Multi-Source Emotion Tagging for Online News

Li Yu*, Xue Zhao*, Ying Zhang*†
*College of Computer and Control Engineering, Nankai University
†College of Software, Nankai University

***Abstract*—With the rapid growth of social media and online news services, users nowadays can respond to online news by rating subjective emotions such as happiness, surprise or anger actively. Once the user ratings is over a certain range, it begins to show up a tendency of what most people think and feel, which can help us understand the preferences and perspectives of most users, and help news providers to provide users with more positive news. Thus it has become a pregnant research problem to tag emotion automatically. This paper tackles the task of emotion tagging for online news with multi-source including news article and comment, as emotion is not only tagged after reading news article, but also can be incorporated in comment with what they feel. In this paper, a novel classification model are proposed with two layer logistic regression. The new approach get outputs from basic classifiers and combine them in a new classifier, making a more accurate prediction when compared with a single source method. An extensive set of experimental results on a real dataset from a popular online news service demonstrate the effectiveness of the proposed approach.**

Keywords*—Sentiment Tagging; Online News; Meta Classification; Multiple Source

I. Introduction

With the explosive growth of the Internet, a large number of social media have been produced and its influence continues to increase. Online news is an important form that attracts millions of users to read and express their feelings among kinds of social media. Users often express their subjective emotions such as sadness, surprise and happiness by rating emotion tags after reading news. Figure 1 shows an example emotion tagging from a popular Chinese news website (i.e. The Society Channel of Sina News[1]).

Fig. 1. An example of emotions and user ratings

Sentiment analysis of online documents such as news articles, blogs and microblogs has received increasing attention in recent years and it is a challenging research problem to do emotion tagging for online news automatically. Naturally, we classify news to each emotion tags based on the content of news article, as the news article is what they read and directly influences their emotion rating. What is more, an increasing number of social news websites provide comment service where users can share what they feel after reading news articles. Thus news comment provide another significant source for emotion tagging.

Sentiment classification model generally fall into two categories, generative models and discriminative models. As the second one has attractive theoretical properties, we adopt logistic regression as our basic method. Two basic classifiers are built to get the outputs of single source. Through utilizing the outputs of multiple sources including news article and comment, we propose a novel approach by two-layer prediction. Since different source has its own effect on emotion prediction in different situations, we match different source with different weighting parameters, making it more adaptable.

In addition, we provide an experiment on a real dataset after detailed analysis. In the experiment, we introduce the experimental environment and give the statistic data on dataset. Our proposed approach with multi-source are compared with basic approaches with single source in different size of dataset. Finally we verify the necessity of multi-source for emotion tagging and the effectiveness of the proposed model after experimenting on the dataset, society channel of Sina News .In particular, the proposed model has shown better performance on Sina dataset than baselines. The robustness of the proposed discriminative meta classification models is evaluated with different sizes of dataset.

The contribution of this paper can be listed as followed,

- Our work consider the difference between different sources and give separated weighting parameters to each source.
- Our work consider to be generalized without the impact of language.
- Our work has been approved to be effective on Sina dataset thus can be applied on other multi-source applications.

The rest of the paper is organized as follows. In Section II, we review the state of the art on emotion tagging and make discussions regarding the difference between our work and previous work. In Section III, we propose our novel method with two layers. The experimental results are presented and discussed in Section IV. Conclusions and future work are given in the last section V.

II. Related Work

The Task of emotion tagging is to give an emotion label to text content. Early work in Ekman [9] defined six basic emotion labels: happiness, sadness, anger, fear, surprise and

[1]http://news.sina.com.cn/society/

disgust, which laying a good foundation in some existing work on emotion tagging including Liu et al. [14], Das and Bandyopadhyay [7, 8], Aman and Szpakowicz [2]. Specifically, Liu et al. [14] used an ensemble of rule-based affect models to classify sentences into six basic emotion labels. Another kind of methods based on supervised learning considering the problem as text categorization or analogous to topic classification Alm et al. [1] has also gone a long way. Along the way of this kind of method, Das and Bandyopadhyay [7, 8] used Naive Bayes and Support Vector Machines based on some online dictionary. Aman and Szpakowicz [2] proposed a method of building an dictionary derived from the *Roget Thesaurus* Roget [17].

In addition, Yang et al. [19] trained SVM model to investigate the emotion classification of web blog. Another work Yang et al. [20] used Yahoo Kimo Blog corpora to build emotion dictionary. By considering the relationships between words in a sentences, Neviarouskaya et al. [16] proposed a syntactical approach to recognition emotion automatically. Mishne and de Rijke [15] used several supervised and unsupervised machine learning techniques on blog data for comparative evaluation. Binali and Potdar [5] compared various text-based emotion detection techniques.

The 4th International Workshop on Semantic Evaluations in 2007 organized an evaluation campaign on Semeval Task for “Affective Text” (Task 14) [2]. The main task is the emotion classication of news headlines downloaded from some well-known news web sites(i.e. New York Times, CNN, and BBC News), as well as from the Google News search engine with Ekman’s six emotion labels. Several participants experimented various techniques for the task. [6] developed a rule-based linguistic system by using a set of lexical resources. [11] gathered statistics from three commercial search engines to determine the kind and the amount of emotion in each headline. Emotion score are obtained by using Pointwise Mutual Information. [10] used a combination of four classifiers including Naive Bayes, nearest neighbor cosine, decision lists, and latent semantic analysis. Synonym expansion on the emotion label words is also performed, using the *Roget Thesaurus* [17]. Based on the Semeval corpus, [18] conducted comparative evaluations of several knowledge-based and corpus-based methods.

Moreover,Bao et al. [3, 4] proposed a topic model by augmenting latent Dirichlet allocation (LDA) with an intermediate layer for emotion modeling whose experiments were conducted on a dataset collected from the news website Sina. Lin et al. [12, 13] studied the classification of news articles into emotions that they invoke on their readers instead of the authors. On the other hand, people try to do emotion tagging for online news comments in Zhang et al. [22, 23], which uses a fixed combination strategy to merge heterogeneous information sources. In this paper, we use a novel model with multi-source on a real-world dataset to evaluate the effectiveness of the proposed approach.

[2]http://nlp.cs.swarthmore.edu/semeval/tasks/task14/summary.shtml

III. Multi-Source Emotion Tagging

In this section, a novel method named Multi-Source Emotion Tagging for online news is proposed based on combining multiple source’ results. We firstly define our research problem in section III-A. Then we apply basic logistic regression on news article and comment in section III-B and III-C. Finally we propose the approach of two-layers for the scenario combining multi-source prediction results to do a second prediction in section III-D.

A. Problem Definition

Here we set the definition of research problem for online news. We have a collection of comment $\mathcal{C}$ which are made by users after reading online news articles $\mathcal{D}$. All features in news article and comment are from $\mathcal{V}$, an emotion word dictionary. We also predefined the emotion set $\mathcal{E} = \{e_k\}(k = 1, \dots, K)$ from which we select emotion for the news. The i_{th} news article d_i is associated with a user-generated comment combination c_i over each emotion e_k. Each news article d has an emotion rating variable $t = \{t_k\}$ denotes how each emotion the reader expresses aftering reading news. As a result, there exists $\sum_{k=1}^{K} t_k = 1$.

We take the problem of emotion tagging for online news into a multi-class classification that classifies news into different emotion tags. We adopt multi-source including the news article and its’ comment combination to achieve the goal.

B. Emotion Tagging with News Article

Sentiment classification model generally fall into two categories, generative models and discriminative models. The discriminative models have attractive theoretical properties. Thus we adopt logistic regression, one of the discriminative probabilistic models to solve the problem of emotion tagging with news article.

Formally, given the i_{th} news article d_i, we denote the conditional probability that the news will be tagged as a predefined emotion e_k by a logistic regression as ξ_{ik}. The parametric form of ξ_{ik} can be expressed as follows in terms of soft-max function over a linear function of features,

$$\xi_{ik} = P(e_k|d_i) = \frac{\exp(\alpha_k s_i)}{\sum\limits_{k=1}^{K} \exp(\alpha_k s_i)} \tag{1}$$

Here s_i represents the feature vector of news d_i, α_k denotes the combination parameters for each term with emotion e_k, after that we need to estimate the parameter by maximizing the likelihood function as follow,

$$\max_{\alpha} \prod_{d_i \in \mathcal{D}} \prod_{k=1}^{K} {\xi_{ik}}^{t_{ik}} \tag{2}$$

For the sake of simplicity, we usually minimization the negative log-likelihood function with L2 (i.e., ridge) regularization, and we call it the loss function as below,

$$\mathcal{L}(\alpha) = -\sum_{d_i \in \mathcal{D}} \sum_{k=1}^{K} t_{ik} \log \xi_{ik} + \lambda R(\alpha) \tag{3}$$

Here L2 regularization term can be written with the following notation,

$$R(\alpha) = \|\alpha\|_2^2 = \sum_{k=1}^{K} \sum_{v_j \in \mathcal{V}} \alpha_{kj}^2 \tag{4}$$

Finally, the loss function can be expressed as below,

$$\begin{aligned} \mathcal{L}(\alpha) &= -\sum_{d_i \in \mathcal{D}} \sum_{k=1}^{K} t_{ik} (\alpha_k s_i - \log \sum_{r=1}^{K} \exp(\alpha_r s_i)) \\ &+ \lambda \sum_{k=1}^{K} \sum_{v_j \in \mathcal{V}} \alpha_{kj}^2 \end{aligned} \tag{5}$$

C. Emotion Tagging with Comment

Similar as section III-B, we describe theemotion tagging with comment in this section. In the i_{th} news, we combine the top news comment together as c_i. It will be tagged as a predefined emotion e_k by the Equation 6, referred as η_{ik}.

$$\eta_{ik} = P(e_k|c_i) = \frac{\exp(\theta_k x_i)}{\sum\limits_{k=1}^{K} \exp(\theta_k x_i)} \tag{6}$$

The combination parameters θ_k can be estimated by minimizing the following loss function with L2 (i.e., ridge) regularization in the training set.

$$\begin{aligned} \mathcal{L}(\theta) &= -\sum_{d_i \in \mathcal{C}} \sum_{k=1}^{K} t_{ik} \log \eta_{ik} + \lambda R(\theta) \\ &= -\sum_{c_i \in \mathcal{C}} \sum_{k=1}^{K} t_{ik} (\theta_k x_i - \log(\sum_{r=1}^{K} \exp(\theta_r x_i))) \\ &+ \lambda \sum_{k=1}^{K} \sum_{v_j \in \mathcal{V}} \theta_{kj}^2 \end{aligned} \tag{7}$$

Here, x_i represents the feature vector of news comment and η_{ik} denotes the probability that the news comments c_i will be tagged as an emotion e_k.

D. Multi-Source Emotion Tagging

As different information source can bring different insight on emotion tagging, our new proposed method named multi-source emotion tagging utilizing multiple sources' prediction outputs to gain more accurate result. Firstly, we do single source emotion tagging training as section III-C and III-B. Then we combine the two outputs and a constant bias as a new feature vector. Let ψ_i denote the combined news features, i.e. $\psi_i = \{\xi_{i1}, \xi_{i2}, ..., \xi_{iK}, \eta_{i1}, \eta_{i2}, ..., \eta_{iK}, 1\}$,we train ψ_i in the train set making a two-layer training approach. Formally, the probability that emotion e_k will be assigned to the i_{th} news which can be estimated as below,

$$\mu_i = \frac{\exp(\omega_k \psi_i)}{\sum\limits_{k=1}^{K} \exp(\omega_k \psi_i)} \tag{8}$$

where ω_k denote the combination parameters for the multiple sources and the extra bias. These parameters can be determined by maximizing the following loss function.

$$\begin{aligned} \mathcal{L}(\omega) &= \sum_{c_i \in \mathcal{C}} \sum_{k=1}^{K} t_{ik} \log \mu_i \\ &= \sum_{c_i \in \mathcal{C}} \sum_{k=1}^{K} t_{ik} (\omega_k \psi_i - \log(\sum_{r=1}^{K} \exp(\omega_r \psi_i))) \end{aligned} \tag{9}$$

Here, μ_i denotes the probability that the i_{th} news will be tagged as an emotion e_k.

IV. Experiments

In this section, we firstly introduce our datasets. Then we describe experiments on real labeled datasets from the society channel of Sina website. Finally we compare the result on our method and single source baseline.

A. Datasets

Web crawler have downloaded most-viewed news article, its matching comment and its user ratings from Society channel of Sina News, one of the largest news providers in China, starting from 2014/5/1 and ending with 2015/5/1. Since we have not find any user rating in English web service, we decide to use Sina dataset alone. However, our novel model is language independent. Among the i_{th} news in Society channel of Sina News, we set a threshold 0f 20 to the user rating and take its top 20 comment as c_i. Finally, we get 6821 news with its matching comment and ratings. Table I gives detailed statistics of the datasets.

TABLE I. The statistics of labeled news on Society Channel of Sina News

Category	Count	Proportion
Touched	1281	18.78%
Shocked	365	5.35%
Amused	1519	22.27%
Sad	940	13.78%
Surprised	165	2.42%
Angry	2551	37.40%
Total	6821	

Text content like news article and comment are segmented to words by ICTCLAS Zhang et al. [21] and emotion words among them is chosen to be the terms, as these terms are more likely to convey the emotion from users. In our text-based classifier, we utilize the emotion terms frequency(tf) in the content of documents as features.

B. Evaluation Metrics

We adopt two measures including Mean Reciprocal Rank (MRR) and Accuracy ($Accu@n$) with a 10-fold cross validation.

1) Mean Reciprocal Rank (MRR)
Given a text content T, let $rank_i$ denote the position of its truth emotion $\overline{e}_i$ in the predicted emotion ranking list $\mathcal{E}_i$, MRR of the collection will be expressed as,

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \tag{10}$$

2) Accuracy ($Accu@m$)
Given a text content T, its truth emotion $\overline{e}_i$ and predicted emotion set $\mathcal{E}_i@m$ including m top ranked emotions, $accu_i@m$ is defined as,

$$accu_i@m = \begin{cases} 1, & \overline{e}_i \in \mathcal{E}_i@m \\ 0, & \overline{e}_i \notin \mathcal{E}_i@m \end{cases} \tag{11}$$

Thus, $Accu@m$ for the entire collection $\{T_i\}(i = 1, \ldots, N)$ is:

$$Accu@m = \frac{1}{N}\sum_{i=1}^{N} accu_i@m \tag{12}$$

where N is the number of content in the dataset.

C. Experimental Methods

The following four methods are compared:

- **Multi-Source Logistic Regression(MSLR)** This approach is proposed in section III-D and based on a two-layers multi-source logistic regression.
- **News Article Logistic Regression(NLR)** This approach is mentioned in section III-B and based on logistic regression with news article.
- **Comment Logistic Regression(CLR)** This approach is mentioned in section III-C and based on logistic regression with news comment.

D. Experimental Results

Table II present the evaluation results in Accu@n(n=1,2,3) on each method and The †‡ symbols indicate statistical significance with p-value < 0.001 with each model in comparison to NLR and CLR respectively. Through this table, our proposed method achieves the best result in Accu@n(n=1,2,3) and shows the robustness of the proposed discriminative meta classification models by evaluating them with different sizes of datasets.

TABLE II. RESULTS OF THREE MODELS ON SOCIETY CHANNEL OF SINA NEWS.

Methods	Accu@1	Accu@2	Accu@3
MSLR	0.5831 †‡	0.7738†‡	0.8642†‡
NLR	0.5640	0.7518	0.8466
CLR	0.5594	0.7433	0.8406

Experiment also explore the influence of the data size to our proposed method and two baselines. 1/3, 2/3 and 3/3 of dataset have been selected randomly and repeated 100 times to observe the impact of dataset size. The MRR and Accu@1 results of the three models in different sizes dataset are plotted in Figure 2 and 3. This shows that more training data will bring more accurate prediction results to all models. LR based on news content predict better than LR based on comment and our two layer method predict better than all baselines.

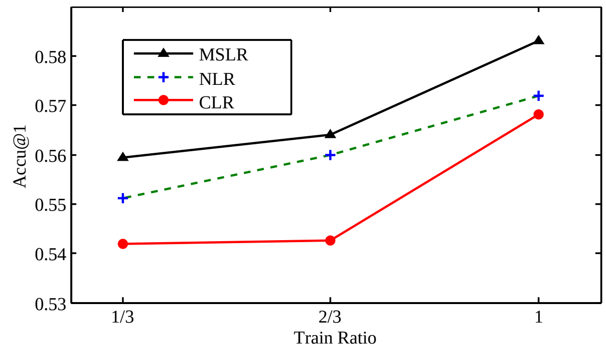

Fig. 2. The $Accu@1$ results on three models in different sizes of Sina dataset

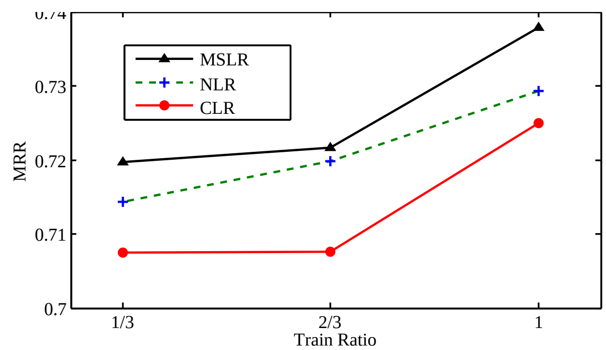

Fig. 3. The MRR results on three models in different sizes of Sina dataset

V. CONCLUSION AND FUTURE WORK

The emergence of online news and social media has promoted the rapid evolution of global conversation. Emotion tagging is a new problem need to be solved. This paper focuses on emotion tagging for online news. We propose a meta classification model by treating emotion tagging for online news as second layer combination problem. Our two-layer method has provided a convenient way to combine multi-source information in one model. Thorough empirical experiments have been conducted on the Sina datasets to show the effectiveness of the proposed probabilistic models for predicting emotions for online news.

This paper is an initial step to the emotion tagging research area. First of all, we plan to extend the work to the collection of online news. In addition, various types of relational information may be available on social media, such as user social network, users tagging news, users voting for comments and so on. The information can be utilized to improve emotion tagging and may lead to better prediction experience on social media. Moreover, it is worthwhile to exploring the emotion dictionary and emotion feature selection.

VI. Acknowledgment

The authors would like to thank the reviewers and the editors for their constructive suggestions. This work is partially supported by National Natural Science Foundation of China under Grant No. 61402243, and Tianjin Municipal Science and Technology Commission under Grant No. 13ZCZDGX02200, 13ZCZDGX01098, and 14JCQNJC00200. This work is also partially supported by the Fundamental Research Funds for the Central Universities of China.

References

[1] C.O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics, 2005.

[2] Saima Aman and Stan Szpakowicz. Using roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 296–302, 2008.

[3] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. Joint emotion-topic modeling for social affective text mining. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 699–704. Ieee, 2009.

[4] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. Mining social emotions from affective text. *IEEE Transactions on Knowledge and Data Engineering*, (99): 1–1, 2011.

[5] Haji Binali and Vidyasagar Potdar. Emotion detection state of the art. In *Proceedings of the CUBE International Information Technology Conference*, pages 501–507. ACM, 2012.

[6] François-Régis Chaumartin. Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 422–425. Association for Computational Linguistics, 2007.

[7] D. Das and S. Bandyopadhyay. Word to sentence level emotion tagging for bengali blogs. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 149–152. Association for Computational Linguistics, 2009.

[8] D. Das and S. Bandyopadhyay. Sentence level emotion tagging on blog and news corpora. *Journal of Intelligent System*, 19(2), 2010.

[9] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4): 169–200, 1992.

[10] Phil Katz, Matthew Singleton, and Richard Wicentowski. Swat-mp: the semeval-2007 systems for task 5 and task 14. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 308–313. Association for Computational Linguistics, 2007.

[11] Zornitsa Kozareva, Borja Navarro, Sonia Vázquez, and Andrés Montoyo. Ua-zbsa: a headline emotion classification through web information. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 334–337. Association for Computational Linguistics, 2007.

[12] K.H.Y. Lin, C. Yang, and H.H. Chen. What emotions do news articles trigger in their readers? In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 733–734. ACM, 2007.

[13] K.H.Y. Lin, C. Yang, and H.H. Chen. Emotion classification of online news articles from the reader's perspective. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 1, pages 220–226. IEEE, 2008.

[14] Hugo Liu, Henry Lieberman, and Ted Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. ACM, 2003.

[15] Gilad Mishne and Maarten de Rijke. Capturing global mood levels using blog posts. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 145–152, 2006.

[16] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Narrowing the social gap among people involved in global dialog: Automatic emotion detection in blog posts. In *ICWSM*. Citeseer, 2007.

[17] Peter Mark Roget. Rogets thesaurus of english words and phrases. *Harlow, Essex, England: Longman Group Limited*, 1852.

[18] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[19] C. Yang, K.H.Y. Lin, and H.H. Chen. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 275–278. Ieee, 2007.

[20] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136. Association for Computational Linguistics, 2007.

[21] H.P. Zhang, H.K. Yu, D.Y. Xiong, and Q. Liu. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 184–187. Association for Computational Linguistics, 2003.

[22] Y. Zhang, Y. Fang, X. Quan, L. Dai, L. Si, and X. Yuan. Emotion tagging for comments of online news by meta classification with heterogeneous information sources. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1059–1060. ACM, 2012.

[23] Ying Zhang, Ning Zhang, Luo Si, Yanshan Lu, Qifan Wang, and Xiaojie Yuan. Cross-domain and cross-category emotion tagging for comments of online news. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 627–636, 2014. doi: 10.1145/2600428.2609587.