Filtering of Brand-related Microblogs using Social-Smooth Multiview Embedding

Yue Gao, Senior Member, IEEE, Yi Zhen, Member, IEEE, Haojie Li, Member, IEEE, and Tat-Seng Chua

Abstract—In recent years, we have witnessed the boom of social media platforms, through which people have been generating a lot of social media data. This data touches almost every aspect of life and may have significant societal and marketing values for a variety of corporations and organizations. Thus, the development of effective techniques for gathering and analyzing social media content has attracted much research attention. As social media data tends to be heterogeneous, conversational and fast evolving in content, a recent work reported a multi-faceted approach to gather comprehensive brand-related data by crawling data using evolving keywords, key users, similar image content and known locations. Although such approach has been found to be effective in gathering representative data, it also brings in a lot of noise. This paper aims to develop an accurate classifier to filter out noise by taking into account the multimedia content and social nature of brand related data. In particular, we develop a microblog filtering method based on a discriminative social-aware multiview embedding. Besides the conventional content-based features, such as textual, low-level visual features, and high-level visual semantic features, that form the three key views of microblogs, we also incorporate the brand and social relations among the microblogs to learn a discriminative and social-aware embedding. With such a learned embedding, an off-the-shelf classifier, such as SVM, can then be trained and applied to microblog filtering. We verify the efficacy of our method on noise filtering in the brand data gathering task on the Brand-Social-Net dataset. Our approach is able to achieve significantly better filtering performance and improve the quality of brand data gathering.

Index Terms—Microblog filtering, Multiview embedding, Social graph.

I. INTRODUCTION

The popularity of microblogging platforms, such as Twitter\(^1\) and Sina Weibo\(^2\), has encouraged users to generate and share a huge amount of social media data known as user-generated contents (UGCs). These UGCs offer real-time information resources for a wide range of topics and are beneficial to a broad range of users and applications\(^{[21],[30]}\). Therefore, extensive research efforts have been dedicated to social media analysis, such as social event summarization\(^{[5]}\), social network analysis\(^{[3],[9],[10]}\), social TV\(^{[16]}\), and sensing the topics\(^{[1]}\). Among the social contents, brand related microblogs, which spread much faster than traditional media content, have significant marketing values for enterprises and governmental organizations\(^{[12],[32]}\). For example, positive or negative comments of a brand may impact the decision of users on whether to purchase the product, especially when the comments come from their friends. As another example, sensitive news or comments can often spread very fast across the entire social network and become viral and uncontrollable. As a result, it is essential for companies and organizations to know what people are talking about them in social network and be able to take preventive actions if necessary.

\(^{1}\)https://twitter.com/\(\)
\(^{2}\)http://www.weibo.com

Fig. 1. A typical framework for brand data gathering.

In social media gathering, brand data gathering is an important and difficult task, which has attracted much research attention (\(^{[11],[20],[33]}\)) in recent years. Figure 1 shows an example framework for this task, which consists of two main stages: extended data gathering and noise filtering. The objective of extended data gathering is to gather a comprehensive set of brand related data from live social media streams using a variety of techniques with the aim to improve data recall; while the goal of the noise filtering step is to remove the noise from the gathered data to achieve high precision. The overall task is challenging due to several factors. First, the short and conversational nature of microblogs leads to rapid changes in content and vocabularies in the related posts. Under such circumstances, the use of a fixed set of keywords is not adequate to track representative data and analyze social exposure for the target brand. Second, the content of microblogs tends to be heterogeneous. For example, it often consists of text, images, and/or videos. Our recent study\(^{[11]}\) shows that about 40% of microblogs contain visual content, while only a small portion of such microblogs (about 30%) contains meaningful text annotations. It is not easy to identify the category of the microblog using just text or image content individually. Considering the rapid changed discussion background in microblog platform, similar content may convey different meanings when the circumstance varies. Concern-
ing all these properties of microblogs, microblog filtering is much challenger than traditional image classification. Figure 2 shows some example microblogs containing text content, visual content only, and both of the text content and visual content together.

![Example microblogs](image1)

Fig. 2. Example microblogs. (a) A microblog containing text only. (b) A microblog containing images only. (c) A microblog containing both text and images.

This phenomenon increases the difficulties in microblog data analysis, in which using text analysis alone is inadequate. Thus, a multi-faceted approach is needed for social media processing to overcome the limitation of textual keyword-based methods. Third, the social media data is often accompanied with social information, such as user friendship relations and geographic network, which may contain valuable information for brand data gathering and analysis.

The multimodal data in social media has also attracted research attention from various fields. In [22], Ntalianis and Doulamis proposed to jointly employ the multimedia content for human life summarization, where the content analysis is comprised of visual content and other associated metadata, such as date, events, likes, and comments. Targeting the task of sensing trending topics in Twitter, Aiello et al. [2] proposed to explore the temporal distributions of concepts to deal with heterogeneous streams.

For data gathering, [11] is the first attempt to use a multi-faceted approach to collect brand-related microblogs in social media streams. In this method, a set of seeds are selected based upon the text-based search and logo detection. After that, the text, social context, and visual content of the seeds are used together to collect more microblogs. In this way, a comprehensive set of relevant microblogs can be obtained in comparison to most existing approaches that rely on text only. It is noted that the improvement on data coverage comes at the cost of bringing in a lot of noise. Therefore, an effective noise filtering step is needed which is the focus of this work.

In this paper, we propose a microblog filtering method, which is able to filter out noisy data from the gathered microblogs for a given brand. The idea behind our brand filtering framework is to map the social media data, microblogs in our case, into a latent subspace that is not only discriminative with respect to the target brand but also consistent with various types of information, including content features, such as textual and visual content, and social features, such as user relationship and spatial-temporal relatedness. Specifically, we propose a multiview embedding approach to jointly model the textual content and visual content represented using the low-level visual features and high-level visual semantic features. We further utilize the brand labels and social features as regularizers in order to take into account both brand and social relations. With the learned embedding in the latent space, an SVM classifier is trained and used for microblog filtering. Following the brand data gathering procedure in [11], we evaluate the proposed microblog filtering method on the Brand-Social-Net dataset. The evaluation is conducted on 3 million microblogs from Sina Weibo with 100 brands. We compare our method with several existing filtering methods and the experimental results demonstrate that our method is able to achieve significantly better performance.

The work in [11] targets on gathering data in social media streams, where a multi-faceted approach is employed to investigate different clues in microblog platforms. Different from [11], this work mainly focuses on filtering out the noise data in microblog platform. [11] aims to improve the coverage of data gathering, while the current work is for the precision of data gathering. Therefore, this work can be used as the next procedure after data gathering [11].

The remainder of this paper is organized as follows. Section II reviews related work. Section III introduces the proposed microblog filtering method named discriminative social-aware multiview embedding. Section IV presents the experimental results, followed by conclusions in Section V.

II. RELATED WORK

The rapidly growing amount of social media data has led to extensive research efforts in recent years. These include research on efficient storage, indexing and ranking of social media content [7], [20]. Different from web data analysis, social media data has several unique characteristics that make its analysis much harder. For example, as revealed by Teevan et al. [33], the Twitter queries are shorter in comparison with web queries, and traditional web query analysis methods cannot be readily applied to achieve satisfactory performance. Furthermore, social media data comprises multi-faceted content, including text, image, as well as social, temporal and geographical relations among tweets and users, and hence the use of only one media type is too limited.

Among the applications, data gathering in social media, which aims at finding relevant social data for a given target, has become a hot topic. Massoudi et al. [18] introduced a dynamic query expansion method with textual and microblog-specific quality indicator to retrieve microblog posts. In this method, the query expansion considers both the topic keywords and the posting time of microblogs. Mishne [19] incorporated time to find the relevant social posts of a news event. Weerkamp and Rijke [37] proposed to use two-level textual
credibility, including the single post level and the blog level, to describe the target. The hashtag patterns were investigated over time to obtain related social posts for a given topic [15]. Nagmoti et al. [20] introduced a ranking approach for microblog search, which employed both the microblog content and meta data. TweetMotif [23], a search application for Twitter, was able to gather messages and further group messages by significant terms. Reutel and Cimiano [28] proposed to utilize the textual content, uploading time, capture time, tags, and geographic location to collect relevant candidate data for social event detection. Uysal and Croft [35] proposed to employ the retweet behavior for personalized tweet ranking. In this method, the coordinate ascent algorithm was used to rank the microblogs. In [39], Zangerle et al. proposed a text-based hashtag recommendation approach, where the text content was used to investigate the correlation between the microblog and the hashtags. Batool et al. [4] proposed a tweet classification method based on text content. In this method, the keywords, entities, synonyms, and parts of speech from tweets were employed for tweet classification. Hu et al. [14] proposed to employ the internal semantics from the original text content and external concepts from world knowledge for microblog text clustering. In this work, a three-level hierarchical structure was used to tackle the data sparsity and generate the feature space with multiple semantic information.

There have been several existing social media monitoring platforms online. Trackur\(^3\) provides social media monitoring, supporting all social media and mainstream news, such as Twitter, Facebook, and Google+. Trackur supports the analysis of trends, keyword discovery, automated sentiment analysis and influence scoring. Naymz\(^4\) measures the influence across social media platforms, such as LinkedIn, Facebook and Twitter. Brandseye\(^5\) provides the service to monitor the data on social media platforms for specific targets. The sentiment across 1000 mentions for each operator can help to recommend better business decisions. Rankur\(^6\) is another social media monitoring and online reputation management service provider. Rankur is able to monitor online reviews, blogs, news, forums and social networks, and also can identify community leaders and customers’ behaviors. SocialMention\(^7\) provides the service of social media search and analysis. SocialMention is able to track and measure the content about a specific target, such as company, product or any other topic. Google Alerts\(^8\) targets on content change detection and notification on social media platforms. New results, such as web pages, blogs, and articles, can be sent to users through emails.

One specific data gathering application in social media is brand tracking. Gao et al. [11] proposed to gather brand related data using a multi-faceted brand tracking method, which was based on not only the evolving keywords but also social context, such as user profile, location and visual content. In this method, a hypergraph is constructed to jointly learn the relevance using both textual and visual content of microblogs for reranking.

Our method is also related to multiview embedding methods that have been widely applied to multimedia data. One basic approach is Canonical Correlation Analysis (CCA) that maps multiview data into a latent space that maximizes the feature correlations [13]. CCA-based approaches have been widely used in cross-modal retrieval [27], cross-lingual retrieval [34], and cross-view clustering [6]. To eliminate shortcomings of CCA such as fixed dimensionality, some sophisticated machine

\(^3\)http://www.trackur.com/  
\(^4\)http://www.naymz.com/  
\(^5\)https://www.brandseye.com/  
\(^6\)https://rankur.com/  
\(^7\)http://www.socialmention.com/  
\(^8\)https://www.google.com/alerts
learning approaches have been developed [26]. Metric learning methods for single view data have also been extended to multiview data which may enjoy the convex-concave property [25] or large-margin property [8].

III. THE PROPOSED MICROBLOGS FILTERING METHOD

A. Outline of Framework

We first illustrate our framework for brand data filtering for microblogs as shown in Figure 3. For a given brand a.k.a. target brand, such as Nike, we first perform brand data gathering, using the method such as that presented in [11], to obtain a comprehensive set of microblogs. Based on the gathered data, the method proposed in this paper aims to filter out noisy records, i.e., the irrelevant microblogs to the target brand. As shown in Figure 3, the proposed framework proceeds in two stages: an offline training stage and an online testing stage.

In the offline training stage, we assume that there is a labeled training set with respect to the target brand; in other words, each microblog is labeled as 1 if it is related to the target brand and 0 otherwise. We first extract content features for each microblog in the training dataset. Specifically, we consider three kinds of content features: textual features, low-level visual features, and high-level visual semantic features.\(^9\) Besides the content features, we also utilize information on brand similarity and social features. As the training data is labeled based on whether they are related to the target brand, the brand similarity is 1 if two items are both related to the brand and 0 otherwise. The social features include the social relations between microblogs, such as user relations, temporal information, and geographical information. Based on the content features and social features, we learn an embedding of the training data via a regularized multiview embedding method. As the data consists of multiple views (one view for each type of content features) and the embedding is obtained by exploiting both the label information and the social features, we call our approach discriminative social-aware multiview embedding. As the final step, an SVM classifier is trained for the target brand in the embedded space.

In the online testing stage, given a set of unlabeled microblogs, we first extract the content and social features from the testing data, and learn the new representations based on the model learned during the offline training stage. Using the new representations, the classifier trained in the offline training stage is employed to predict whether the testing microblogs belong to the target brand.

In what follows, we use lowercase boldface letters such as \(a\), uppercase boldface letters such as \(A\), and calligraphic letters such as \(\mathcal{A}\) to denote vectors, matrices, and sets, respectively. For a matrix \(A\), \(A^\top\) is its transpose and \(\|A\|_F\) is its Frobenious norm.

B. Content Features

For each microblog, we extract three types of content features: the textual feature, the low-level visual feature, and the high-level visual semantic feature. For the two visual features, they are available only from the microblogs containing images. In following, we describe how these features are extracted.

1) Textual Features: Most microblogs contain a set of textual words which is often the most direct way for users to express their view on a target brand. For example, a user may use a sentence like “I like my new Nike shoes” to show his/her positive comment to the brand Nike. To extract the textual feature, we first perform text parsing for each microblog’s text\(^10\) and encode each word using a dictionary \(D_{text}\). By removing the stop words and the 200 most frequently used words\(^11\), we keep 5,000 words with the highest frequency to form \(D_{text}\). We then employ the commonly used \(tf-idf\) term weighting scheme [31] to assign weight for each term, giving rise to a 5,000-dimensional term vector for each microblog text. For example, the weight for the \(i\)th word is:

\[
f_i = \text{tf}(i) \times \log_2 \left( \frac{N}{\text{df}(i)} \right),
\]

where \(\text{tf}(i)\) is the normalized frequency of word \(i\) appeared in the current microblog text, \(\text{df}(i)\) is the number of microblogs that contains word \(i\), and \(N\) is the total number of microblogs.

2) Low-Level Visual Features: In the Brand-Social-Net dataset [11], about one third of the microblogs contain images. For example, a user may post a picture of his/her new car to show his/her excitement, but without any meaningful accompanying text. In such cases, images in microblogs may be the only indicator of the relationship between the microblogs and the target brand. In this work, we extract the spatial pyramid image feature [38] for each microblog image, as it has been found to be discriminative on spatial layout and local information. Specifically, we extract a set of dense SIFT features [17] (about \(1,000 - 2,000\) SIFT features) for each image and learn a visual dictionary \(D_{image}\) of size 1024 via sparse coding [38]. A spatial pyramid feature is then generated by multi-scale max pooling. In other words, the spatial pyramid structure consists of three levels and each level consists of different numbers of blocks. Specifically, we use 16 blocks in the first level, 4 blocks in the second level, and 1 block in the third level. Therefore, there are \(16 + 4 + 1 = 21\) blocks in total. With max pooling, we obtain a \(1,024 \times 21 = 21,504\) dimensional feature vector for each image. We finally reduce the feature space to 200 by using Principle Component Analysis (PCA) as the low-level visual features.

3) High-Level Visual Semantic Features: The visual feature extracted in the previous subsection may be not informative enough since they only capture the low-level visual information. We further extract the high-level visual semantic information from the microblog images.

Our semantic features are based on ImageNet,\(^12\) a well-known image database that is organized semantically. ImageNet contains 1,000 image concepts collected from WordNet.

\(^9\)Note that our framework is not limited by the number of content features. Actually, it can be easily extended to use more than three kinds of features.

\(^10\)Since our testing dataset is crawled from Sina Weibo, the text is mostly in Chinese and a Chinese language parser is needed.

\(^11\)These most frequently used words are not brand related.

\(^12\)www.image-net.org
Net\textsuperscript{13} which is a human knowledge database of English words, and, for each concept, there are 1,000 to 1,300 images which are semantically related to the concept. To obtain the high-level visual semantic features, we train 1,000 SVM classifiers\textsuperscript{14}, one for each concept, to predict the concept labels for each image. The image features are combined to train the semantic classifiers include HOG, SIFT, Local Binary Patterns (LBP) [24], and color (histogram or color moment) features. Note that, by tuning the feature complexity, we can control the trade-off between the prediction speed and accuracy. Generally, it costs about 1 second to process one image for the 1000 concepts. As there are quite a number of concepts about animal topics which are irrelevant to our target brands, we manually select 369 concepts from the ImageNet, such as jeep, restaurant, and notebook, in our experiments.

The other important high-level visual semantic for brands is the logo information. Here we employ the cascaded classifier which was jointly trained using Adaboost and SVM [11] for logo detection. In this method, a recursive process as presented in [36] is employed for training by producing a cascaded classifier consisting of multiple node classifiers. To detect logos, each testing image is split into sub-windows at multiple scales and the sliding window method with 1 pixel stride on two directions is employed for scanning. The detection process is fast due to that each node of the cascaded classifier can eliminate a large amount of negative windows. It is noted that some brands may have multiple logos. The logo information is set to 1 when any of these logos is detected.

Overall, the two types of high-level visual semantic features, i.e., the selected 369 visual concepts and the logo information, are extracted for each microblog containing images.

C. Similarity Graphs

To incorporate the brand information and social features into the embedding, we introduce two similarity graphs: the brand similarity graph and the social similarity graph. While the brand similarity graph enforces that the embedding is discriminative with respect to the target brand, the social similarity graph implicitly requires that the embedding is smooth with respect to the social relationships which are generated from users friendship and geographical-temporal information. In the following two subsections, we will introduce the construction of the two similarity graphs. The objective of the brand-similarity graph is to explore the relationship among different brands, as there are underneath connections among such brands. The social similarity graph aims to explore the relationship among different microblogs through the social connection.

1) Brand Similarity Graph: The brand similarity graph is constructed from the training data which has been labeled as relevant or irrelevant to a given brand. The objective of the brand similarity graph is to supervise the training phase with the label information of the brand. Given the labeled training microblogs, the brand similarity graph is defined by an adjacency matrix \( W_b \in \mathbb{R}^{N \times N} \) and \( W_b(i, j) = 1 \) if the two microblogs \( T(i) \) and \( T(j) \) are both relevant to the target brand or \( i = j \); otherwise \( W_b(i, j) = 0 \). Here \( L_b \) is the graph Laplacian defined on \( T \):

\[
L_b = I - D_b^{-\frac{1}{2}} W_b D_b^{-\frac{1}{2}},
\]

where \( D_b \in \mathbb{R}^{N \times N} \) is a diagonal matrix with \( D_b(i, i) = \sum_{i'=1}^{N} W_b(i, i') \) and \( I \) is an \( N \times N \) identity matrix.

2) Social Similarity Graph: The social similarity graph encodes the social relationship among microblogs, because social connection is also an import indicator of microblog similarities. Here we take the training data as an example. We extract three types\textsuperscript{15} of social similarity graphs from the training data \( T \): (a) the user connection-based adjacency matrix, where \( W_{s1}(i, j) = 1 \) if \( T(i) \) and \( T(j) \) are posted by the same user or the two users are connected, such as the two users are under the relation of follower/followee or have comments on each other’s microblogs; otherwise \( W_{s1}(i, j) = 0 \); (b) the temporal information-based adjacency matrix, \( W_{s2}(i, j) = 1 - \frac{(t_i - t_j)}{\tau} \), where \( t_i \) and \( t_j \) are the time-stamps of \( T(i) \) and \( T(j) \), respectively, and \( \tau \) is a normalized factor; and (c) the location-based adjacency matrix, \( W_{s3}(i, j) = 1 - H(g_1, g_2) \), where \( H(x, y) \) is the Haversine formula [29] and \( g_1 \) and \( g_2 \) are the geographical locations of \( T(i) \) and \( T(j) \).

We then define the social-aware graph Laplacian for the training data as

\[
L_s = I - \sum_{i=1}^{3} \varsigma_i D_{s i}^{-\frac{1}{2}} W_{s i} D_{s i}^{-\frac{1}{2}},
\]

where \( D_{s i} \in \mathbb{R}^{N \times N} \) is a diagonal matrix with \( D_{s i}(i, i) = \sum_{i'=1}^{N} W_{s i}(i, i') \), \( \varsigma_i \) is the weight for the \( i \)th social information. In our experiments, \( \varsigma_i \) is set as 1/3.

Similarly, given a set of testing data, the corresponding social similarity graph can be generated.

These social features are used here to explore the social similarities among microblogs. The features are designed based on the premise that a user, or connected users tend to have relatively high possibility to post microblogs with related content that will have same relevance or irrelevance to the target brand. The same premise can be observed for the location and temporal information. With the generated social similarity graph \( \text{Laplacian} \), we can formulate the social connection among microblogs, which can be regarded as complementary to microblog content analysis based on the content features.

These three types of information, \textit{i.e.}, user connection, temporal information and location data, generate three adjacency matrices, respectively, which are jointly used with the brand-similarity graph and the content features to build the connection among microblogs. The social information is used as a direction to explore the microblog relationship, just like the content features. It indicates that two microblogs who are

\textsuperscript{13}wordnet.princeton.edu

\textsuperscript{14}Note that linear classifiers such as logistic regression and probit regression, and nonlinear classifiers such as neural networks and random forests, are also applicable.

\textsuperscript{15}Note that our method is not limited in the number of social graphs. More social graphs can be constructed if the data have richer information.
close in the social connections should have high probability to be with similar labels.

D. Offline Training

Given a training set $\mathcal{T}$ with $N$ microblogs, we extract the textual features, and if the image is available, the low-level visual features and high-level visual semantic features for each microblog. We group the three feature vectors for all training data respectively into three matrices: $V_1 \in \mathbb{R}^{N \times D_1}$, $V_2 \in \mathbb{R}^{N \times D_2}$, and $V_3 \in \mathbb{R}^{N \times D_3}$. In other words, each microblog has three views: the textual view, the visual view, and the semantic view (note that for those microblogs without images, there is only the textual view).

Given the training data with multiple views (in our case, three views), our goal is to learn an embedding that incorporates information from different views. Besides, we expect the embedding to be discriminative with respect to the class labels (whether relevant to a target brand) and social relations between the microblogs. Specifically, we can achieve the above goal by formulating the following optimization problem:

$$
\min_{P_i, Y} \sum_{i=1}^{3} \alpha_i \|V_i P_i - Y\|_F^2 + \beta \text{tr}(Y^T L_b Y) + \eta \text{tr}(Y^T L_s Y)
$$

s.t. $P_i \Sigma_{ij} P_j = I$, \forall $i, j = 1, 2, 3$

(3)

where $P_i \in \mathbb{R}^{D_i \times K}$ is the projection matrix for the $i$th view; $Y$ is the embedding of all microblogs; $K$ is the dimensionality of the latent embedded space; and $L_b$ and $L_s$ are two Laplacian matrices constructed based on brand similarity and social similarity, respectively. $\Sigma_{ij}$ is the covariance matrix between $V_i$ and $V_j$, and $p_{ik}$ is the $k$th column of $P_i$. \{\alpha_i\}, \beta$ and $\eta$ are empirically set parameters that respectively control the relative importance of projection consistency, smoothness on the brand similarity graph and smoothness on the social similarity graph.

To solve the optimization problem in Eq. (3), we derive an alternating algorithm. Specifically, we first update $Y$ while fixing $\{P_i, i = 1, 2, 3\}$; we then update $\{P_i, i = 1, 2, 3\}$ while fixing $Y$.

The optimization problem of updating $Y$ can be rewritten as:

$$
\min_{Y} \sum_{i=1}^{3} \alpha_i \|V_i P_i - Y\|_F^2 + \beta \text{tr}(Y^T L_b Y) + \eta \text{tr}(Y^T L_s Y)
$$

(4)

This is a convex problem for which a local optimal solution is also the global optimum. We can thus use any gradient based method to find a local optimum, and the gradient can be calculated as follows:

$$
\nabla Y = \sum_{i=1}^{3} 2\alpha_i (Y - V_i P_i) + 2\beta L_b Y + 2\eta L_s Y.
$$

Given $Y$, we update each column of $\{P_i, i = 1, 2, 3\}$ sequentially while fixing the remaining parameters. Specifically, the objective function with respect to $p_{ik}$ can be rewritten as:

$$
\min_{P_{ik}} \|V_i P_{ik} - y_k\|_F^2
$$

(5)

where $y_k$ is the $k$th column of $Y$. This is again a convex problem that can be readily solved by existing solvers such as CVX\textsuperscript{10}.

The above two steps are proceeded alternately until convergence. Once $Y$ is learned, we train an SVM classifier\textsuperscript{17} for this brand. The workflow for the offline training procedure is shown in Algorithm 1.

**Algorithm 1** The workflow for the offline training procedure.

Input: The training data $\mathcal{T}$.

Brand similarity graph Laplacian $L_b$.

Social similarity graph Laplacian $L_s$.

Output: The projection matrices $\{P_i, i = 1, 2, 3\}$.

Procedure:

Solve the optimization task

$$
\min_{P_i, Y} \sum_{i=1}^{3} \alpha_i \|V_i P_i - Y\|_F^2 + \beta \text{tr}(Y^T L_b Y) + \eta \text{tr}(Y^T L_s Y)
$$

s.t. $P_i \Sigma_{ij} P_j = I$, \forall $i, j = 1, 2, 3$

(6)

**Step 1.** Fix $\{P_i, i = 1, 2, 3\}$ and update $Y$.

$$
\min_{Y} \sum_{i=1}^{3} \alpha_i \|V_i P_i - Y\|_F^2 + \beta \text{tr}(Y^T L_b Y) + \eta \text{tr}(Y^T L_s Y)
$$

Step 2. Fix $Y$ and update $\{P_i, i = 1, 2, 3\}$.

$$
\min_{P_{ik}} \|V_i P_{ik} - y_k\|_F^2
$$

Step 3. Repeat Step 1 and Step 2 until convergence.

E. Online Filtering

Given a group of $M$ testing microblogs $\mathcal{Z}$, e.g., the brand data gathering results from [11], the goal of online filtering is to predict whether each microblog is relevant to the given brand. First of all, we extract the three features for each microblog when available, and let $V_1$, $V_2$, and $V_3$ respectively denote the feature matrices for different views for all testing data.

With $P_i \{i = 1, 2, 3\}$ learned in the training phase, we can obtain the embedding $Y$ for all testing data by solving:

$$
\min_{Y} \sum_{i=1}^{3} \alpha_i \|V_i P_i - \tilde{Y}\|_F^2 + \beta \text{tr}(\tilde{Y}^T \tilde{L}_b \tilde{Y}) + \eta \text{tr}(\tilde{Y}^T \tilde{L}_s \tilde{Y})
$$

(7)

where $\tilde{L}_s$ is a Laplacian matrix constructed based on social similarity information of the testing data $\mathcal{Z}$. We then predict the class labels of $\mathcal{Z}$ based on $Y$ and the classifiers learned in the offline training stage.

To better illustrate the whole procedure, we summarize it in Algorithm 2.

\textsuperscript{10}http://cvxr.com/cvx/

\textsuperscript{17}Here the standard learning algorithm of SVM is used and we omit the details due to space limitation.


### Algorithm 2 The workflow for the proposed microblog filtering method

**Input:**
- The training data $T$.
- The testing data $Z$.

**Output:**
- The trained SVM $f(\cdot)$.
- The filtering results for $Z$.

**Procedure:**

1. **Step 1.** Extract features for $T$.
2. **Step 2.** Generate graph Laplacians $L_b$ and $L_s$.
3. **Step 3.** Learn $\{P_i, i = 1, 2, 3\}$ and $Y$ for $T$.
4. **Step 4.** Train SVM classifier $f(\cdot)$ with $Y$.
5. **Step 5.** Extract features for $Z$.
6. **Step 6.** Generate graph Laplacians $\tilde{L}_s$.
7. **Step 7.** Predict class labels for $Z$ using $f(\cdot)$ and $\tilde{Y}$.

---

### IV. Experimental Evaluation

In this section, we introduce the testing dataset, evaluation measures, experimental settings, results and discussions.

#### A. Dataset

To evaluate the performance of the proposed microblog filtering method, we employ the Brand-Social-Net dataset\(^{18}\) \cite{11}, abbreviated as BSN, as the test bed. BSN is a recently published dataset to facilitate research on social media analysis. It consists of 3 million microblogs that were crawled from one of the largest social network sites in the world, the Sina Weibo, during the period of June and July 2012. Most microblogs are text based, with a large proportion containing images. For example, there are 1.2 million images in 3 million microblogs. In this dataset, there are 100 popular brands from a variety of categories, including automobile, sport, electronic products, and cosmetics. Figure 4 presents the logos of the brands in the dataset. All microblogs are manually labeled as relevant or irrelevant to the 100 brands. Furthermore, the logos of these brands are also annotated for the microblog images. There are hundreds to thousands of relevant microblogs for each brand in the dataset. In addition to microblogs, the dataset contains the information about 1 million users and their social relations, such as who generated the microblogs.

#### B. Experimental Settings

The experiments are set under the brand data gathering framework in \cite{11}, i.e., extended data gathering (EDG) based on social context and visual content, which is the most recent method for data gathering in social media. In experiments, given a target brand, the extended data gathering procedure as presented in \cite{11} is conducted to collect microblogs. In this method, a set of text-based results are first obtained, and then logo detection is conducted on the images from these text-based results, which further generates a group of seed microblogs based on the logo detection results. These seed microblogs are further used to gather more candidate relevant microblogs via social context, such as active users, known location, and visual contents. The obtained microblog set is employed for our microblog filtering experiments. It should be noted that our filtering framework can be applied to any set of microblogs gathered using any method.

We compare our method with extended data gathering method (EDG) presented in \cite{11}. Given a group of testing microblogs, EDG utilizes the hypergraph structure \cite{40} to model the relationship among the microblogs. Then a semi-supervised learning algorithm is performed on the hypergraph structure to learn the relevance of the testing microblogs to the given brand. We also report the performance of SVM with individual content features and several variants of our method, including:

- SVM with textual features ($SVM_t$).
- SVM with low-level visual features ($SVM_v$).
- SVM with high-level visual semantic features ($SVM_s$).
- Multiview embedding (MVE).
- Multiview embedding with social regularization, i.e., the proposed method (MVE+SR).

The parameters of the compared methods have been tuned via grid search, and the best performance is reported in this paper. For SVM methods with individual content features and our method, we try with several variants of training set $T$ for each brand by randomly selecting $n_{pos}$ positive samples and $n_{neg}$ negative samples from the entire dataset as the training set $T$ for each brand respectively. To compare these methods, 100 positive samples and 500 negative samples for each brand are used as the training samples. In our method, the embedded latent space dimensionality $K$ is set to 100, $\alpha_i (i = 1, 2, 3)$, $\beta$

---

\(^{18}\)www.nextcenter.org/Brand-Social-Net/
and $\eta$ are all set to be 1.

As for the text feature, the weight for each word is important for microblog content representation. We have conducted data sampling to select different portions of data, such as 0.5 million, 1 million, 2 million, for word weight calculation. We observe that given different number of randomly selected data from the whole dataset, the weights are relatively steady and the filtering performance keeps almost the same accordingly.

C. Evaluation Measures

We use the following metrics to evaluate the performance for different microblog filtering methods:

- **Recall** ($Re$). Recall measures the percent of correctly classified microblogs for a target brand, calculated by:
  $$Re = \frac{\# \text{Correct Relevant Microblogs}}{\# \text{All Relevant Microblogs}},$$
  where $\# \text{Correct Relevant Microblogs}$ is the number of correctly detected relevant microblogs for the target brand, and $\# \text{All Relevant Microblogs}$ is the number of relevant microblogs in the BSN dataset for the target brand.

- **Precision** ($Pr$). Precision measures the accuracy of the microblog filtering, calculated as:
  $$Pr = \frac{\# \text{Correct Relevant Microblogs}}{\# \text{All Classified Microblogs}},$$
  where $\# \text{All Classified Microblogs}$ is the number of microblogs which are classified as relevant to the brand.

- **F1-Measure** ($F1$). F1-Measure jointly considers Recall and Precision, defined as:
  $$F1 = \frac{2 \times Re \times Pr}{Re + Pr}.$$

D. On Comparison with the State-of-the-art Methods

Figure 5 presents the comparative results of different methods for all the 100 brands data in terms of Recall, Precision and F1-Measure, respectively. The proposed method achieves an average $F1$-Measure of 0.74 ± 0.05, with the best and the worse performance of the proposed method of 0.85 and 0.63, respectively.

Here we first compare the proposed method with the baselines (the SVM methods using different content features).

As shown in Figure 5, all MVE methods outperform SVM methods by a large margin. Here we take SVM$_v$ as an example. One variant of the proposed method, MVE+SR, achieves an improvement of 69.00%, 83.36%, and 78.54% as compared to SVM$_v$ in terms of Recall, Precision and F1-Measure, respectively. These results demonstrate that the proposed method can achieve better microblog filtering results in comparison with the SVM based methods. Compared with EDG, which is the most recent method for brand data gathering and noise filtering, the proposed method also achieves much better results with improvements of 13.12%, 38.87%, and 25.37% for Recall, Precision and F1-Measure, respectively. The results show that the proposed discriminative social-aware multiview embedding is effective on noisy data removal. We also note that the proposed method can improve the precision of microblog filtering much more than recall. This indicates that there is only very few false classification results, which means that the proposed method is able to extract the discriminative features for brands.

As shown in the results, the performance of using visual content only (SVM$_v$) performs worse than that of using textual content only (SVM$_t$). As introduced before, the texture and visual content may be not consistent for the data in microblogs, which leads to the fact that the visual content can be not directly related or not related to the brand. Under such circumstance, filtering using just visual content can perform worse than that of using textual content only.

![Fig. 5. Comparison of different methods on all the 100 brands in terms of Recall, Precision and F1-Measure, respectively.](image)

![Fig. 6. Experimental results for six exemplar brands, i.e., Apple, KFC, Suzuki, Hershey, Nivea, and Olay, in terms of F1-Measure.](image)
brand defined Laplacian and the social network information. In following subsections, we will explore the impact of these factors individually.

E. On False Positives and False Negatives

Most of false positives come from the false leads on the brand information, such as texture content misunderstanding and inaccurate logo detection. A few false positives come from social connections, where some users may post related microblogs and other microblogs in similar time and location for the same/related users becomes false positive. For false negatives, most of this type of mistakes come from the missing of brand information. For example, some brand-related microblogs do not contain direct key words, but are connected using social connections. Although the proposed method is able to explore social connections to detect brand-related microblogs, there can be also false negatives if the social connections are not strong enough or even fail. Another type of false positives for the microblogs with images comes from the missing of logos. We note that some logos are too small or blurred and cannot be detected accurately. In such circumstance, the logo information will lose and false negatives may come.

F. On Multiview Embedding

In this subsection, we evaluate the properties of the proposed multiview embedding method for microblog filtering, which considers three kinds of content features, i.e., textual features, low-level visual features, and high-level visual semantic features. Specifically, we make comparison between MVE with SVM using individual features to justify the effectiveness of the use of the combination of all three views.

As shown in Figure 5, MVE can achieve an improvement of 7.11%, 71.80%, and 60.67% from SVM_t, SVM_v, and SVM_s in terms of F1-Measure, respectively. These results can be attributed to two reasons. First, the high-level visual semantic features can describe the brand related visual context from a higher conceptual level, and hence they are better than the low-level visual features that only capture low level information. Beside the high-level visual semantic features, other views for microblog description can be further employed in the proposed framework, such as the topic information, which can be extracted using the topic models learned from the microblog texts. This can be regarded as another form of high-level textual semantic features. Second, the proposed multiview embedding framework can jointly learn the relevance among three content views, which is able to obtain the optimal discriminative features.

The proposed method is able to jointly explore multiple features in the learning process, which leads to better performance compared with the use of single feature. The proposed method is also flexible to deal with the missing of multidimensional features. In such circumstance, only one or two types of features can be used in the proposed method.

G. On Social Regularization

In this part, we evaluate the impact of the proposed social regularization term for the multiview embedding framework on microblog filtering. The social regularization, including user connection, temporal correlation, and geographical location similarity, is employed in both the offline training and the online filtering procedures. The social regularization term can help to jointly investigate the social information and the multiview feature embedding of microblogs. Here we compare the experimental results between MVE and MVE+SR to evaluate the influence of social regularization.

As shown in Figure 5, the use of social regularization can improve the performance by 3.65%, 0.72%, and 3.93% from MVE to MVE+SR in terms of Recall, Precision and F1-Measure, respectively. The performance gain comes from the employment of social connections in social media platforms.

Another interesting observation is that the effectiveness of the social regularization varies a lot for different brand categories. We analyze the number of relevant microblogs and all microblogs for each user to each brand. We find that when there are influential users for the brands, such as the official or sales accounts, who have very big social network and a large number of microblogs about the brands, the social regularization can exploit much information from the social network, which further improves the microblog filtering performance more than the other brands. For example, there are several influential users for Apple, KFC, and Suzuki, who post many relevant microblogs in the dataset and also have hundreds to thousands of other microblogs. For these three brands, the social regularization works better, where the improvement of using social regularization are 5.73%, 8.39%, and 5.34% from MVE to MVE+SR in terms of F1-Measure, respectively. On the other hand, for Hershey, Nivea, and Olay, there is no such big users, and the improvement that can be obtained from social regularization is relatively smaller as compared to that in Apple, KFC, and Suzuki.

H. On the Embedded Latent Space Dimensionality

In this subsection, we evaluate the influence of dimensionality $K$ of the embedded latent space in the proposed method. Here we vary $K$ from 10 to 200, and the experimental results are shown in Figure 7.

![Fig. 7. The overall performance curves with respect to different $K$.](image-url)

As shown in Figure 7, we can observe that when $K$ is too small, such as when $K = 10$, the microblog filtering performance is unsatisfactory. With the increasing of $K$, the performance improves. When $K$ is large enough, such as for...
I. On the Training Data

In this subsection, we evaluate the influence of the number of training data on the proposed discriminative social-aware multiview embedding method. For the training data, we vary the number of positive samples, i.e., $n_{pos}$ from 50 to 300 (if available) and train the multiview embedding method, where the number of negative samples keeps at about five times that of the positive samples. The experimental results are shown in Figure 8.

![Fig. 8. The overall performance curves with respect to different $n_{pos}$](image)

From Figure 8, we can observe that the proposed method can achieve satisfactory results with even a small set of training data. The increasing size of training samples can lead to better results while the general trend is smooth with respect to the increasing of the training set. Generally, 100 to 200 positive training samples are sufficient to achieve satisfactory performance.

J. On the Weights of Regularizers

In our multiview embedding formulation, there are several key parameters that control the relative impact of different information sources: $\alpha_i (i = 1, 2, 3)$ for the content features, $\beta$ for the brand related similarity, and $\eta$ for the social similarity among microblogs. Specifically, $\alpha_i$ modulates the effect of the loss term, $\beta$ weights the brand graph regularizer of the training data, and $\eta$ weights the social graph regularizer for both the training data and the testing data.

![Fig. 9. The performance evaluation by varying key parameters.](image)

To evaluate the influence of $\alpha_i$ on the microblog filtering task, Figure 9 (a) compares the performance by varying $\alpha_i$ while fixing $\beta = 1$ and $\eta = 1$. As shown in the results, we can see that the performance curves mainly exhibit a “wedge” shape as $\alpha_i$ varies. When $\alpha_i$ is too small, such as $\alpha_i = 0.01$, the effect of content feature embedding will be weak, and the other two factors play more important roles in the offline training and the online filtering stages. Without the content features, the microblog filtering performance cannot be satisfactory by using only the brand related similarity and the social similarity. When $\alpha_i$ is too large, the brand related similarity and the social similarity will have very little impact on the microblog filtering results. Therefore, when $\alpha_i$ is large, such as $\alpha_i > 10$, the microblog filtering results will be a bit worse than the best results but remain stable.

We next evaluate the influence of different $\beta$ values by fixing $\alpha_i = 1(i = 1, 2, 3)$ and $\eta = 1$. Figure 9 (b) compares the performance by varying $\beta$. Here, we find that when $\beta$ is too small or too big, the performance is not good. Thus, the performance curve is also a “wedge” shape. This observation is due to the fact that a small $\beta$ value will reduce the effect of labeled training data, which makes the training data not discriminative enough to the given brand and leads to unsatisfactory results. A too high $\beta$ value also leads the degradation of performance, which reduces the influence of content features and the social regularizer.

Finally, we fix $\alpha_i = 1(i = 1, 2, 3)$ and $\beta = 1$, and vary $\eta$. The experimental results are presented in Figure 9 (c). As shown in these results, when $\eta$ is very small, the social regularizer will have very little contribution to the proposed method, where only the content features and the training labels have effect on the microblog filtering. Therefore, the performance is steady. With the increasing value of $\eta$, the social regularizer will have more impact on the results, which leads to improvement in the microblog filtering performance. However, when $\eta$ is too large, such as $\eta > 10$, the performance will degrade dramatically, which indicates that the social connection itself is not adequate for data analysis.

We note that in our current work, different content features are regarded with equal weight $\alpha_i$. To better represent microblogs, the weights for different content features should be learnt accordingly, which can lead to better performance. For the microblogs from different brands, the textual data and the visual data can play different roles in microblog representation. For example, the textual and visual features should have different weights for BMW and KFC. How to learn an optimal combination weights for different content features is an important task and it will be investigated in our future work.

V. CONCLUSION

In this paper, we proposed a microblog filtering method, which can be used as the noise filtering step for the brand data gathering task. The key component of our method is a discriminative social-aware multiview embedding approach, which maps the microblog content, consisting of three (or more) views, into a latent space, while taking into account the brand information and social relations of microblogs. Extensive experiments conducted in the Brand-Social-Net dataset with
100 famous brands demonstrated that the proposed microblog filtering method can achieve better performance in comparison to the state-of-the-art methods. We also discovered one interesting property of social information from the experiments that it makes more impact to microblog filtering for brands that have influential users with large social connections and followings.

It is noted that data labeling requires a lot of manual work. Although we have fully annotated all the data for evaluation in this work, it is not affordable to label all the coming data. Confronting the changing circumstances in microblog platform, one possible solution is to actively keep monitoring the related data which can be used to automatically update the training dataset without additional human annotations.

There are still several open issues on microblog filtering. First, learning the brand-related visual context is an important issue, and the ability to extract target object information from the images and discover the associated high-level semantic feature may greatly improve the microblog filtering results. Second, the deep structure of social context such as conversational graph structure of the microblogs should be exploited to boost the precision and recall of brand data prediction. Last but not least, exploring the multiple-topic information of microblogs via topic models may provide more semantic level information, and can be incorporated to enhance our multiview embedding framework.

For our current work, there are also limitations. For the brand-similarity grand and the social-similarity graph, they are with high computational complexity. For our task, our objective is to filter the microblogs for a given target, such as a brand. The current framework is able to handle tens or hundreds of thousands microblogs each time but has the limitation on handling more data, such as 1 million data, which is a limitation of our work. Another limitation of our current work is that the correlation among different brands has not been investigated, which is also important for microblog content analysis. For example, the microblogs from brand “Mazda” and brand “BWM” should share similar visual content about cars. Therefore, the brand correlation should be taken into consideration in the future.

REFERENCES


Yue Gao (SM’14) received the B.S. degree from the Harbin Institute of Technology, Harbin, China, and the M.E. and Ph.D. degrees from Tsinghua University, Beijing, China.

Yi Zhen is a Senior Research Scientist with Baidu Research. He received his Bachelor degree from Nankai University, Master degree from Tsinghua University, and PhD degree from the Hong Kong University of Science and Technology. He has worked as a research scientist with Georgia Tech, a research associate with Duke University, and a postdoctoral research fellow in Huawei Noah’s Ark Lab. His research interests are machine learning and its applications to data mining, health informatics, and computer vision.

Haojie Li (M’08) is a Professor in the School of Software, Dalian University of Technology. His research interests include social media computing and multimedia information retrieval. He has co-authored over 50 journal and conference papers in these areas, including IEEE TCSVT, TMM, TIP, ACM Multimedia, ACM ICMR, etc.

Dr. Li received the B.E. and the Ph. D. degrees from Nankai University, Tianjin and the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 1996 and 2007 respectively. From 2007 to 2009, he was a Research Fellow in the School of Computing, National University of Singapore. He is a member of IEEE and ACM.

Tat-Seng Chua received the B.S. and Ph.D. degrees from the University of Leeds, Leeds, U.K., in 1979 and 1983, respectively. He then joined the National University of Singapore, where he is now a full Professor at the School of Computing. He was the Acting and Founding Dean of the School of Computing during 1998-2000. He spent three years as a research staff member at the Institute of Systems Science (now I²R) in the late 1980s. His main research interest is in multimedia information processing, in particular on the extraction, retrieval, and question-answering (QA) of text and video information. He focuses on the use of relations between entities and external information and knowledge sources to enhance information processing. He is currently working on several multi-million-dollar projects: interactive media search, local contextual search, and event detection in movies. His group participates regularly in TREC-QA and TRECVID news video retrieval evaluations.

Dr. Chua is active in the international research community. He has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia, and text processing. He is the conference co-chair of ACM Multimedia 2005, CIVR (Conference on Image and Video Retrieval) 2005, and ACM SIGIR 2008. He serves in the editorial boards of: ACM Transactions of Information Systems (ACM), The Visual Computer (Springer), and Multimedia Tools and Applications (Springer). He is a member of the steering committee of CIVR, Computer Graphics International, and Multimedia Modeling conference series; and as member of International Review Panels of two large-scale research projects in Europe.