Topic and User Based Refinement for Competitive Perspective Identification

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

***Abstract*—*The competitive perspectives implied in online texts reflect people's conflicts in their stances and viewpoints. Competitive perspective identification aims to determine people's inclinations to one of multiple competitive perspectives, which is an important research issue and can facilitate many security-related applications. As the word usage of different perspectives is distinct in various topics, in this paper, we first propose a supervised topic-refined method for competitive perspective identification. Our method refines perspective classifiers with the document-topic distributions mined from texts. To reduce human labor in data annotation, we further extend our work in a semi-supervised manner and propose a user-based bootstrapping framework. As the perspectives people hold are relatively stable, our bootstrapping process leverages the user-level perspective consistency to select high-quality classified texts from unlabeled corpus and boost the perspective classifier iteratively. Experimental studies show the effectiveness of our proposed approach in identifying the competitive perspectives of online texts.*

Keywords—Competitive perspective identification; Topic-based refinement; User-based bootstrapping

I. INTRODUCTION

People often publish texts on the Web to express their attitudes and exchange opinions. The perspectives from which people write online texts reflect the fundamental stances and essential viewpoints they stand, and competitive perspectives, in particular, reflect the conflicts in people's stances and viewpoints [1]. Identifying the competitive perspectives people hold can provide valuable information about their intrinsic judgments or inclinations. Therefore, competitive perspective identification is an important research issue and can facilitate many security-related applications, such as decision making, policy suggestion and emergency response.

Competitive perspective identification was rarely studied in previous research, while some efforts have been made in a related research field, stance detection [2-4]. The stances people take often manifest as their supportive or unsupportive attitudes towards certain topics or entities. Different from the stances, the competitive perspectives people hold indicate their inclinations to one of multiple competitive entities (such as political parties and organizations), which are more intrinsic and macroscopic. The perspectives people hold can influence their stances towards different topics. Specifically, stance detection aims to classify the stance of a text towards one certain topic or entity as supportive, unsupportive or none. However, competitive perspective identification focuses more on determining the perspective of a text among multiple competitive entities. The texts people write to express their perspectives usually cross various topics, and there can be two or more competitive perspectives. Thus the computational methods developed for stance detection are not directly applicable to competitive perspective identification.

Previous work on stance detection typically uses machine learning and takes n-grams as features [2, 3]. Syntactic and semantic features have also been incorporated to improve the performance of stance classifiers [4]. Similarly, previous work on competitive perspective identification applies machine learning techniques to build the identification model based on word features [5]. On the basis of this, Lin et al. [1] mine latent topic information in texts to take into account different word usage under various topics.

In the work by Lin et al. [1], topic information under each perspective is mined for identifying the specific perspective. Meanwhile, at the general topic level, the word usage of different perspectives is distinct. To make use of the general topic information across different perspectives, in this paper, we first propose a topic-refined method which mines perspective-independent topics for classification. On the other hand, the existing methods of competitive perspective identification are supervised [1, 5], which need large labeled corpora for model training. To reduce human labor in labeling training data, we further propose a semi-supervised bootstrapping framework for competitive perspective identification. As the perspectives people hold are relatively stable, the perspectives of different texts written by the same author should be consistent. Therefore, our bootstrapping process leverages the user-level perspective consistency to select high-quality classified texts from unlabeled corpus and boost the perspective classifier iteratively. Experimental studies show the effectiveness of our topic-refined model and user-based bootstrapping process in identifying competitive perspectives of online texts.

II. PROPOSED APPROACH

To solve the competitive perspective identification problem, we propose a user-based bootstrapping framework, in which we also develop a topic-refined identification method. The main process of our proposed approach is given in Fig. 1. This bootstrapping process adopts an iterative strategy to boost the perspective classifier. In the beginning, we initialize the training corpus with a few labeled seed texts. In each iteration, we

first mine the latent topic information in the training texts and construct the topic-refined perspective classifier. Then we use the constructed classifier to identify the perspectives of the unlabeled texts. To expand the training corpus with high-quality texts, we leverage the user information to measure the classification confidence of the unlabeled texts, and add the highly confident classified texts to the training corpus so as to reconstruct the perspective classifier in the next iteration.

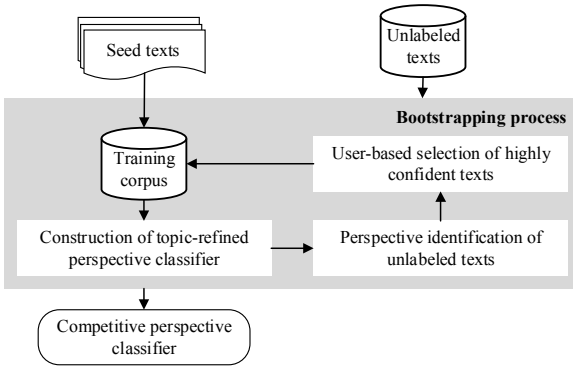

Fig. 1. Main process of the proposed approach

A. Construction of topic-refined perspective classifier

To refine the perspective classifier with perspective-independent topic information, we apply the Non-negative Matrix Factorization (NMF) based topic model to acquire the document-topic distributions of all the training texts, and use these distributions to construct a fine-grained model for competitive perspective identification. The NMF based topic model decomposes the original term-document matrix V into a term-topic matrix W and a topic-document matrix H so that $V \approx WH$ and $W, H \geq 0$. By normalizing each column in H, we can get the topic distribution $\boldsymbol{h_i}$ of each text d_i.

As the word usage in perspective expression is different under various topics, we train a perspective classifier for each topic respectively. When training the i-th topic-related perspective classifier, we take the topic distributions $h_{1i}, h_{2i}, \dots, h_{Ni}$ of texts $d_1, d_2, \dots, d_N$ as the sample weights. In doing so, we can distinguish the importance of different texts to the classifiers. In addition, we train a general perspective classifier to capture the common word usage in expressing perspectives.

B. Perspective identification of unlabeled texts

We use the trained topic model and perspective classifiers to identify the competitive perspectives of unlabeled texts. For an unlabeled text d', we apply the trained NMF model to acquire its topic distribution $(h'_1, \dots, h'_K)$, where K is the predefined number of topics. Meanwhile, we use the general and topic-related perspective classifiers to identify the perspective of this text, denoted as $\boldsymbol{p_0}$ and $\boldsymbol{p_1}, \dots, \boldsymbol{p_K}$. Note that $\boldsymbol{p_0}$ and $\boldsymbol{p_1}, \dots, \boldsymbol{p_K}$ are one-hot vectors where the element corresponding to the identified perspective is 1 and the other elements are 0. Finally, to determine the perspective of text d' based on the outputs of the general and topic-refined perspective classifiers, we adopt a topic-based voting scheme, which is given by equation (1) and (2), where $perspective(d')$ is the identified perspective of text d' by our method, and α is a weighting factor which measures the importance of the general classifier. It can be seen that we use the topic distribution as the voting weights in determining the final perspective of an unlabeled text.

$$\boldsymbol{p}' = \sum_{i=1}^{K} h'_i \cdot \boldsymbol{p_i} + \alpha \cdot \boldsymbol{p_0} \tag{1}$$

$$perspective(d') = \underset{i}{argmax}\, p'_i \tag{2}$$

C. User-based selection of highly confident texts

To iteratively boost the perspective classifier with additional labeled texts, we add the classified texts of high confidence to the training corpus at the end of each iteration. As the perspectives of different texts written by the same author are usually consistent, we leverage the author information to measure the classification confidence of the unlabeled texts. Specifically, we calculate the confidence scores of all user u's texts by equation (3), where c_i^u is the number of u's texts classified as perspective i and σ is a shrinking factor. This equation indicates that if most of user u's texts are classified as the same perspective, and this user has written adequate texts, the classification confidence of this user's texts is high.

$$confidence(u) = \frac{\max_i c_i^u}{\sum_i c_i^u} \times \frac{1}{1+e^{-\frac{\sum_i c_i^u}{\sigma}}} \tag{3}$$

Based on the perspective classification results, we group authors by their perspectives and sort the authors of each perspective in the descending order according to their confidence scores. Note that we take the majority perspective of a user's texts as his/her perspective. Then for each perspective, we fetch the first author in the sorted author list iteratively and expand the training corpus with his/her texts, until the amount of the added texts is more than a pre-defined threshold $TOPN$. To ensure the balance of different classes, the bootstrapping process ends once there exists a perspective that no more corresponding texts can be added to the training corpus. Finally, to make full use of all the unlabeled texts, we classify the perspectives of the remaining unlabeled texts by the latest trained perspective classifier and add them to the training corpus, based on which we construct the final perspective classifier.

III. Experiment

A. Dataset

We use a dataset crawled from the "bitterlemons.org" website for our experimental studies, which was also used in the related work [1, 5]. This website publishes online documents about the Palestinian-Israeli conflict, and it provides the perspective of each document, either *Palestinian* or *Israeli*. We crawl all the documents along with their perspectives and author information. Our dataset contains 1765 documents written by 335 users from 2001 to 2012. Among the authors, two of them are resident editors and the others are guests. To leverage more author information for our evaluation, we use the documents written by the guests as the training data and use the documents written by two editors as the testing data. In our dataset, there are 436 *Palestinian* documents and 437 *Israeli* documents in the training corpus, and there are 446 *Palestinian* documents and 446 *Israeli* documents in the testing corpus.

B. Evaluating topic-refined perspective identification method

To test the effectiveness of our topic-refined competitive perspective identification method (abbreviated as "TR_CPI"),

we employ typical machine learning algorithms to construct the perspective classifiers and examine whether our topic-refined method can boost their performances, including Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF). To be consistent for comparison, all these methods take word unigrams as features.

For our method, we determine the parameter values by 5-fold cross validation on the training data, and set the topic number $K = 50$ and the weighting factor $\alpha = 0.5$. The accuracies of our method with different perspective classifiers are given in Table I, where "Original" refers to the original classifiers without our topic-based refinement.

TABLE I. ACCURACIES (%) OF TR_CPI AND ORIGINAL METHODS

Method	NB	LR	SVM	RF
Original	95.07	96.08	95.63	96.08
TR_CPI	**95.27**	**96.19**	**95.74**	**97.09**

From Table I, we can see that our topic-refined method can improve the performances of typical machine learning based perspective classifiers. Among all the original perspective classifiers, Random Forest achieves the highest accuracy. Our topic-refined method with Random Forest outperforms all the other methods, which demonstrates the effectiveness of our topic-based refinement for competitive perspective identification.

C. Evaluating the user-based bootstrapping process

To evaluate the performance of the user-based bootstrapping process (UBP) for semi-supervised competitive perspective identification, we randomly select 10% of the training data as the labeled seed texts and use the remaining training texts in an unsupervised manner. Specifically, we compare UBP with three methods. The first method is fully supervised (FS) and it only takes the seed texts for model training. The second method is semi-supervised (SS), which first trains an initial classifier based on the seed texts, and then uses this classifier to annotate the unlabeled texts in the training corpus. It finally uses the annotated texts along with the seed texts to train a classifier for perspective identification. The third method is similar to the second method except for adjusting the perspective of each annotated text to the majority perspective of its author (SS_U).

To examine whether the proposed bootstrapping process works well with different perspective classifiers, we take the topic-refined versions of typical machine learning based methods (abbreviated as "T_NB", "T_LR", "T_SVM" and "T_RF") as the perspective classifiers in our framework UBF and the comparative methods FS, SS and SS_U. In addition, we take two methods in the related work [1, 5] (i.e. LSPM and NB_T) as the perspective classifiers in UBF, FS, SS and SS_U.

For our bootstrapping process and topic-refined method, we determine the parameter values by 5-fold cross validation on the training data, and set the topic number $K = 30$, the weighting factor $\alpha = 0.5$ and the shrinking factor $\sigma = 2$. To balance the efficiency and effectiveness of the bootstrapping process, currently we set the amount of the added texts in each iteration (i.e. $TOPN$) to 30. The accuracies of our bootstrapping process and comparative methods with different perspective classifiers are given in Table II.

TABLE II. ACCURACIES (%) OF UBP AND COMPARATIVE METHODS

Method	T_NB	T_LR	T_SVM	T_RF	LSPM	NB_T
FS	84.97	85.43	85.54	90.36	82.53	83.28
SS	84.69	84.53	84.64	88.68	82.17	82.67
SS_U	84.69	88.90	90.92	92.83	82.83	83.88
UBP	**93.47**	**93.51**	**92.78**	**95.70**	**91.79**	**93.34**

From Table II, we can see that our user-based bootstrapping process outperforms all the comparative methods. The semi-supervised method (i.e. SS) seldom achieves better performance than the fully supervised method (i.e. FS). One reason for this is that the texts annotated by the semi-supervised method is too noisy. The semi-supervised method with author-based perspective adjustment (i.e. SS_U) outperforms the purely semi-supervised method in most cases. It can also be seen that our bootstrapping process works well with different perspective classifiers. This indicates the generality of our user-based bootstrapping process in identifying competitive perspectives.

IV. CONCLUSION

Competitive perspective identification is an important research topic and beneficial for many security-related applications. As the word usage of different perspectives is distinct in various topics, we first propose a supervised topic-refined method for competitive perspective identification. In this method, we mine general topics implied in texts and refine the perspective classifiers with document-topic distributions. As supervised method usually needs large labeled corpus for model training, which is time-consuming and labor-intensive, we further extend our work in a semi-supervised manner. Based on the consideration that the perspectives people hold are relatively stable, we propose a user-based bootstrapping framework which requires only a few labeled seed texts. In the bootstrapping process, we leverage the user-level perspective consistency to select highly confident classified texts and boost the perspective classifier iteratively. Experimental studies demonstrate the effectiveness of our topic-refined method and user-based bootstrapping process in identifying competitive perspectives.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Junjie Lin, Wenji Mao, Daniel Zeng, Competitive perspective identification via topic based refinement for online documents, in Proc. 2016 IEEE Conference on Intelligence and Security Informatics, 2016, pp. 214-216.

[2] Saif M Mohammad, Parinaz Sobhani, Svetlana Kiritchenko, Stance and sentiment in tweets, arXiv preprint arXiv:1605.01655, 2016.

[3] Javid Ebrahimi, Dejing Dou, Daniel Lowd, A joint sentiment-target-stance model for stance classification in tweets, in Proc. the 26th International Conference on Computational Linguistics, 2016, pp. 2656-2665.

[4] Kazi Saidul Hasan, Vincent Ng, Stance classification of ideological debates: data, models, features, and constraints, in Proc. the 2013 International Joint Conference on Natural Language Processing, 2013, pp. 1348-1356.

[5] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, Alexander Hauptmann, Which side are you on?: Identifying perspectives at the document and sentence levels, in Proc. the 10th Conference on Computational Natural Language Learning, 2006, pp. 109-116.