Sentiment-driven Community Profiling and Detection on Social Media

Amin Salehi, Mert Ozer, Hasan Davulcu
Computer Science and Engineering
Arizona State University, Tempe, AZ, USA
{asalehi1, mozer, hdavulcu}@asu.edu

ABSTRACT
Web 2.0 helps to expand the range and depth of conversation on many issues and facilitates the formation of online communities. Online communities draw various individuals together based on their common opinions on a core set of issues. Most existing community detection methods merely focus on discovering communities without providing any insight regarding the collective opinions of community members and the motives behind the formation of communities. Several efforts have been made to tackle this problem by presenting a set of keywords as a community profile. However, they neglect the positions of community members towards keywords, which play an important role for understanding communities in the highly polarized atmosphere of social media. To this end, we present a sentiment-driven community profiling and detection framework which aims to provide community profiles presenting positive and negative collective opinions of community members separately. With this regard, our framework initially extracts key expressions in users’ messages as representative of issues and then identifies users’ positive/negative attitudes towards these key expressions. Next, it uncovers a low-dimensional latent space in order to cluster users according to their opinions and social interactions (i.e., retweets). We demonstrate the effectiveness of our framework through quantitative and qualitative evaluations.

ACM Reference Format:

1 INTRODUCTION
With the advent of social media platforms, individuals are able to express their opinions on a variety of issues online. Like-minded users form online communities by interacting with each other and expressing similar attitudes towards a set of issues. While many methods [31] have been proposed to detect online communities, most of them do not provide insights into the collective opinions of community members. To shed light on such opinions, few efforts have focused on profiling communities, but a large body of work has been devoted to user profiling [17, 19, 26]. Indeed, "the founders of sociology claimed that the causes of social phenomena were to be found by studying groups rather than individuals" [18].

Turner et al. [43] suggest that individuals come together and form communities by developing shared social categorization of themselves in contrast to others. Therefore, to profile a community, we need to uncover the collective opinions of its members which makes them distinguishable from the members of other communities. Tajfel [40] suggests focusing on unit-forming factors (e.g., similarities, shared threats, or common fate) which function as cognitive criteria for segmentation of the social world into discrete categories. Accordingly, the controversial issues on which users have different opinions can be taken into account in order to discover the motives driving the segmentation of social media and the formation of communities. As a result, the profile of a community should present its important issues on which its members generally have the same position. Such community profiles can be found useful in a broad range of applications such as recommender systems [38], community ranking [6, 16], online marketing [20], interest shift tracking of communities [47], and community visualization [8]. For example, a group recommender system [3] can suggest more relevant items to communities by knowing the collective opinions of their members.

Many community detection methods [1, 5, 29, 30, 32, 33, 37, 46, 47] which are capable of community profiling have been proposed. However, these methods usually present a set of frequent keywords used by the members of a community as the community profile. However, it is common in social media that the members of different communities use the same keywords in their messages. Therefore, keywords alone might not be enough to differentiate communities in which their members have similar word usage. For instance, in the course of the US presidential election of 2016, Republicans and Democrats have used many common keywords such as Trump, Clinton, and Obamacare but with different sentiments. To differentiate and understand these two parties, not only keywords but also the collective attitude of community members towards these keywords should be taken into account.

In this paper, we tackle the aforementioned problem by proposing a sentiment-driven community profiling and detection framework which utilizes user-generated content and social interactions. Our framework first captures key expressions in users’ messages as representative of issues by utilizing a POS-tagger and built-in features of social media platforms (i.e. hashtags and user accounts). Next, it identifies users’ attitudes towards the extracted key expressions. Finally, we employ a novel graph regularized semi-nonnegative matrix factorization (GSNMF) technique to cluster users according to both their opinions and social interactions.
GSNMF uncovers not only communities but also their sentiment-driven profiles. The main contributions of the paper are as follows:

- Providing sentiment-driven community profiles which separately present the positive and negative collective attitudes of the members of each community towards their important key expressions;
- Achieving higher performance in detecting communities compared to several existing state-of-the-art community detection methods.

The rest of the paper is organized as follows. We review related work in Section 2. In Section 3, we propose our sentiment-driven community profiling and detection framework. To demonstrate the efficacy of our framework, we conduct quantitative and qualitative experiments by using real-world social media datasets in Section 4. Section 5 concludes the paper and discusses future work.

2 RELATED WORK

Community detection methods can fall into three broad categories: link-based, content-based and hybrid methods. Most of the existing works belong to the first category and utilize only social interactions [2, 7]. However, they neglect to utilize valuable user-generated content in which users express their opinions. On the other hand, content-based methods only utilize user-generated content [22]. Nevertheless, the content on social media is extremely noisy, resulting in the failure in detecting communities effectively. To alleviate these challenges, hybrid community detection methods are proposed. These methods are the most related work to our study since they not only exploit both user-generated content and social interactions but are also capable of profiling communities. These methods roughly fall into two categories: probabilistic graphical models and non-negative matrix factorization (NMF) based methods.

2.1 Probabilistic Graphical Models

Community User Topic (CUT) models [46] are one of the earliest works for detecting communities using probabilistic graphical models. The first proposed model (CUT_1) assumes that a community is a distribution over users, while the second one (CUT_2) considers a community as a distribution over topics. To discover communities, CUT_1 and CUT_2 are biased towards social interactions and user-generated content, respectively. Community Author Recipient Topic (CART) [32] is an unbiased model which assumes the members of a community discuss topics of mutual interests and interact with one another based on these topics. CART considers users as both authors and recipients of a message. However, in well-known social networks such as Twitter and Facebook, the number of recipients for a message can be very large. To make community detection scalable, Topic User Community Model (TUCM) [37], considering users as authors not recipients, is proposed. Since CART and TUCM consider users as authors, recipients, or both, they are limited to certain types of social interactions (e.g., retweet and reply-to in Twitter). The link-content model [29] solves this problem by ignoring the assumption that messages can be related to each other using social interactions. It is also capable of using different types of social interactions (e.g., friendship in Facebook and followship in Twitter). Furthermore, COCOMP [47] is proposed to model each community as a mixture of topics about which a corresponding group of users communicate. [5] is another model which detects and profiles communities in the domains having user-user, user-document, and document-document links.

2.2 NMF-based Methods

In order to encode graphs as local geometric structures, many methods extending standard NMF are proposed. LLNMF [15] introduces a regularizer, imposing the constraint that each data point should be clustered based on the labels of the data points in its neighborhood. GNMF [4] further incorporates a graph regularizer to encode the manifold structure. Moreover, DNMF [39] is proposed based on the idea that not only the data, but also the features lie on a manifold. The graph regularizers proposed by the above methods have been utilized by several other works [30, 33] to detect communities on social media. Moreover, another work [1] proposes a NMF-based approach utilizing a graph regularizer to exploit different social views (i.e., different social interactions and user-generated content) as well as prior knowledge in order to detect and profile communities.

3 THE PROPOSED FRAMEWORK

3.1 Problem Statement

We first begin with the introduction of the notations used in the paper as summarized in Table 1. Let \( \mathcal{U} = \{u_1, u_2, ..., u_n\} \) be the set of \( n \) users, \( \mathcal{C} = \{c_1, c_2, ..., c_k\} \) indicate the set of \( k \) communities, and \( \mathcal{S} = \{s_1, s_2, ..., s_m\} \) denote the set of \( m \) key expressions. \( X \in \mathbb{R}^{m \times n} \) indicates the matrix of users’ attitudes towards key expressions, where \( X_{ij} \) corresponds to the attitude of user \( u_i \) towards key expression \( s_j \). Furthermore, \( U \in \mathbb{R}^{n \times k} \) indicates the community membership matrix, in which \( U_{ik} \) corresponds to the membership strength of user \( u_i \) in community \( c_k \). \( V \in \mathbb{R}^{n \times m} \) further denotes the community profile matrix, where \( V_{lk} \) corresponds to the contribution strength of key expression \( s_l \) in the profile of community \( c_k \). Moreover, \( W \in \mathbb{R}^{m \times n} \) indicates the social interaction matrix, in which \( W_{ij} \) represents the number of social interactions between user \( u_i \) and user \( u_j \). We use \( \tilde{W} \) to denote the symmetric normalization of \( W \) (i.e., \( \tilde{W} = D^{-1/2}WD^{-1/2} \), where \( D \) is the degree matrix of \( W \)).

By using the above notations, the problem of detecting and profiling communities can be defined as: Given user opinion matrix

<table>
<thead>
<tr>
<th>Notation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{U} )</td>
<td>The set of users</td>
</tr>
<tr>
<td>( \mathcal{C} )</td>
<td>The set of communities</td>
</tr>
<tr>
<td>( \mathcal{S} )</td>
<td>The set of key expressions</td>
</tr>
<tr>
<td>( n )</td>
<td>The number of users</td>
</tr>
<tr>
<td>( m )</td>
<td>The number of key expressions</td>
</tr>
<tr>
<td>( k )</td>
<td>The number of communities</td>
</tr>
<tr>
<td>( X )</td>
<td>User opinion matrix</td>
</tr>
<tr>
<td>( U )</td>
<td>Community membership matrix</td>
</tr>
<tr>
<td>( V )</td>
<td>Community profile matrix</td>
</tr>
<tr>
<td>( W )</td>
<td>Social Interaction matrix</td>
</tr>
<tr>
<td>( D )</td>
<td>Symmetrically normalized matrix ( W )</td>
</tr>
</tbody>
</table>
X and social interaction matrix W, we aim to obtain community membership matrix U and community profile matrix V.

3.2 Extracting Key Expressions as Issues
Social media presents an opportunity to utilize user-generated content in which individuals express their opinions on various issues. The first step towards understanding users’ opinions is the extraction of the issues they discuss. To this end, many efforts [28, 35, 45] have been made to extract issues or related aspects. However, these methods require enough training samples for a specific domain to work accurately. Due to the lack of such a dataset for our required experiments, we follow a simple approach to extract key expressions. We utilize the built-in features common among well-known social media platforms. In such social networks, hashtags and user account mentions, which usually indicate issues, are perpendicularly tagged by ‘#’ and ‘@’, respectively. However, the built-in features are not enough to detect all issues. To tackle this problem, we employ a part-of-speech (POS) tagger to extract proper nouns and noun phrases (two or more nouns in a row) as representative of issues. If some proper nouns are in a row, they are considered as a single key expression. We utilize the POS tagger proposed in [13] proven to perform well for the content on social media.

3.3 Capturing Users’ Opinions
The position individuals take towards issues reflects their opinions. Many efforts [34, 41] have been made to detect users’ sentiments towards issues. However, these methods work effectively when enough training samples for a specific domain are given. However, there is no such a dataset for our required experiments so we apply a simple approach although a sophisticated approach can improve the result of our framework. First, a window with a certain size centered at each positive/negative sentiment word is created. Next, the nearest key expression to the sentiment word is selected, and the positivity/negativity of the sentiment word determines user’s positive/negative attitude towards that key expression. For instance, in the message “Conservatives seem angry every time economy adds jobs”, we assume the author has a negative sentiment towards key expression “Conservatives” because it is the closest key expression to the negative sentiment word “angry” if we consider the window size to be at least two. To generate matrix X, we need to apply the above procedure for all messages. Therefore, for each message if author $u_i$ takes a positive/negative attitude towards key expression $s_l$, we add the sentiment strength of the corresponding sentiment word to $X_{il}$, respectively. We utilize SentiStrength [42] to discover positive and negative words as well as their sentiment strength.

3.4 Exploiting Users’ Opinions
After extracting users’ attitudes towards key expressions, the next major objective is sentiment-driven community profiling and detection of like-minded users. To accomplish this, we exploit seminonnegative matrix factorization [10] as follows:

$$\min_{U,V} \|X - UV^T\|_F^2$$

s.t. $U \geq 0.$

(1)

Since the non-negativity constraint in Eq. (1) only holds on matrix U, matrix V can contain both positive and negative values. A positive/negative value of $V_{lk}$ denotes that the members of community $c_l$ have a collective positive/negative attitude towards key expression $s_k$. The larger the positive value of $V_{lk}$ is, the more the members of community $c_l$ have a collective positive attitude towards key expression $s_k$. The lower the negative value of $V_{lk}$ is, the more the members of community $c_l$ have a collective negative attitude towards key expression $s_k$. This property of matrix V also results in the categorization of key expressions into positive and negative categories according to the sign of the corresponding elements of key expressions in matrix V. Therefore, key expressions in a community profile are divided into two positive and negative categories. Moreover, the key expressions in each category can also be ranked by their values in matrix V in order to show how important they are to the members of the corresponding community.

3.5 Exploiting Social Interactions
Social interactions (e.g., retweets in Twitter and friendship in Facebook) are one of the most effective sources of information to detect communities [31]. To utilize social interactions, NMF-based methods exploit graph regularizers. Gu et al. [14] suggest that graph regularizers used in GNMF [4] and DNMF [39] suffer from the trivial solution problem and the scale transfer problem. When the graph regularizer parameter is too large, the trivial solution problem occurs and results in similarity among the elements of each row of community membership matrix U. The scale transfer problem, in which $\{V^*,U^*\}$ stands as the optimal solution for Eq. (1), results in a smaller objective value for the scaled transferred solution $\left(\frac{V}{\beta},\frac{U}{\beta}\right)$, for any real scalar $\beta > 1$.

To avoid these problems, we propose using the following graph regularizer,

$$\max_{U} \text{Tr}(U^T W U)$$

s.t. $U \geq 0, U^T U = I.$

(2)

where I is the identity matrix with the proper size. Eq. (2) clusters users into k communities, with the most interactions within each community and the fewest interactions between communities. In fact, Eq. (2) is equivalent to the nonnegative relaxed normalized cut as put forth in [9].

3.6 The Proposed Framework GSNMF
In the previous sections, we introduced our solutions to exploit social interactions and users’ attitudes toward key expressions. Using these solutions, our proposed framework simultaneously utilizes users’ opinions and social interactions to uncover communities and their profiles. The proposed framework requires solving the following optimization problem,

$$\min_{U,V} \mathcal{F} = \|X - UV^T\|_F^2 - \lambda \text{Tr}(U^T W U)$$

s.t. $U \geq 0, U^T U = I.$

(3)

where $\lambda$ is a non-negative regularization parameter controlling the contribution of the graph regularizer in the final solution. Since the optimization problem in Eq. (3) is not convex with respect to variables U and V together, there is no guarantee to find the global
Algorithm 1 The Proposed Algorithm for GSNMF

<table>
<thead>
<tr>
<th>Input: user opinion matrix X and social interaction matrix W</th>
</tr>
</thead>
<tbody>
<tr>
<td>output: community membership matrix U and community profile matrix V</td>
</tr>
<tr>
<td>1: Initialize U and V randomly where U ≥ 0</td>
</tr>
<tr>
<td>2: while not convergent do</td>
</tr>
<tr>
<td>3: Update U according to Eq. (4)</td>
</tr>
<tr>
<td>4: Update V according to Eq. (6)</td>
</tr>
<tr>
<td>5: end while</td>
</tr>
</tbody>
</table>

Optimal solution. As suggested by [21], we introduce an alternative scheme to find a local optimal solution to the optimization problem. The key idea is optimizing the objective function with respect to one of the variables U or V, while fixing the other one. The algorithm keeps updating the variables until convergence.

Optimizing the objective function \( F \) with respect to \( U \) leads to the following update rule,

\[
U = U \odot \left( \frac{(X^T V)^+ + [U(V^T V)^+] + \lambda W U + U \Gamma^-}{(X^T V)^- + [U(V^T V)^+] + U \Gamma^+} \right)
\]

where \( \odot \) denotes the Hadamard product, and

\[
\Gamma = U^T X^T V - V^T V + \lambda U^T W U
\]

We separate the negative and positive parts of a matrix \( A \) as \( A^- = (|A| - A)/2 \) and \( A^+ = (|A| + A)/2 \), respectively. The details regarding the computation of Eq (4) are given in Appendix A.

Moreover, optimizing the objective function \( F \) with respect to \( V \) leads to the following updating rule,

\[
V = XU(U^T U)^{-1}
\]

The details are given in Algorithm 1. In line 1, it randomly initializes \( U \) and \( V \). From lines 2 to 5, it updates \( U \) and \( V \) until convergence is achieved.

3.7 Algorithm Complexity

In Algorithm 1, the most costly operations are the matrix multiplications in update rules Eq. (4) and Eq. (6). Therefore, we provide the time complexity of these two updating rules as follows:

- The time complexity of Eq. (4) is \( O(nmk + mk^2 + n^2k + nk^2) \).
- Since the inversion of small matrix \( U^T U \) is trivial, the time complexity of Eq. (6) is \( O(nmk + nk^2) \).

Accordingly, the time complexity of Algorithm 1 is \( O(iknm + mk + n^2 + nk) \) where \( i \) is the number of iterations. Our framework can be applied to large scale social network platforms by exploiting the distributed approaches outlined in [12, 23, 25].

4 EXPERIMENTS

To evaluate the efficacy of our framework, we need to answer the following two questions:

1. How effective is the proposed framework in detecting communities compared to the state-of-the-art community detection methods?
2. How effective is our framework in profiling communities?

4.1 Data Description

We take politics as an example to evaluate our framework. In this regard, we used Twitter search API to crawl politicians’ tweets from three different countries, namely United States, United Kingdom, and Canada. However, Twitter API imposes the limitation of retrieving only the latest 3200 tweets for each user. To overcome this limitation, we crawled politicians’ user accounts several times during the time each dataset covers. The datasets are described as follows,

- **US Dataset** consists of the tweets posted by 404 politicians from two major political parties (Republican Party and Democratic party) in the United States from August 26 to November 29, 2016.
- **UK Dataset** consists of the tweets posted by 317 political figures from five major political parties (Conservative Party, Labour Party, Scottish National Party, Liberal Democratic Party, and UK Independence Party) in the United Kingdom from January 1 to September 30, 2015.
- **Canada Dataset** consists of the tweets posted by 102 politicians from three major political parties (Liberal Party, Conservative Party, and New Democratic Party) from January 1 to November 18, 2016.

All users in the datasets have discussed at least 15 key expressions. Moreover, the key expressions used by less than 15 users and stop words are eliminated. As a window size, we experimentally determine the threshold of 3 for the nearest keywords on both sides of each sentiment word. Furthermore, the party to which a user belong is labeled as ground truth. The statistics for the datasets are shown in Table 2. The GSNMF code and users’ Twitter accounts as well as their ground truth labels used in this paper are available.

4.2 Community Detection Evaluation

4.2.1 Baselines. In order to demonstrate the effectiveness of our framework, we compare GSNMF with the following state-of-the-art community detection methods,

- **GNMF** [4] is a hybrid method utilizing both user-generated and social interactions by incorporating a graph regularizer into standard NMF.

<table>
<thead>
<tr>
<th>Table 2: The statistics of the datasets.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>US</strong></td>
</tr>
<tr>
<td># of tweets</td>
</tr>
<tr>
<td># of retweets</td>
</tr>
<tr>
<td># of distinct words</td>
</tr>
<tr>
<td># of distinct key expressions</td>
</tr>
<tr>
<td># of users</td>
</tr>
<tr>
<td># of baseline communities</td>
</tr>
</tbody>
</table>

In the next sections, we first describe the datasets used in this study. Next, the performance of GSNMF is compared with several state-of-the-art community detection methods. Then, we qualitatively evaluate the community profiles uncovered by our framework.
• Louvain [2] is a link-based method optimizing modularity using a greedy approach.
• Infomap [36] is a link-based method built upon information theory to compress the description of random walks in order to find community structure.
• DNMF [39] is a hybrid method utilizing both user-generated content and social interactions by incorporating two regularizers (i.e., a graph regularizer and a word similarity regularizer) into standard NMF.
• Soft Clustering [44] is a link-based method that assigns users to communities in a probabilistic way.
• CNM [7] is a link-based method based on modularity optimization.

4.2.2 Evaluation Metrics. To evaluate the performance of the methods, we utilize three metrics frequently used for community detection evaluation; namely, Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and purity.

4.2.3 Experimental Results. For this experiment, we use all three datasets. We also utilize the party membership of each politician as ground truth in our evaluation. For the methods providing soft community membership, like our framework, we select the community with the highest membership value for each user as the community to which she/he belongs. Regularization parameters of NMF-based methods are set to be all powers of 10 from 0 to 9 to find the best configuration for each of these methods. We run each method 10 times with its best configuration and then report the best result. According to the results shown in Table 3, we can make the following observations,

• Our proposed framework achieves the highest performance in terms of NMI and ARI for all three datasets. In terms of purity, it also achieves the best in the Canada and US datasets. In the UK dataset, Louvain, Infomap, and CNM obtain higher purity compared to our framework since they generate an artificially large number of communities for sparse graphs such as social media networks. For instance, Louvain detects 21 communities for UK dataset.
• Exploiting both user-generated content and social interactions does not necessarily result in achieving better performance compared to link-based methods. For example, the Soft Clustering method achieves better results compared to GNMF and DNMF in terms of all three used metrics. However, link-based methods do not uncover any community profile.
• All NMF-based methods achieve their highest performance with large values (i.e., from $10^6$ to $10^9$) for the graph regularizer parameter.

4.3 Community Profiling Evaluation
In this section, we evaluate the effectiveness of our proposed framework in profiling communities by using US and UK datasets. In this regard, we first label each community detected by our framework with the party to which the majority of community members belong. Next, we evaluate how effectively the profile of a community represents its corresponding ground truth party. To this end, two graduate students who have knowledge of US and UK politics are assigned to label the results of community profiling methods. It is asked that each key expression in a community profile to be assigned to one of the following categories:

• Supported: A key expression is labeled as supported if the majority of community members have a positive attitude towards it or support it.
• Opposed: A key expression is labeled as opposed if the majority of community members have a negative attitude towards it or oppose it.
• Concerned: A key expression is labeled as concerned if the majority of community members are concerned about it.
• Unrelated: A key expression is labeled as unrelated if the annotators cannot find a strong relevance between the community (party) and the key expression.

In the tables representing community profiles, we color (and mark) supported, opposed, and concerned key expressions with green (+), red (−), and blue (±), respectively. We also leave unrelated key expressions uncolored (and unmarked).

In the following experiments, we expect our proposed framework to achieve three goals:

(1) Uncovering community profiles which represent the collective opinions of community members into two positive/negative categories;
(2) Assigning supported key expressions and opposed/concerned ones to positive/negative categories, respectively;
(3) Minimizing the number of unrelated key expressions in community profiles.

In Sections 4.3.1 and 4.3.2, we evaluate the results of GSNMF according to the first and second goals by using US and UK datasets. To evaluate the performance of the third goal, Section 4.3.3 compares GSNMF with the baselines with regard to their effectiveness in extracting relevant key expressions.

4.3.1 US Politics. The US dataset covers many events such as occurrences of gun violence, police brutality (e.g., the shooting of Terence Crutcher), the Flint water crisis, and the death of Fidel Castro; but the major event is the US presidential election of 2016. To give brief background knowledge, two major US parties during the election are described as follows [24],

• Democratic Party: A liberal party focusing on social justice issues. In 2016, Hillary Clinton was nominated as the presidential candidate of the party with Tim Kaine as her vice president. Moreover, Barrack Obama, the incumbent Democratic President, was a strong advocate for Hillary Clinton.
• Republican Party: A conservative party, known as the GOP, which had the majority of congressional seats in 2016 and embraces Judeo-Christian ethics. Moreover, Donald Trump
was nominated as the party candidate for the presidency with Mike Pence as his vice president.

During the campaign, Republicans—especially Donald Trump—mainly criticized President Obama and his policies (e.g., Obamacare, tax plans, and Iran deal) in order to discredit Hillary Clinton, whom they claimed was going to continue the Obama legacy and uphold the status quo [24]. On the other hand, Clinton’s campaign brought the issue of gun violence into the contest, and also focused on human rights for groups such as women and LGBTQ [24].

Table 4 shows the profiles of two communities detected by our framework in the US dataset as well as their corresponding ground truth political parties. According to the provided background, the community on the left highly resembles the Democratic Party since its members have generally expressed: 1) positive attitudes towards Hillary Clinton, the U.S. president (i.e., POTUS), Tim Kaine, and human rights issues (e.g., #WomensEqualityDay, #LGBTQ equality, and #NationalComingOutDay), and 2) negative attitudes towards the Republican Party (e.g., @HouseGOP and GOP), Donald Trump, and gun violence, police brutality (e.g., the shooting of Terence Crutcher). On the other hand, the community on the right highly resembles the Republican Party since its members have generally expressed: 1) positive attitudes towards the Republican Party (e.g., @HouseGOP and @SpeakerRyan), Donald Trump, Mike Pence, Congress, and religious (i.e., God), and 2) negative attitudes towards President Obama and his policies (i.e., Obamacare, tax code, Iran, Guantanamo Bay detention camp (i.e., GITMO)) as well as Hillary Clinton.

Negative sentiment implies both opposition and concern. If necessary, our framework can differentiate opposition from concern by providing the sentiment words frequently expressed by the members of a community towards each key expression. For example, Democrats’ negative sentiment towards Donald Trump mainly comes from the sentiment words “unfit”, “low”, and “dangerous” which suggest opposition. On the other hand, their negative sentiment towards #Trans (i.e., transgender people) mainly originates from the sentiment words “discrimination” and “murder” which indicate concern.

To demonstrate the advantage of our community profiling method, we compare the profiles of typical community profiles usually provided by retrospective studies with those uncovered by our framework. Table 5 shows the profiles of two communities detected by GNMF and DNMF in the US dataset as well as their corresponding ground truth political parties. As we observe, it is almost impossible for a non-expert individual to recognize the party associated with each profile since the position of the communities towards the key expressions are not taken into account. For example, in profiles corresponding to the Democratic Party and the Republican Party, many key expressions related to Trump, Clinton, and Obama exist, but there is no information regarding collective attitude of community members toward such key expressions. However, Table 4 shows that our proposed method correctly divides opposed/concerned key expressions and supported ones into the correct categories. Therefore, our framework makes it easy not only to differentiate and understand communities better but also to associate online communities with their real-world counterparts (if exist).

4.3.2 UK Politics. The UK dataset covers many events such as the rise of terrorism and terrorist attacks (e.g., CharlieHebdo and Tunisia attack), and many natural disasters (e.g., Nepal earthquake and Eboa) that happened in the first nine months of 2015. However, the major event in this period of time is the UK general election. Brief background knowledge about five major UK parties during the general election are provided as follows [27].

- **Conservative Party:** This party is also known as Tory and was led by David Cameron in 2015. David Cameron also led the UK government before and after the election of 2015. George Osborne, Nicky Morgan, and Jeremy Hunt were some of his secretaries.
- **Labour Party:** Ed Miliband was the leader of the Labour party for the election and selected Tom Watson as his deputy chair and campaign coordinator. Jeremy Corbyn, Yvette Cooper, Liz Kendall, and Andy Burnham were among the prominent members of the party.
- **Liberal Democrat Party:** Nick Clegg led the Liberal Democrat Party in 2015. Norman Lamb, John Leech, Nick Harvey, Tim Farron, and Charles Kennedy were some of the party’s parliamentarians.

| Table 4: The profiles of two communities detected by our framework in the US dataset. |
|---------------------------------|---------------------|---------------------|
| **Democrats** | **Republicans** |
| Positive | Negative |
| Positive | Negative |
| Hillary Clinton | Ziki | America | Obamacare |
| POTUS | Trump | @SpeakerRyan | ± BetterWay |
| America | @HouseGOP | Congress | ± Zina |
| #WomensEqualityDay | Donald Trump | @MikePence | Iran |
| NationalComingOutDay | Gun Violence | @RayTomPriece | Obama |
| Americans | #Trans | Tax code | ± Guantanamo Bay |
| #LaborDay | Gun Violence | Texas | ± Breast Cancer |
| TimKaine | Climate Change | ± VeteransDay | President Obama |
| Hillary | #Trajedores | ICTY | ± GITMO |
| Americans | TerenceCrutcher | Senate | Islamic |
| Cuba | GOP | ± LaborDay | State Sponsor |
| Halloween | Violence Situations | God | POTES |
| Veterans | ISIS | Constitution Day | ISIS |
| Florida | #TumpPin | USMAC | Hillary |
| ± LGBTQ equality | Donald | Thanksgiving | Fidel Castro |

Note: All colors, signs, and the name of parties in the table are ground truth.

| Table 5: The profiles of two communities detected by GNMF and DNMF in the US dataset. |
|---------------------------------|---------------------|---------------------|
| **Democrats** | **Republicans** |
| Positive | Negative |
| Positive | Negative |
| POTUS | Trump | @SpeakerRyan | ± BetterWay |
| America | @HouseGOP | Congress | ± Zina |
| #WomensEqualityDay | Donald Trump | @MikePence | Iran |
| Americans | #Trans | Tax code | ± Guantanamo Bay |
| #LaborDay | Gun Violence | Texas | ± Breast Cancer |
| TimKaine | Climate Change | ± VeteransDay | President Obama |
| Hillary | #Trajedores | ICTY | ± GITMO |
| Americans | TerenceCrutcher | Senate | Islamic |
| Cuba | GOP | ± LaborDay | State Sponsor |
| Halloween | Violence Situations | God | POTES |
| Veterans | ISIS | Constitution Day | ISIS |
| Florida | #TumpPin | USMAC | Hillary |
| ± LGBTQ equality | Donald | Thanksgiving | Fidel Castro |
| Hillary Clinton | Donald | ± Trump | DonaldTrump |
| Pence | #Trans | ISIS | Hillary |
| Americans | Russia | ± Putin |
| America | China | Iran | ± Venezuela |
| #WomensEqualityDay | #Trans | ± Trump |
| Americans | Russia | ± Putin |
| Pence | #Trans | ISIS | Hillary |
| ± Trump | Obama | ± Trump |
| Pence | Russia | ± Putin |
| America | China | Iran | ± Venezuela |
| #Trans | ISIS | Hillary | ± Trump |
| Pence | Russia | ± Putin |
| America | China | Iran | ± Venezuela |
| #WomensEqualityDay | #Trans | ISIS | Hillary |
| Pence | Russia | ± Putin |

Note: All colors, signs, and the name of parties in the table are ground truth.
• Scottish National Party: The SNP is a Scottish Nationalist party led by Nicola Sturgeon in 2015. Alan Brown and Neil Gray were some of the party’s parliamentarians.

• UK Independence Party: UKIP was led by Nigel Farage in 2015. The party embodies opposition to both United Kingdom EU membership and immigration.

Table 6 shows the profiles of five communities detected by our framework in the UK dataset as well as their corresponding ground truth political parties. As shown in the table, all parties have a common key expression, the general election of 2015 (e.g, GE2015 and GE15). We can also observe that the members of each party have generally expressed: (1) positive attitudes towards their party and also their prominent members and (2) negative attitudes towards other parties and their prominent members due to election competition [27]. Moreover, the government was a coalition between the Conservative party and the Labour party before the election. This coalition explains why they expressed positive sentiments towards the government related issues (i.e., Govt and Cameron) [27]. According to the negative attitudes of almost all parties, we can determine that the Conservative party and the Labour party are the ones towards which other parties expressed most negative sentiments. Furthermore, these two parties expressed a high negative sentiment towards each other. The reason behind this antagonism is that these parties are the two biggest parties having the highest chance of winning an outright majority in the election [27]. Moreover, UKIP’s negative view on Calais, the city in France where immigrants enter the UK, and Mediterranean immigration reflects its anti-immigration stance. Moreover, UKIP’s positive sentiment on Brexit and its negative sentiment on Greece indicates its anti-EU orientation.

Table 7 shows the profiles of five communities detected by GSNMF in the UK dataset.

4.3.3 Quantitative results. In this section, we aim to compare GSNMF with GSNMF and DNMF in terms of their effectiveness in extracting relevant key expressions for community profiles. Figure 1 shows the accuracy of all methods in US and UK datasets by considering different number of top key expressions as community profiles. As we observe, GSNMF outperforms GSNMF and DNMF in all experiments. For instance, by considering top 30 key expressions as community profiles, 93% of key expressions extracted by GSNMF in the US dataset are relevant compared to 82% in GSNMF and 85% in DNMF. Similarly, 83% of key expressions extracted by GSNMF in the UK dataset are relevant compared to 65% in GSNMF and 73% in GSNMF. The experiments also suggest that GSNMF achieves better associate the community profiles to the parties. However, as Table 6 suggests, the community profiles detected by our framework shows that the community associated to the Conservative party has positive attitude towards David Cameron but negative attitude towards Labour and Miliband. On the other hand, the community associated to the Labour party have the positive attitude towards Labour, UKLabour, and Miliband but negative attitude towards Tories and David Cameron. Since this corresponds to our ground truth, we can conclude that sentiment information can play an essential role in providing better community profiles.
accuracy with lower number of top key expressions as community profiles. This implies that the higher a key expression is ranked by GSNMF, the more likely it is relevant. Following these observations, sentiment-driven community profiling produces key expressions which are more relevant than its sentiment insensitive counterparts.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented a sentiment-driven community profiling and detection framework. Our framework uncovers a low-dimensional latent space in order to cluster users according to their opinions and social interactions. It also provides community profiles reflecting positive/negative collective opinions of their members. Experimental results on real-world social media datasets demonstrated: (1) our framework obtains significant performance in detecting communities compared to several state-of-the-art community detection methods, and (2) our framework presents a sentiment-driven community profiling approach that provides better insights into the collective opinions of community members by dividing key expressions into positive and negative categories.

Our future work includes the following directions. First, the current sentiment analysis is not capable of differentiating between opposition and concern. There is a need to propose new methods to differentiate between opposition and concern. Second, identifying the dynamics of communities sheds light on their temporal behavior. Therefore, we will focus our efforts on detecting and profiling the dynamics of communities.

A APPENDIX

A.1 Computation of U

Optimizing the objective function $F$ in Eq. (3) with respect to U is equivalent to solving

$$
\min_U F_U = ||X - UV^T||_F^2 - \lambda \text{Tr}(U^T\tilde{W}U)
$$

s.t. $U \geq 0, U^T U = I$. (7)

Let $\Gamma$ and $\Lambda$ be the Lagrange multiplier for constraints $U^T U = I$ and $U \geq 0$ respectively, and the Lagrange function is defined as

$$
\min_U \mathcal{L}_U = ||X - UV^T||_F^2 - \lambda \text{Tr}(U^T\tilde{W}U) - \text{Tr}(\Gamma(U^T U - I))
$$

The derivative of $\mathcal{L}_U$ with respect to $U$ is

$$
\frac{\partial \mathcal{L}_U}{\partial U} = -2X^T \tilde{V} + 2UV^T \tilde{V} - 2\lambda \tilde{W}U - \Lambda + 2U \Gamma
$$

By setting $\frac{\partial \mathcal{L}_U}{\partial U} = 0$, we get

$$
\Lambda = -2X^T \tilde{V} + 2UV^T \tilde{V} - 2\lambda \tilde{W}U + 2U \Gamma
$$

With the KKT complementary condition for the nonnegativity of $U$, we have $\Lambda_{ij} U_{ij} = 0$. Therefore, we have

$$
(-X^T \tilde{V} + UV^T \tilde{V} - \lambda \tilde{W}U + U \Gamma)_{ij} U_{ij} = 0
$$

where $\Gamma = U^T X^T \tilde{V} - V^T \tilde{V} + \lambda U^T \tilde{W}U$.

Matrices $\Gamma$, $X^T \tilde{V}$, and $V^T \tilde{V}$ take mixed signs. Motivated by [10], we separate positive and negative parts of any matrix $A$ as $A_{ij}^+ = (|A_{ij}| + A_{ij})/2$, $A_{ij}^- = (|A_{ij}| - A_{ij})/2$.

Thus, we get

$$
((X^T \tilde{V})^+ + [U(V^T \tilde{V})^+] + U \Gamma^+))_{ij} U_{ij} = 0
$$

which leads to the updating rule of $U$ in Eq (4).

A.2 Computation of V

Optimizing the objective function $F$ in Eq. (3) with respect to $V$ is equivalent to solving

$$
\min_V F_V = ||X - VU^T||_F^2
$$

The derivative of $F_V$ with respect to $V$ is

$$
\frac{\partial F_V}{\partial V} = -2XU + 2VU^T U
$$

Setting $\frac{\partial F_V}{\partial V} = 0$, we get the updating rule of $V$ in Eq (6).

Figure 1: The accuracy of community profiling methods in extracting relevant key expressions.
Sentiment-driven Community Profiling and Detection on Social Media

REFERENCES


[4] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. 2011. Graph regularized nonnegative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 8 (2011), 1548–1560.


