An Enhanced Topic Modeling Approach to Multiple Stance Identification

Junjie Lin[1,2], Wenji Mao[1,2] and Yuhao Zhang[1]
[1]State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, University of Chinese Academy of Sciences
[2]School of Computer and Control Engineering, University of Chinese Academy of Sciences Beijing, China
{linjunjie2013, wenji.mao, zhangyuhao2012}@ia.ac.cn

ABSTRACT

People often publish online texts to express their stances, which reflect the essential viewpoints they stand. Stance identification has been an important research topic in text analysis and facilitates many applications in business, public security and government decision making. Previous work on stance identification solely focuses on classifying the supportive or unsupportive attitude towards a certain topic/entity. The other important type of stance identification, multiple stance identification, was largely ignored in previous research. In contrast, multiple stance identification focuses on identifying different standpoints of multiple parties involved in online texts. In this paper, we address the problem of recognizing distinct standpoints implied in textual data. As people are inclined to discuss the topics favorable to their standpoints, topics thus can provide distinguishable information of different standpoints. We propose a topic-based method for standpoint identification. To acquire more distinguishable topics, we further enhance topic model by adding constraints on document-topic distributions. We finally conduct experimental studies on two real datasets to verify the effectiveness of our approach to multiple stance identification.

CCS CONCEPTS

·Information systems →Content analysis and feature selection; Document topic models; Sentiment analysis

KEYWORDS

Multiple stance identification; topic modeling; constrained Nonnegative Matrix Factorization

1 INTRODUCTION

People often exchange opinions and express their beliefs and attitudes online. The stances implied in online texts reflect the essential viewpoints they stand. Mining the stances in online texts can help us understand what people think and believe, and provide valuable information of their intrinsic judgments or inclinations. Therefore, stance identification is both an important research topic and beneficial to applications in business, public security, government decision making and many others.

Previous work on stance identification [1-8] has focused on analyzing people's attitudes towards a single topic or entity, and usually identifies stance as supportive or unsupportive attitude based on literal, syntactic and semantic information. Multiple stance identification, which aims to identify different standpoints of multiple parties, was largely ignored in previous research. In multiple stance identification, there may exist the standpoints of three or more parties. For example, during the 2016 American presidential election, people may take the standpoint of Hillary Clinton, Donald Trump, Bernie Sanders, Ted Cruz or other candidates. Thus computational methods developed for traditional stance classification are not directly applicable to the problem of multiple stance classification. Even for the case of two parties, if one manifests unsupportive attitude towards one party, it does not follow that he or she holds the standpoint of the other party.

Among the traditional work on stance identification, machine learning techniques play an important role. For example, Abbott et al. [1] train classifiers with linguistic features to determine the stances of texts regarding a single topic. Besides, some work leverages the relations between texts to improve the classification performance [4]. More recently, deep learning models have also been applied [7, 8]. Other work applies sentiment analysis to acquire the positive or negative sentiments towards a specific entity, and use them for entity-level stance identification [5].

To the best of our knowledge, the only work closely related to multiple stance identification was done by Lin et al. [9], which classifies the two-sided standpoints of online documents about the Palestinian-Israeli conflict. They propose a probabilistic model LSPM to recognize standpoint-related sentences to identify the standpoint of a document. However, their method is based on literal words and neglects the latent semantic information in texts. As their work explicitly considers the standpoints of two parties, it can be viewed as a special case of multiple stance identification.

In this paper, we aim at solving the problem of multiple stance identification. As people tend to support their standpoints through diverse topics, topics can provide distinguishable information for standpoint identification. Thus we propose a method to recognize distinct standpoints based on the topic information in texts. To acquire more distinguishable topics for

CIKM'17, November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133145

standpoint identification, we further enhance topic modeling by incorporating standpoint information and adding constraints on the mined document-topic distributions. Because of the wide usage and the flexibility of adding constraints directly, we choose to use the Nonnegative Matrix Factorization (NMF) based topic model [10] in our method. As it is often difficult to acquire labeled texts for standpoint identification, we also develop an automatic process to annotate the standpoints of texts with relatively high precision.

Our work has made several contributions. We first address the problem of multiple stance identification, which was almost totally ignored in previous research. We propose a topic-based method for multiple stance classification, and enhance the NMF based topic model to mine more distinguishable topics of multiple standpoints. We empirically evaluate our proposed method based on two online datasets with different standpoints.

2 PROPOSED METHOD

We define the problem of multiple stance identification as follows. Given a set of candidate standpoints and the input texts which belong to these standpoints, identify the standpoint of a new text. Standpoints can be represented as terms for specific entities or abstract beliefs/values, etc., depending on the application domain.

Below we propose a topic-based method to solve the multiple stance identification problem. Our method applies NMF based topic model to mine the latent topics in texts, and classifies the stances of texts based on topic similarities. To mine more distinguishable topics of different standpoints, we leverage the standpoint information to generate constraints on document-topic distributions and use them to enhance the topic model. Besides, as labeled texts are often difficult to acquire in practice, we develop an automatic process to annotate the texts with clear standpoints.

2.1 Topic-based standpoint identification

We adopt a four-step process for topic-based standpoint identification. First, a topic model is trained to mine the latent topics in input texts. Then the main topics of each standpoint and those of the new text are determined based on the trained model. Finally, the similarities between the main topics of the new text and those of multiple standpoints are calculated to identify the standpoint of the new text.

Step 1. We train an NMF based topic model to mine the latent topics in texts. The NMF based topic model factorizes a term-document matrix V into two non-negative matrixes W and H, so that $V \approx WH$, where W and H represent the term-topic and document-topic distributions.

Step 2. To determine the main topics of each standpoint, we first calculate its centric topic vector, which is the averaged topic distributions of the texts belonging to this standpoint. Then we choose the topics with top T values in the centric topic vector of this standpoint as its main topics, where T is the pre-defined number of main topics.

Step 3. For a new text $\boldsymbol{v}'$, we apply the trained NMF based topic model to acquire its topic distribution $\boldsymbol{h}'$. Similarly, we choose the topics with top T values in $\boldsymbol{h}'$ as the main topics of $\boldsymbol{v}'$.

Step 4. We calculate the similarities between the main topics of the new text and those of multiple standpoints, and choose the one with the maximum similarity as the identified standpoint. In the topic distribution of a text, the value of the dominant topic is sometimes much larger than those of other topics. In such case, other topics would not take effect in identifying standpoints. Thus, we choose to use ranking-based topic similarity instead of the similarity of topic distributions. The similarity measure is calculated on two ranked main topic lists, as given in equation (1), where l' is the ranked topic list of text $\boldsymbol{v}'$, l_i is the ranked topic list of standpoint s_i and $rank(t, l)$ is the ranking of topic t in list l.

$$similarity(l', l_i) = \sum_{t \in l' \wedge t \in l_i} \frac{1}{(1+rank(t,l')) \times (1+rank(t,l_i))} \quad (1)$$

Notice that $1 \leq rank(t, l) \leq T$, where T is the length of list l. We can see that two topic lists are similar if they share many common topics and the common topics rank high in both two lists.

2.2 Constrained NMF based topic modeling

The standard NMF model leverages no standpoint information in mining latent topics. To acquire more distinguishable topics for standpoint identification, we enhance the NMF-based topic model by adding standpoint-related constraints. We use the texts with standpoint labels to produce pairwise must-links and cannot-links as constraints. More specifically, we generate the must-link between two texts belonging to the same standpoint, and generate the cannot-links between two texts belonging to different standpoints. The must-link constraints help aggregate the common aspects of document-topic distributions of the same standpoint, while the cannot-link constraints help make the mined document-topic distributions of different standpoints distinct from each other.

The loss function $f(W, H)$ of the proposed constrained NMF model is given in equation (2). $M_H[i]$ and $C_H[i]$ denote the sets of must-link and cannot-link constraints about the i-th document d_i. $H_{:i}$ denotes the i-th column of H, which represents the topic distribution of document d_i. ρ_{MH} and ρ_{CH} are weighting factors measuring the importance of must-link and cannot-link constraints. ρ_W and ρ_H are regularization factors. This loss function indicates that if two documents are connected by a must-link, the Euclidean distance of their topic distributions should be close. On the contrary, if two documents d_i and d_j are connected by a cannot-link, the similarity between their topic distributions should be low.

$$\begin{aligned} f(W,H) = & \frac{1}{2}||V - WH||_2^2 + \rho_W||W|| + \rho_H||H|| \\ & + \frac{1}{2}\rho_{MH}\sum_i \sum_{j \in M_H[i]} ||H_{:i} - H_{:j}||_2^2 \\ & + \rho_{CH}\sum_i \sum_{j \epsilon C_H[i]} H_{:i} \cdot H_{:j} \\ & s.t.\ W, H \geq 0 \end{aligned} \quad (2)$$

Since this loss function is non-convex, we try to find a local optimization solution by the coordinate descent method, which is given in **Algorithm 1** below.

Algorithm 1: Coordinate descent method for constrained NMF

Input: term-document matrix V; must-link set M_H; cannot-link set C_H; weighting factors ρ_W, ρ_H, ρ_{MH}, ρ_{CH}; ε (set to 0.001)

Output: term-topic matrix W; topic-document matrix H

1. Initialize W and H with random values in [0.0,0.01);
2. **While** true **do:**
3. **For** each element W_{it} in W **do:**
4. $W_{it} \leftarrow \max\left(0, W_{it} - \frac{(WHH^T - VH^T)_{it} + \rho_W}{(HH^T)_{tt}}\right)$
5. **For** each element H_{tj} in H **do:** // $|\cdot|$ is the size of the set
6. $H_{tj} \leftarrow \max\left(0, H_{tj} - \frac{(W^TWH - W^TV)_{tj} + \rho_H + \rho_{MH}\sum_{k\in M_H[j]}(H_{tj} - H_{tk}) + \rho_{CH}\sum_{k\in C_H[j]} H_{tk}}{(W^TW)_{tt} + \rho_{MH}|M_H[j]|}\right)$
7. **If** $||\nabla^P f(W,H)||^2 \le \varepsilon||\nabla^P f(W^0,H^0)||^2$ **then break**; // $\nabla^P f(W,H)$ is the projected gradient for $f(W,H)$, W^0 and H^0 are the initial W and H
8. **Return** W, H.

2.3 Automatic standpoint annotation

In practice, it often lacks manually labeled texts for standpoint identification, thus we develop a simple but effective process to automate standpoint annotation of texts, and examine whether our proposed method works well with noisily-labeled texts. As the sentiments people express towards certain parties often reveal their standpoints, we annotate the texts with clear standpoints by dictionary-based sentiment analysis. We first take the root of a sentence's syntactic dependency tree as its core word, and use the sentiment of the core word as the sentence's main sentiment. To find the subject (i.e. the targeting party in stance identification) of the sentence's main sentiment, we then take the noun phrase nearest to the core word in the sentence's phrase-structure tree. We finally annotate the standpoint of a text based on its sentences that express non-negative sentiments towards the standpoint parties. Moreover, as a person's standpoint is relatively stable, we annotate the standpoint of an unlabeled text of the person using the majority standpoint of all his/her annotated texts.

For short texts, syntactic parsing may be inaccurate as the language usage is quite informal. Considering that the sentiment expressions in short texts are often direct and explicit, we annotate the standpoint of a sentence if it mentions standpoint parties and its overall sentiment is non-negative.

3 EXPERIMENTS

3.1 Datasets

We use two datasets for our experimental studies. The first dataset contains online documents about the Palestinian-Israeli conflict collected from the "bitterlemons.org" website ("Bitterlemons" dataset), which has been used in the related work of Lin et al. [9]. This website provides the standpoint of each document, either "Palestinian" or "Israeli". We crawl all the 1765 documents, which are published by 335 users from 2001 to 2012. The numbers of documents with "Palestinian" and "Israeli" standpoints are 882 and 883 respectively.

The second dataset is the Twitter dataset about the 2016 American presidential election. We take the names of four candidates (i.e. Hillary Clinton, Bernie Sanders, Donald Trump and Ted Cruz) as standpoint keywords to crawl the relevant tweets. All the crawled tweets are published from March 30th to June 6th in 2016, among which we randomly select 25000 tweets to construct our dataset. To collect the test data, we invite two senior graduates majored in Social Media Analysis to label the standpoints of 4000 randomly chosen tweets.

3.2 Experimental results and analysis

We first evaluate the automatic standpoint annotation process using all the documents from the "Bitterlemons" dataset and the labeled tweets from the Twitter dataset. The precisions of automatic standpoint annotation are 78.01% on the "Bitterlemons" dataset and 85.07% and 87.66% on the Twitter dataset labeled by two raters respectively. It can be seen that our automatic annotation process can achieve relatively high precision to facilitate further topic-based multiple stance identification.

To test the performance of our approach to multiple stance identification, we first take traditional work on stance detection for comparison, including two representative sentiment analysis based methods (i.e. WordDist and DepDist) [5] and three state-of-the-art machine learning based methods (i.e. SVM$^+$ [6], pkudblab [7] and BiConditional [8]). WordDist and DepDist identify standpoints based on the sentiment words in sentences [5]. SVM$^+$ and pkudblab are top-ranked methods in the stance detection task A and B of SemEval 2016. BiConditional is a deep learning model, which has been a widely used strong comparative method in stance detection. Besides, we take the closest related work [9] to multiple stance identification (i.e. LSPM) for comparison.

We use both the "Bitterlemons" and Twitter datasets for the evaluation of multiple stance identification. For the former, we randomly choose 66% of the documents for training and use the left for testing. For the latter, we use the standpoint-related tweets labeled by the two raters as the testing data, and use the rest of the tweets as the training data. We use the above training data for automatic standpoint annotation and the training of constrained NMF model and the comparative standpoint classifiers.

In our method, we determine the values of parameters by 5-fold cross validation on the annotated data. For the "Bitterlemons" dataset, we set the topic number $K = 50$, $\rho_W = \rho_H = 0.01$ and $\rho_{MH} = \rho_{CH} = 0.1$. For the Twitter dataset, we set the topic number $K = 100$, $\rho_W = \rho_H = 0.5$ and $\rho_{MH} = \rho_{CH} = 1.0$. We set the number of main topics in calculating topic similarity to 10. In particular, we also compare the performances of our NMF based topic models without must-link and cannot-link constraints, with only must-link constraints and with only cannot-link constraints. The macro-averaged F1-values of standpoint identification by our method and comparative methods are given in Table 1.

Table 1: Macro-F1 of our method and comparative methods

Methods		Bitterlemons dataset	Twitter dataset	
			Rater 1	Rater 2
WordDist		30.41%	63.02%	66.03%
DepDist		45.50%	66.07%	69.43%
SVM+		73.42%	69.39%	72.90%
pkudblab		67.66%	66.85%	71.81%
BiConditional		78.41%	74.75%	76.25%
LSPM		73.80%	71.28%	73.23%
Our method	without constraint	62.64%	73.51%	78.45%
	must-link only	71.27%	75.55%	78.80%
	cannot-link only	70.05%	75.11%	79.07%
	both constraints	**83.41%**	**77.33%**	**79.25%**

We can see from Table 1 that among the comparative methods, sentiment analysis based methods achieve relatively low F1-value. This indicates that sentiment information is insufficient for standpoint identification. Our topic-based method without any constraints outperforms most comparative methods in the Twitter dataset, indicating that the mined topics can provide useful information for standpoint identification. Moreover, our method with both must-link and cannot-link constraints performs the best among all the methods in both datasets. One possible reason is that most comparative methods take the noisily-labeled texts as the ground-truth data to train classifiers, while our constrained NMF model uses the noisily-labeled texts to generate soft constraints on document-topic distributions, which makes our method more robust. In general, the experimental results verify the effectiveness of our approach to multiple stance identification.

Table 2 provides some illustrative examples to show the mined main topics in the "Bitterlemons" dataset by our method. Each topic is represented by several keywords.

Table 2: Example main topics in "Bitterlemons" dataset

Standpoints	Topics
Palestine	P1:{iraq, america, war, weapon, afghanistan, ⋯}
	P2:{occupation, territory, illegal, settlement, ⋯}
	P3:{sharon, likud, labor, coalition, dismantling, ⋯}
Israel	I1:{terrorist, gaza, hamas, suicide, bombing, ⋯}
	I2:{economic, government, failed, concession, ⋯}
	I3:{bombing, terrorist, barrier, security, build, ⋯}

From Table 2, we can see that the mined main topics reflect the opinions or facts behind people's standpoints. Topics P1 and I1 are both about the "terrorism" issue. In Topic P1, Palestinians criticize Israel for providing weapons to America to attack Afghanistan, while in Topic I1, Israelis criticize the suicide bombing conducted by Hamas, a Palestinian organization. Besides, Palestinians express negative opinions towards Israel's territory occupation (P2). Palestinians also talk about the dismantlement of Likud, which is an unfavorable fact to Israel (P3). Israelis criticize the severe economic condition of Palestine (I2), and support their standpoints by talking about building the barrier walls to defend the attack from Palestine (I3). In summary, the mined topics can provide distinguishable information for standpoint identification.

4 CONCLUSION

Stance identification is an important research topic and beneficial for many applications. Previous research on stance identification focuses on binary attitude classification towards a certain topic or entity. Multiple stance identification, which recognizes different standpoints of multiple parties in texts, reveals people's intrinsic judgments and inclinations. In this paper, we first address the problem of multiple stance identification, and propose a method to identify the standpoints of texts based on topic information. In order to make the mined topics distinguishable for identification, we enhance the NMF based topic model by generating constraints using texts of different standpoints. Experimental studies verify the effectiveness of our approach to multiple stance identification. Finally, our future work shall incorporate sentiment information to further enhance our approach.

ACKNOWLEDGMENTS

This work is supported in part by NSFC Grant #71621002, #61671450 and #71472175, the Ministry of Science and Technology of China Major Grant #2016QY02D0205, CAS Key Grant #ZDRW-XH-2017-3 and grant #2017A074.

REFERENCES

[1] R. Abbott, M. Walker, P. Anand, J.E.F. Tree, R. Bowmani, and J. King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*. 2-11.

[2] K.S. Hasan and V. Ng. 2013. Stance classification of ideological debates: data, models, features, and constraints. In *Proceedings of the 2013 International Joint Conference on Natural Language Processing*. 1348-1356.

[3] L. Wang and C. Cardie. 2014. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of ACL Workshop on Computational Approaches To Subjectivity, Sentiment and Social Media Analysis*. 97-106.

[4] D. Sridhar, L. Getoor, and M. Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 109-117.

[5] H.L. Hammer, P.E. Solberg, and L. Øvrelid. 2014. Sentiment classification of online political discussions: A comparison of a word-based and dependency-based method. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 90-96.

[6] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 2016 International Workshop on Semantic Evaluation*. 31-41.

[7] W. Wei, X. Zhang, X. Liu, W. Chen, and T. Wang. 2016. Pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 2016 International Workshop on Semantic Evaluation*. 384-388.

[8] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 876-885.

[9] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. 2006. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning*. Association for Computational Linguistics. 109-116.

[10] D. Kuang, J. Choo, and H. Park, 2015. Nonnegative matrix factorization for interactive topic modeling and document clustering. *Partitional Clustering Algorithms*, 215-243.