification, to test its performance on spammer detection in the experiment.

- *MFSD*: We test the performance of the unsupervised learning method by employing the basic matrix factorization model *MFSD*. This is a variant of our proposed method without introducing any knowledge learned from external sources. As a common initialization for clustering methods, we randomly assign initial centroids and an initial class indicator matrix for *MFSD*.

Experimental results of the methods on the two datasets, TweetH and TweetS, are respectively reported in Table 4 and 5. To avoid bias brought by the sizes of the training data,[6] we conduct two sets of experiments with different numbers of training instances. In the experiments, "External Data I (50%)" means that we randomly chose 50% from the whole training data. "External Data II (100%)" means that we use all the data for training. Also, "gain" represents the percentage improvement of the methods in comparison with the first baseline method *Least_Squares*. In the experiment, each result denotes an average of 10 test runs. By comparing the spammer detection performance of different methods, we observe the following:

(1) From the results in the tables, we can observe that our proposed method *CSD* consistently outperforms other baseline methods on both datasets with different sizes of training data. Our method achieves better results than the state-of-the-art method *MFTr* on both datasets. We apply two-sample one-tail t-tests to compare *CSD* to the four baseline methods. The experiment results demonstrate that the proposed model performs significantly better (with significance level $\alpha = 0.01$) than the four methods.

(2) The performance of our proposed method *CSD* is better than the first three baselines, which are based on different strategies of using resources from the other media. This demonstrates the excellent use of cross-media knowledge in the proposed framework for spammer detection.

[6] Similar to the definitions in machine learning literature, training data here refers to the labeled data from the external sources, and testing data represents the unlabeled microblogging data.

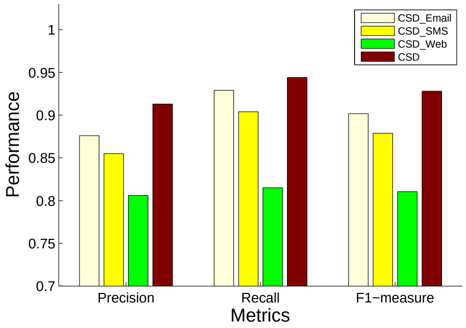

**Figure 2: Results on TweetH Dataset**

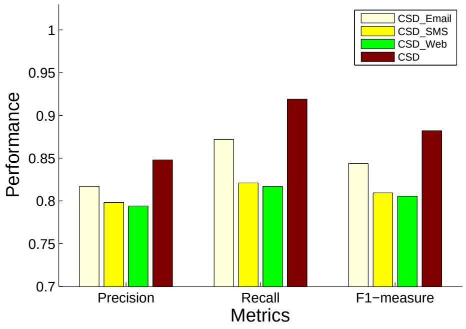

**Figure 3: Results on TweetS Dataset**

(3) Among the baseline methods, *MFTr* achieves the best results. It demonstrates that the knowledge transferred from other media help the task of spammer detection in microblogging. *Lasso* performs better than *Least_Squares*. This shows that, for high-dimensional textual data from email, SMS and web, feature selection is necessary for a supervised learning method for this task we study.

(4) The method *MFSD* achieves the worst performance among all the baseline methods. It shows that learning based on microblogging data itself can not discriminant well between spammers and legitimate users. It further demonstrates that the knowledge learned from external sources is helpful to build an effective model to tackle the problem.

In summary, with the effective use of data from the other media, our proposed framework outperforms the baseline methods in spammer detection. Next, we investigate the effects of different resources on the spammer detection task.

## 5.3 Effects of External Information

In this subsection, we study the effects of the external information from the other media on our proposed framework, accordingly answering the second question asked in the beginning of Section 5.

We first evaluate the performance of the proposed framework with data from only one of the three media. In particular, we learn a lexicon based on one of the three types of media, i.e., email, SMS and web, and perform spammer detection on the microblogging datasets. We do not have legitimate web pages in the original Web dataset. To build a classifier *CSD_Web*, following the data construction procedure proposed in [18], we randomly sample 20,100 web snippets with BingAPI as legitimate data. The experimental results of the methods on the two microblogging datasets are plotted in Figures 2 and 3, respectively. In the figures, the first three bars represent the performance of the baselines with one type of external information. The last is the method with all three types of external information. From the figures, we observe the following:

(1) With the integration of all three types of external information, *CSD* consistently achieves better performance than the three baselines with only one type of information. It demonstrates that the proposed method uses beneficial information to perform effective spammer detection.

(2) Among the three baseline methods, *CSD_Email* and *CSD_SMS* achieve better performance than *CSD_Web*. It shows that, as external resources, email and SMS data are more suitable to be used for the spammer detection in microblogging than the web data. This result is consistent with the linguistic variation analysis in Section 2.

To further explore the effects of different media sources on the performance of spammer detection in microblogging, we employ a "knockout" technique in the experiment. Knockout has been widely used in many fields, e.g., gene function analysis, to test the performance variance brought by one component when it is made inoperative in the framework [10]. We conduct the experiments by knocking out one type of the external information from the proposed framework. The results are summarized in Table 6. In the table, "loss" represents the performance decrease of the methods as compared to the setting "Default" which is learned based on data from all three media sources. The three columns in the middle are experimental settings, in which "0" means this resource is knocked out. The last two columns are the F$_1$-measure results under different experimental settings. From the table, we observe the following:

(1) By knocking out one of the external sources, performance of the proposed framework decreases. This suggests that all the three types of external information are useful for spammer detection in microblogging.

(2) Knocking out email from the resources incurs the most performance decrease among all the experimental settings. This demonstrates that email is the most effective source among the three types of information. This finding is consistent with our discussion above.

In summary, the use of data from the other media shows the effectiveness in spammer detection task. The superior performance of the proposed method *CSD* validates its excellent use of knowledge from the other media.

## 5.4 Parameter Analysis

As discussed in Section 5.1, three positive parameters, i.e., $\lambda$, $\alpha$ and $\beta$, are involved in the proposed framework. We first examine the effects brought by $\lambda$, which is to control the contribution of knowledge from other media to the learned lexicon. In previous subsections, for general experimental purposes, we empirically set $\lambda = 0.1$. We now conduct experiments to compare the spammer detection performance of the four methods introduced in Section 5.3 with different settings of $\lambda$. The experiment results on the TweetH dataset are plotted in Figure 4. From the figure, we observe the following: (1) The general trends of the four methods are similar with the variation of different parameter settings. They achieve relatively good performance when setting $\lambda$ in the range of [0.1, 10]. (2) In most cases, performance of

Table 6: Learning from Different Media for Spammer Detection in Microblogging

| | Email | SMS | Web | TweetH (loss) | TweetS (loss) |
|---|---|---|---|---|---|
| Default | 1 | 1 | 1 | 0.928 (N.A.) | 0.882 (N.A.) |
| Knock Out One Term | 0 | 1 | 1 | 0.881 (-5.09%) | 0.843 (-4.43%) |
| | 1 | 0 | 1 | 0.911 (-1.86%) | 0.856 (-2.96%) |
| | 1 | 1 | 0 | 0.923 (-0.57%) | 0.860 (-2.50%) |

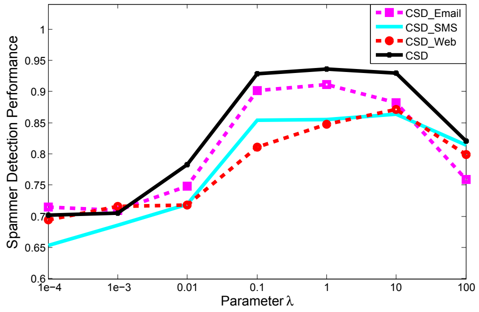

**Figure 4: Performance with Different $\lambda$ Settings**

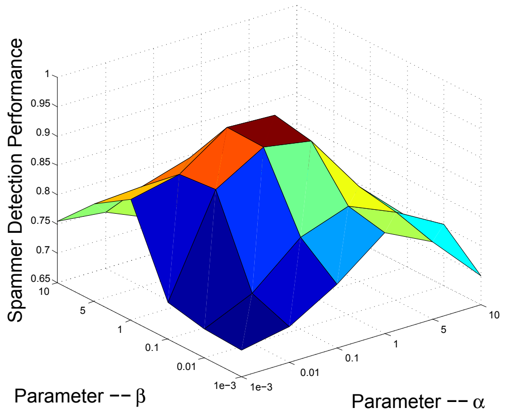

**Figure 5: Impact of Content Information ($\alpha$) and External Information ($\beta$)**

the proposed *CSD* is better than the other three methods. It demonstrates that the combination of the three resources improve the spammer detection performance.

We further examine the effects of the parameters $\alpha$ and $\beta$ discussed in Eq. (28) on the proposed framework. $\alpha$ is to control the contribution of content information and $\beta$ is to control the effects of external information from the other media. To understand the effects brought by the parameters, we compare the spammer detection performance of the proposed *CSD* on the Twitter datasets with different parameter settings. The results on the TweetH dataset are plotted in Figure 5. From the figure, we observe that the proposed method *CSD* performs well when $\alpha \in [0.1, 5]$ and $\beta \in [0.1, 1]$. Generally, the performance of *CSD* is not quite sensitive to the parameters. The proposed framework can perform well when choosing parameter settings in a reasonable range. Similar results have been observed for the two sets of experiments on the TweetS dataset; we omit the results owing to lack of space.

# 6. RELATED WORK

Significant efforts have been devoted to detecting spammers in various online social networks, including Facebook [5], Twitter [19, 20, 22], Renren [32], Blogosphere [24], etc. One effective way to perform spammer detection is to use the social network information. The assumption is that spammers cannot establish a large number of social trust relations with normal users. This assumption might not hold in many social networks. Yang *et al.* [32] studied the spammers in Renren, and found that spammers can have their friend requests accepted by many other users and thus blend into the Renren social graph. Different from Facebook-like OSNs, microblogging systems feature unidirectional user bindings because anyone can follow anyone else without prior consent from the followee. Ghosh *et al.* [13] show that spammers can acquire many legitimate followers. Besides the methods based on social networks, some efforts [22] have also been devoted to study characteristics related to tweet content and user social behavior. By understanding spammer activities in social networks, features are extracted to perform effective spammer detection. These methods need a large amount of labeled data, which is hard to obtain in social media.

Spammer detection on emails [4], SMS [14] and the web [31] has been a hot topic for quite a few years. The spams are designed to corrupt the user experience by spreading ads or driving traffic to particular web sites [31]. A popular and well-developed approach for anti-spam applications is learning-based filtering. The basic idea is that we extract effective features from the labeled data and build a classifier. We then classify new users / messages as either spam or ham according to their content information for filtering. The attempts have been done in these areas and the abundant labeled resources are the major motivation of our work.

Some efforts have been made to employ domain adaption and transfer learning in various applications, e.g. sentiment analysis [23] and text classification [27]. Our work started the investigation of leveraging knowledge from other media for spammer detection in microblogging. Different from traditional methods, based on the quantitatively linguistic variation analysis, our proposed framework naturally combines knowledge learned from internal and external data sources in a unified model. In addition, some work has been done to study the linguistic challenges of social media texts. It is accepted that texts in social media are noisy, but it is also reported by researchers that the texts are not as noisy as what people expected [2]. The language used in Twitter is more like a projection of the language of formal media like news and blogs with shorter form [21], and it is possible to make use of normalization and domain adaption to "clean" it [11]. The evidence provided by linguists also motivate us to explore the language differences of spams across different media, and make use of resources from other media to help spammer detection in microblogging.

## 7. CONCLUSIONS AND FUTURE WORK

Texts in microblogging are short, noisy, and labeling processing is time-consuming and labor-intensive, which presents great challenges for spammer detection. In this paper, we first conduct a quantitative analysis to study how noisy the microblogging texts are by comparing them with spam messages from other media. The results suggest that microblogging data is not significantly different from data from the other media. Based on the observations, a matrix factorization model is employed to learn lexicon information from external spam resources. By incorporating external information from other media and content information from microblogging, we propose a novel framework for spammer detection. The experimental results demonstrate the effectiveness of our proposed model as well as the roles of different types of information in spammer detection.

This work suggests some interesting future directions. Different types of medium resources have different effects on the spammer detection performance. It would be interesting to quantify the contributions of different types of sources to spammer detection in microblogging. This could be an important support for source selection in spammer detection.

## Acknowledgments

We truly thank the anonymous reviewers for their pertinent comments. This work is, in part, supported by ONR (N000141410095) and ARO (#025071).

## 8. REFERENCES

[1] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of DocEng*, 2011.

[2] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of IJCNLP*, 2013.

[3] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: automated identity theft attacks on social networks. In *WWW*, 2009.

[4] E. Blanzieri and A. Bryl. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92, 2008.

[5] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In *ACSAC*, 2011.

[6] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] H. M. Breland. Word frequency and word difficulty: A comparison of counts in four corpora. *PSS*, 1996.

[8] F. Chung. *Spectral graph theory*. Number 92. Amer Mathematical Society, 1997.

[9] C. Ding, T. Li, and M. Jordan. Convex and semi-nonnegative matrix factorizations. *TPAMI*, 2010.

[10] T. Egener, J. Granado, and M. Guitton. High frequency of phenotypic deviations in physcomitrella patens plants transformed with a gene-disruption library. *BMC Plant Biology*, 2:6, 2002.

[11] J. Eisenstein. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, 2013.

[12] J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning, 2008.

[13] S. Ghosh, B. Viswanath, F. Kooti, N. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. Gummadi. Understanding and combating link farming in the twitter social network. In *WWW*, 2012.

[14] J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García. Content based sms spam filtering. In *Proceedings of DocEng*, 2006.

[15] M. A. Halliday and C. M. Matthiessen. An introduction to functional grammar. 2004.

[16] M. Hart. *Project gutenberg*. Project Gutenberg, 1971.

[17] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of SIGIR*, 1999.

[18] X. Hu, N. Sun, C. Zhang, and T.-S. Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *CIKM*, 2009.

[19] X. Hu, J. Tang, and H. Liu. Online social spammer detection. In *AAAI*, 2014.

[20] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in microblogging. In *IJCAI*, 2013.

[21] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *WSDM*, 2013.

[22] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots + machine learning. In *Proceedings of SIGIR*, 2010.

[23] T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of ACL*, 2009.

[24] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *AirWeb*, 2007.

[25] V. Metsis, I. Androutsopoulos, and G. Paliouras. Spam filtering with naive bayes-which naive bayes? In *Proceedings of CEAS*, 2006.

[26] D. O'Callaghan, M. Harrigan, J. Carthy, and P. Cunningham. Network analysis of recurring youtube spam campaigns. In *Proceedings of ICWSM*, 2012.

[27] S. J. Pan and Q. Yang. A survey on transfer learning. *TKDE*, pages 1345–1359, 2010.

[28] D. Seung and L. Lee. Algorithms for non-negative matrix factorization. *NIPS*, pages 556–562, 2001.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[30] R. Wardhaugh. *An introduction to sociolinguistics*, volume 28. Wiley. com, 2011.

[31] S. Webb, J. Caverlee, and C. Pu. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *CEAS*, 2006.

[32] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Zhao, and Y. Dai. Uncovering social network sybils in the wild. In *Proceedings of IMC*, 2011.

[33] S. J. Yates. Oral and written linguistic aspects of computer conferencing. *Pragmatics and beyond New Series*, 1996.

[34] Y. Zhu, X. Wang, E. Zhong, N. Liu, H. Li, and Q. Yang. Discovering spammers in social networks. In *Proceedings of AAAI*, 2012.