Learning Salient Samples and Distributed Representations for Topic-Based Chinese Message Polarity Classification

Xin Kang[1,2]* **Yunong Wu**[3]* **Zhifei Zhang**[4]*
[1]Department of Electronics and Information, Tongji University
[2]Faculty of Engineering, Tokushima University
[3]Department of Research and Development, Business Big Data Co., Ltd.
[4]Department of Computer Science and Operations Research, University of Montreal
[1]`xkang@tongji.edu.cn`, [2]`kang-xin@tokushima-u.ac.jp`
[3]`wuyunong@brandbigdata.com`, [4]`zhanzhif@iro.umontreal.ca`

Abstract

We describe our participation in the Topic-Based Chinese Message Polarity Classification Task, based on the restricted and unrestricted resources respectively. In the restricted resource based classification, we focus on the selection of parameters in a multi-class classification model with highly-biased training data. In the unrestricted resource based classification, we explore the distributed representation of Chinese words through unsupervised feature learning and the annotation of salient samples through active learning, with a raw corpus of over 90 million messages extracted from Chinese Weibo Platform. For two classification subtasks, our submitted results ranked the 4th and the 2nd respectively.

1 Introduction

The ZWK team participates in the Topic-Based Chinese Message Polarity Classification Task, the purpose of which is to predict the message polarities in the Positive, Negative, and Neutral classes towards particular topics. Learning classification models on the training corpus with bag-of-words features is very challenging, given the fact that the class labels are highly-biased in the corpus and that the number of training samples is an order of magnitude lower than the number of observed word features. Therefore, our work focuses on the active learning and unsupervised feature learning algorithms, to avoid over-fitting the parameters of a linear classification model. To predict polarities with respect to specific topics, we re-evaluate the features with respect to their distances to topical words in a message.

Because the class labels are highly-biased in the training corpus, most of which are Neutral, we explore an active learning algorithm to incrementally obtain the knowledge of different polarities from a large raw corpus. In the iterative procedure of active learning, salient samples are firstly selected from a large raw corpus, based on the amount of information in their polarity predictions, their representativeness within the raw corpus, and their distinctiveness in the selection. The selection procedure ensures that samples of the minor classes are more probably selected than samples of the major class(es) and that the extension of training data with these samples has the most potential to improve the current classification model. Then, class labels are annotated to the salient samples by querying oracles, and all labeled samples are appended to the training corpus to update the classification model before the next iteration in active learning. We select and append the salient samples in a batch-mode, to efficiently re-balance the training corpus and incrementally improve the polarity classification model.

And because the number of training messages (around 5K) turns much smaller than the number of unique words (17.5K), a linear classification model can be easily over-fitted with bag-of-word features. To avoid over-fitting, we project the 17.5K-dimensional word space to a 200-dimensional vector space through an unsupervised feature learning. We employ word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov et al., 2013c) as the unsupervised feature learning algorithm, based on a raw corpus of over 90 million messages extracted from Chinese Weibo Platform. One of the most significant advantage of learning with word2vec is that the vector representations are additively composable, which means we can represent the semantic composition of multiple words by adding the respective vector representations. For the topic-based polarity classifica-

*These authors contributed equally to this work.

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 68–73,
Beijing, China, July 30-31, 2015. ©2015 Association for Computational Linguistics and Asian Federation of Natural Language Processing

tion problem, we only compose words around the specific topics as features, with an exponentially decreasing weight along the word sequence.

The rest of this paper is arranged as follows: section 2 reviews the related work of polarity classification, section 3 describes our active learning algorithm for retrieving salient samples, section 4 illustrates the unsupervised learning algorithm for reducing feature dimensions, section 5 shows our experiment results on polarity classification and discusses the over-fitting problem, and section 6 concludes our work.

2 Related Work

Polarity classification has been a popular field in natural language processing. In polarity classification, the main difficulty is to find effective language features for distinguishing positive, negative, and neutral sentiments (Kiritchenko et al., 2014). Because overwhelming ambiguities exist in word polarity expressions, polarity prediction results based on lexicons (Taboada et al., 2011) could be unreliable.

To incorporate such ambiguity in sentiment modeling, a few studies resort to the hierarchical Bayesian models, in which the ambiguity of sentiments in words has been transformed into the joint probability of words, word clusters (topics), and sentiments (Ren and Kang, 2013; Wu et al., 2014; Rao et al., 2014). Another solution of resolving such ambiguity in sentiment classification is to directly represent words in sentimental vectors (Maas et al., 2011; Socher et al., 2013; Tang et al., 2014; Kalchbrenner et al., 2014; Kim, 2014). Compared to the Bayesian models, vectorized representation relates the semantic information directly to each entry of the word vector, and the results can be easily transformed in a simple classifier. We employ an unsupervised algorithm word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov et al., 2013c) to learn such vectorized word representation from a large raw corpus.

In most of these sentiment classification researches, the class labels are not as skewed as the polarity labels in this task. To rebalance the training data, we develop an novel active learning (Settles, 2010) algorithm which automatically selects the salient samples from a large raw corpus and costs the minimum labor for annotation. TwitterHawk (Boag et al., 2015) notably places 1st in topic-based sentiment classification subtask of the SemEval-2015 shared task on Sentiment Analysis in Twitter, which uses many hand-crafted features and a classic classifier. The overall solution of our work is basically consist with it.

3 Active Learning of Salient Samples

Active learning (Settles, 2010) is a subfield of machine learning. An active learning algorithm will automatically select salient samples from the unlabeled data set, and will incrementally improve machine learning by obtaining knowledge from these samples and merge them to the training data.

In the polarity classification problem, we developed an active learning algorithm for obtaining the knowledge of different polarities from a large raw corpus of over 90 million messages. The algorithm begins with a restricted corpus L, in which the polarity labels are highly-biased i.e., 394 positive labels, 538 negative labels, and 3,973 neutral labels. By iterating through three sample selection steps, the algorithm incrementally adds salient messages to L after querying labels from oracles, and generates a less-biased corpus finally with 1,003 positive labels, 1,060 negative labels, and 4,242 neutral labels.

Before the first step of sample selection, a multi-label Logistic Regression classifier is trained on L, and a batch of 1,000,000 messages is extracted from the raw corpus as an unlabeled pool U. We get probabilistic prediction y for each message x in U, and evaluate the amount of information in its probabilistic prediction by entropy

$$E(x) = -\sum_{y} p(y|x) \log p(y|x). \qquad (1)$$

The largest entropy $E(x)$ is approached by those x with the most evenly distributed predictions in y. Because the classifier is trained on a biased corpus, its prediction would favor the major label of neutral. Therefore, the true labels for messages in U with larger entropies are more probably positive and negative than neutral, since a truly neutral x will get odd probabilistic predictions and locates far from the large entropies. Our algorithm selects the top 10,000 messages for S_1 as the first step.

For the second step, the algorithm calculates Euclidean distances between every pair of messages x_i and x_j in S_1 by

$$d(x_i, x_j) = \sqrt{x_i \cdot x_i - 2x_i \cdot x_j + x_j \cdot x_j}, \qquad (2)$$

where $\cdot$ is the dot product of two message vectors. We evaluate the representativeness of a message x

by its average distance between all other messages in S_1 as

$$R(x) = \frac{1}{|S_1| - 1} \sum_{x_i \in S_1} d(x, x_i), \tag{3}$$

and select the smallest 1,000 messages for S_2. This is because a representative x must be surrounded by many similar x_i's in the Euclidean space, and $R(x)$ is usually smaller than $R(x')$ for x' in a very sparse region[1]. We select the representative messages for S_2 because they are potentially more general samples.

In the third step, the algorithm iteratively select the most distinctive messages from S_2 and move them to an empty set S_3. For x in S_2, its distinctiveness is evaluated by the minimum Euclidean distance between x and every x_i in $L \cup S_3$

$$D(x) = \min_{x_i \in L \cup S_3} d(x, x_i). \tag{4}$$

Then the message with the largest distinctiveness

$$x^* = \arg\max_{x \in S_2} D(x) \tag{5}$$

is moved from S_2 to S_3. This procedure selects 100 most distinctive x^* for S_3, by ensuring the diversity in selected samples. The active learning algorithm then queries oracles (i.e., human experts) for polarity labels in S_3, and merges the labeled S_3 to L at the end of this step.

4 Unsupervised Learning of Word Features

The bag-of-word feature is simple for usage in learning a polarity classification model. However, the feature dimension is an order of magnitude higher than the size of training data. In such case, the trained classifier is only sensible to messages in the training corpus, but not generalizable to new messages, which is an over-fitting problem. And a further problem in bag-of-word feature is that the semantic information in words is not fully represented by single feature indexes.

We employ a dimension reduction method to solve this problem, which projects the large word space to a small vector space through unsupervised learning of distributed word representations. The algorithm for unsupervised learning is word2vec[2], and we employ its python implementation[3] to learn a 200-dimensional vector representation for words with a 90-million-message corpus. The algorithm learns word representations by constructing a recurrent neural network with each word and its context associated with as a layer (vector) of neurons respectively and fitting a 3-layer neural network to recurrently predict the next word given the current word and its context. More detailed implementations are described in (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov et al., 2013c). The algorithm learned vector representations for 1 million words.

An important property of the word2vec algorithm is that both the vector representation and the addition (subtraction) on vector representations are semantically meaningful. This can be examined by the word pair relationships (Mikolov et al., 2013a) as follows. We calculate the semantic relation between words w_1("China") and w_2("Beijing") by subtracting their vector representations, and use this to examine if a same relationship exists between w_3("American") and w_4("Washington, D.C."), by searching through the learned words in V

$$w^* = \arg\max_{w \in V} \cos(w_1 - w_2 + w_4, w). \tag{6}$$

w^* equals w_3 with the cosine similarity 0.6433, which ensures the additive compositionality exists in our learned model.

We assume words around the topical word have greater impact to the message polarity than the distant words. To compose the semantic information in feature vector x for polarity prediction, we attach exponentially decreasing weights around the topical word

$$x = \sum_{i \neq t} \exp(-|i - t|/l) w_i, \tag{7}$$

where i and t are the word and topic locations in a message. l controls the decreasing speed in weights, which is set to 5 in our work.

5 Experiment and Discussion

The Topic-Based Chinese Message Polarity Classification task provides 4,095 topic-message pairs from Chinese Weibo Platform for developing a basic polarity classifier over positive, negative,

[1] This is not always true in Euclidean space, but has been employed in many active learning algorithms.

[2] https://code.google.com/p/word2vec/

[3] https://radimrehurek.com/gensim/models/word2vec.html

		RUN0			RUN1			RUN2			RUN3	
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Neg	30.47	18.52	23.04	40.65	**24.54**	**30.60**	**43.07**	16.74	24.10	40.72	5.25	9.30
Neu	77.25	88.43	82.46	**78.98**	87.08	82.83	78.01	91.98	84.42	76.08	**97.88**	**85.61**
Pos	23.35	9.20	13.20	19.08	**18.06**	18.55	**24.73**	16.06	**19.47**	19.93	2.00	3.63
Mac	43.69	38.72	39.57	46.24	**43.22**	**44.00**	**48.60**	41.49	42.67	45.54	35.04	32.85
Acc		70.68			71.30			73.42			**74.89**	

Table 1: Polarity classification results.

and neutral sentiments, and 19,469 topic-message pairs for evaluating the classification results. In this task, further resources are required to improve the classifier.

We employ a raw corpus of 90 million messages from Chinese Weibo Platform, for developing salient samples with active learning and for learning distributed representation of words with unsupervised feature learning. All these messages are randomly collected from April to September in 2013.

Based on the One-vs-All Logistic Regression algorithm from scikit-learn[4], we construct several polarity classifiers $\mathrm{clf_i}$ with different features. For the basic classifier clf_0, we explore the bag-of-word feature by collecting words which occur more than "min_occur" times in the training corpus and by removing the most frequent "stop_num" words in the collection. We select model parameters "C", "penalty", "class_weight" and feature parameters "min_occur", "stop_num" through grid search with 5-fold cross validation on the training corpus.

We employ an active learning algorithm to generate a less-biased training corpus as shown in Figure 2. Class labels have been significantly balanced after 14 loops of sample selection. Classifier clf_1 is trained on this corpus, with a similar parameter selection procedure as clf_0.

We employ the word2vec algorithm to project the large word space to a small vector space. The algorithm has learned a 200-dimensional distributed representation for 1 million different words in the raw corpus. Classifier clf_2 is trained on the basic corpus with composed word2vec features as in Eq. 7.

To evaluate the classification results, we calculate precision, recall, and F1 scores for each polarity, the macro average of these scores, and the overall accuracy. Table 1 shows the evaluation of

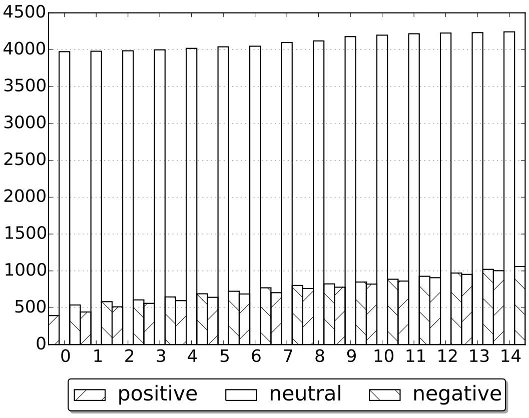

Figure 1: Label counts in active learning.

4 RUNs on the test corpus, with each RUN described below.

- RUN0 generates predictions from clf_0.
- RUN1 generates predictions from clf_1.
- RUN2 summarizes probabilistic predictions by

$$Y = \arg\max \sum_{i \in \{0,1,2\}} \mathrm{pclf_i}(X)$$

where $\mathrm{pclf_i}$ generates the probabilistic predictions over (negative, neutral, positive) for $\mathrm{clf_i}$, and $\arg\max$ generates the class label with the largest accumulated probabilistic prediction.

- RUN3 combines probabilistic predictions by

$$Y = \mathrm{clf}_3\left([\mathrm{pclf}_0(X); \mathrm{pclf}_1(X); \mathrm{pclf}_2(X)]\right)$$

where clf_3 takes three probabilistic predictions as features and generates polarity predictions in Y. clf_3 has been trained on the labeled corpus, with parameters optimized over classification accuracy.

[4] http://scikit-learn.org/dev/index.html

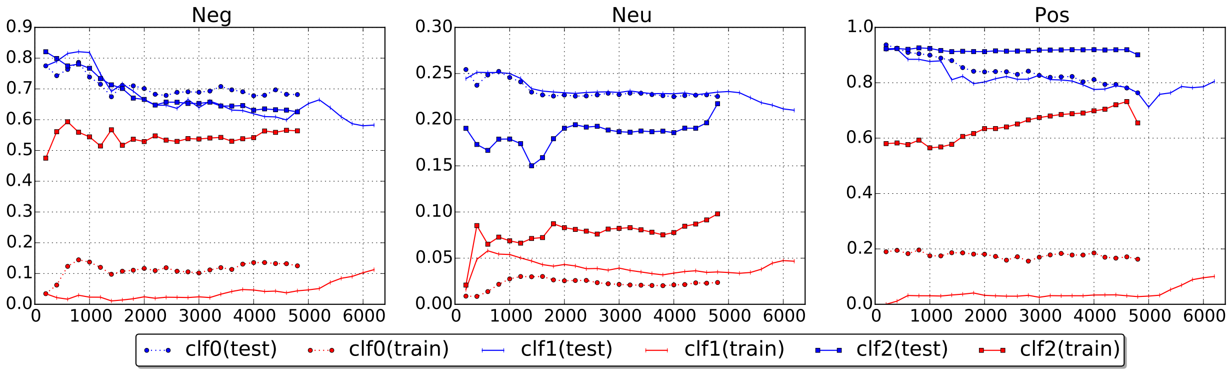

Figure 2: Learning curves.

RUN3 achieves the highest accuracy since its classifier is optimized over classification accuracy on training data. RUN1 yields the highest macro recall and F1 scores, which suggests that our active learning has effectively selected salient samples for training the polarity classifier. RUN2 yields the highest macro precision by summarizing the probabilistic predictions from three classifiers. Among the results from all participants for the restricted and unrestricted source based classifications, our submitted results in RUN0 and RUN3 have been ranked the 4th and the 2nd, respectively.

To further examine the problems in learning procedure we plot learning curves for each class label. A learning curve represents the error rates of a classifier, trained with different sizes of data. Learning curves of clf_0 and clf_1 suggest an over-fitting problem since the models fit well on the training data but generalize poorly on the test data. Compard to clf_0, clf_1 is more generalizable with extra samples selected by active learning. The learning curves of clf_2 on negative and positive labels suggest an under-fitting problem, which implies that the composed word2vec features have lost some important information for predicting these labels. Improvement is possible to be achieved by increasing the dimension of word vectors in the word2vec algorithm.

6 Conclusion

We report our approach for solving the Topic-Based Chinese Message Polarity Classification problem. The basic polarity classifier is over-fitted with highly-biased labels in the training data. We employ an active learning algorithm to select salient samples from a large raw corpus, and improve the learning procedure with less-biased labels in a larger training data. We then resort to a dimension reduction technique, by reducing the feature dimension from 17.5K to 200 with the word2vec algorithm, to further relief the over-fitting problem. However, because the feature reduction loses some important information, the model suffers an under-fitting problem. We believe developing the topic-based features in a properly low dimension and incrementally selecting salient samples would help improving the classification results. Moreover, we want to analyze the function of sentence syntactics for topic-based polarity classification in the future, since the syntactic structures can better interpret the significance of a feature relevant to a specified topic. Last but not least, we hope to further improve the classification algorithm based on the distributed representations of words as features.

Acknowledgments

This work has been partially supported by the China Postdoctoral Science Foundation funded project, under Grant No. 2014M560351, and the Quebec-China Postdoctoral Scholarship (File No. 188040).

References

William Boag, Peter Potash, and Anna Rumshisky. 2015. TwitterHawk: A feature bucket based approach to sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 640–646. ACL.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665. ACL.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751. ACL.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. ACL.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751. ACL.

Yanghui Rao, Qing Li, Xudong Mao, and Wenyin Liu. 2014. Sentiment topic models for social emotion mining. *Information Sciences*, 266:90–100.

Fuji Ren and Xin Kang. 2013. Employing hierarchical Bayesian networks in simple and complex emotion topic analysis. *Computer Speech & Language*, 27(4):943–968.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. ACL.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565. ACL.

Yunong Wu, Kenji Kita, and Kazuyuki Matsumoto. 2014. Three predictions are better than one: Sentence multi-emotion analysis from different perspectives. *IEEJ Transactions on Electrical and Electronic Engineering*, 9(6):642–649.