Multi-label Sentiment Analysis Base on BERT with modified TF-IDF

1st Zeyi Jin
School of Logistics Engineering
Wuhan University of Technology
Wuhan, China
elevenjzy@126.com

2nd Xin Lai
School of Logistics Engineering
Wuhan University of Technology
Wuhan, China
fairylai@126.com

3rd Jingjig Cao
School of Logistics Engineering
Wuhan University of Technology
Wuhan, China
bettycao@whut.edu.cn

2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN) | 978-1-6654-1461-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISPCE-CN51288.2020.9321861

***Abstract*—With the evolution of web technologies, sentiment analysis, especially aspect-based sentiment analysis (ABSA), as a technology to identify and extract user opinions, has attracted widespread attention. According to the characteristics of user comments, we convert ABSA problem into a multi-label classification problem, and propose a classification model based on BERT with modified TF-IDF for ABSA. In the feature extraction phase, a modified TF-IDF method is proposed to better reflect the words' importance in multi-label classification problem by calculating the different weights in individual classes. And a new feature is generated by combining the BERT embedding with modified TF-IDF for classification task. Then the feature is input into a fully connected layer to fine-tune BERT model for the multi-label classification task. The effectiveness of supervised TF-IDF and proposed model is validated by experiments of multi-label classification on a restaurant reviews dataset.**

***Index Terms*—sentiment analysis, BERT, TF-IDF, multi-label classification**

I. Introduction

With the growth of online shopping, various e-commerce websites have emerged, Taobao.com, Amazon.com, etc., all of which provide the platforms for customers to express their usage experiences on products, resulting online reviews have grown exponentially. Huge volumes of reviews data on the e-commerce websites are of great value for customers, business owners, and e-commerce websites. Analyzing customer reviews not only can guide consumers to make purchasing decisions, but also can obtain product feedback and customer preferences, which provides strong support in precision marketing and product design. Online reviews have been regarded as one of the most effective sources in facilitating potential consumers online purchase decisions about products or services [1]–[5].

However, due to the large number of online reviews from consumers, it is impractical to manually deal with the opinions in these online reviews. Sentiment analysis, a branch of affective computing research, is an significant approach to opinion mining that efficiently analyzes such reviews [3]. To better understand reviews and well extract hidden feature, aspect-based sentiment analysis (ABSA) is proposed [6], which aims to identify fine-grained polarity towards a specific aspect [7].

Approaches for ABSA mainly include rule-based method such as lexicon-based method [8]–[10], and machine learning method which includes traditional machine learning methods [4], [7], [11] and deep learning methods [12]–[15]. Compared with traditional machine learning methods, deep learning can automatically extract features by utilizing low-dimensional and dense vectors to implicitly represent language semantic features. Nevertheless, on the ground that the datasets in the majority of ABSA tasks is of small scale, the performance of DNN are prone to overfitting on this case, resulting in unsatisfactory results in practice [16]. Over the last a decade or so, the pre-trained language models, such as ELMo [17], OpenAI GPT [18], and BERT [19], have shown outstanding performance on ABSA problem [2] [20]. The pre-trained language models can learn universal language representations by pre-training on the large unlabeled corpus to avoid overfitting on small size data.

In the light that a review includes the sentiments toward one or more aspects, we treat ABSA as a multi-label classification problem by label encoding in this paper. For instance, "the food quality is decent, but the price is very steep", a comment contains two aspect-sentiment pairs: "food, positive" and "price, negative". For the purpose of exploiting more, we try to integrate the features learned by pre-training model and manually extracted by traditional machine learning methods. In this paper, a multi-label classification model based on BERT with modified TF-IDF for ABSA is proposed. The model mainly includes two steps: (1) feature extraction based on BERT with modified TF-IDF; (2) fine-tune model for the multi-label classification task.

The main contributions of this paper include:

- We propose a multi-label classification framework for the problem of aspect-based sentiment analysis in restaurant customer review, which mainly includes text preprocessing, feature extraction utilizing BERT and modified TF-IDF, and fine-tuning.
- In the multi-label classification model proposed in this paper, a new feature extraction way is discussed. We try to integrate the features learned by pre-training model and manually extracted by traditional machine learning methods. For the purpose of considering labels information, a modified TF-IDF method in a supervised method manner for multi-label classification is proposed.
- The effect of feature extraction based on modified TF-

IDF, BERT and fusion of the two methods are compared in experiment. And experiment result proves that the multi-label classification based on BERT and modified TF-IDF model is the most effective.

The remainder of this paper is structured as follows. Section II describes the background information on pre-trained language models and aspect based sentiment analysis. The proposed method is then presented in Section III. In Section IV, we conduct experiments on a real-world restaurant reviews dataset and present the results. Finally, Section V concludes the overall study.

II. Related Work

A. Pre-trained Language Models

With the advent of pre-trained language models, such as ELMo [17], OpenAI GPT [18], and BERT [19], natural language processing (NLP) has entered a new era. Due to the transfer learning capabilities enabled by pre-trained language models, people can use another pre-trained language model as the basis and only fine-tune it to solve the specific NLP task instead of training the model from scratch. ELMo is a deep contextualised word representation that models syntax and semantic of words as well as their linguistic contexts. The model has been pre-trained on a huge text-corpus and learned functions from deep bi-directional models (biLM) [17]. The ELMo model combines left-to-right information of a word, but is only the realization of splicing. The vector splicing method is weak in fusion of contextual features. BERT, short for Bidirectional Encoder Representations, considers the context from both sides of a word based on bidirectional transformer, which can gain a much better understanding of the context in which the word was used. Additionally, BERT is designed to do multi-task learning, that is, it can perform different NLP tasks simultaneously [19]. In this paper, we adopt the BERT model and extend the BERT model by integrating traditional supervised TF-IDF features.

B. Aspect Based Sentiment Analysis

Sentiment analysis refers to the process of extracting the sentiment polarity of opinions expressed in text data [2]. To have a more granular understanding of opinions, the aspect-based sentiment analysis (ABSA) was proposed. The goal of ABSA is to identify the fine polarity of a specific aspect. This task allows users to evaluate every aspect of a given product or service [20], [21].

Traditional approaches for ABSA mainly contains two types, lexicon based and machine learning based [22]. In [10], Zubair et al. proposed a lexicon-enhanced method based on rule-based classification scheme by integrating emojis, modifiers, and domain-specific terms to analyze comments posted by online communities. In [7], Manek et al. proposed a feature selection method with Support Vector Machine based on Gini Index for sentiment classification for large data set. However, traditional approaches focus on designing features, and it almost reaches its performance bottleneck [21].

The deep learning can automatically learn features from big data instead of manually designed features. Hence, lots of deep learning method applied on ABSA, such as convolutional neural network [21], recurrent neural network [13].

III. Methodology

A. Problem Statement

In this paper, we focus on the aspect-based sentiment analysis problem. This task can be formulated as a multi-label classification problem by label encoding. We regard an aspect-sentiment pair as a label. For instance, "the food quality is decent, but the price is very steep", it contains two aspect-sentiment pairs: (food, positive) and (price, negative). Namely this sentence has two labels, one is (food, positive) another is (price, negative). A fixed aspect set is noted $A = \{a_1, \ldots, a_m\}$, and the sentiment polarity set is noted $P = \{p_1, \ldots, p_n\}$, where m is the number of aspects, n is the number of sentiment polarities. Aspect-sentiment pair $y_{a_i p_j}$ is defined by $\{(a_i, p_j) \mid a_i \in A, p_j \in P, 1 \leq i \leq m, 1 \leq j \leq n\}$.

Let $\mathcal{X} = \mathcal{R}^d$ denote the domain of instances and $\mathcal{Y} = \{y_{a_i p_j}\}$ denote the label space consisting of k class labels, $k = m * n$. Given the training corpus $\mathcal{D}_{tr} = \{(x_i, Y_i) \mid x_i \in \mathcal{X}, Y_i \subseteq \mathcal{Y}\}$, the task is to predict the corresponding label set for test instance x.

B. Framework Overview

In this section, we elaborate the framework of the proposed approach. The overall framework is shown in Fig. 1. The multi-label classification model based on BERT with modified TF-IDF for ABSA proposed in this paper mainly contains three parts, text preprocessing, feature extraction by BERT and modified TF-IDF, and fine-tune model for multi-label classification task. Customer comments on the Internet are inherently noisy, preprocessing the text before feature extraction helps to generate better features and semantics. We perform the following preprocessing steps, including word segmentation, stem extraction, and removal of stop words. After preprocessing, feature extraction which is the difficulty of natural language processing is carried out, and the quality of the feature extraction determines the quality of the classification effect. For a multi-label classification task, labels include important information but word embedding by BERT model, a unsupervised method, does not contain the information of labels. Thus, we combine BERT with a modified TF-IDF which is supervised in the feature extraction phase. Obviously, feature extraction includes two aspects, one is sentences embedding based on BERT pre-training model, and the other is feature generation by modified TF-IDF. Fusion of two aspects of the features as a new feature is entered into a fully connected layer for multi-label classification.

C. Text Preprocessing

Data preprocessing is the foundation of data mining, an indispensable step, because the quality of data will affect the final data mining effect. Effective data preprocessing before mining can save a lot of time and space. Text preprocessing

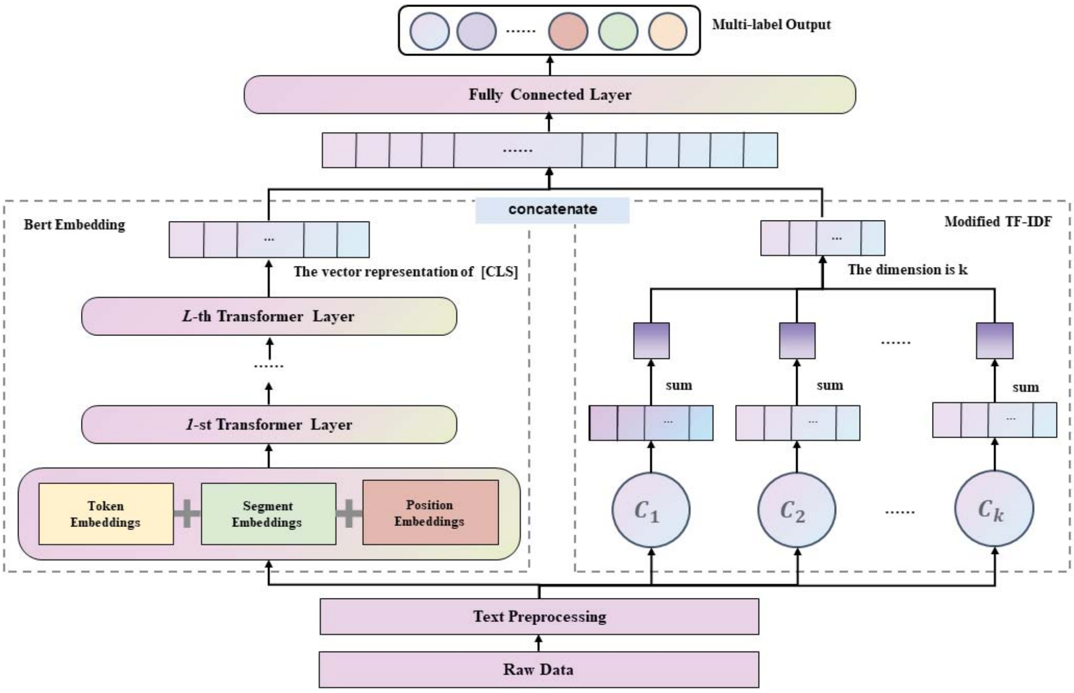

Fig. 1. The framework of the multi-label classification model based on BERT with supervised TF-IDF.

at first is helpful to feature extraction and classification task. In particular we perform the following preprocessing steps including the steps of word segmentation, punctuation removal, stop word removal, and word form restoration:

- Word segmentation. Because the smallest unit of text is words, we divide the text into the smallest unit to be processed. The data set used in this paper is in English, so we segment sentence by spaces.
- Remove punctuation. Punctuation play a major role in sentence segmentation and separation. But there is no substantive effect, so we remove the punctuation in the sentence during data preprocessing.
- Remove stop words. There are many words that have no practical meaning in English expressions, such as articles, conjunctions, prepositions and so on. These are all called stop words. Stop words are abundant in the text, but they are useless for sentiment analysis. On the contrary, it will increase the feature dimension, and confuse the classification, so stop words need to be removed.
- Word form reduction. Word form reduction is to restore the vocabulary to its original form, and the restored words can express the complete semantics. Similar to word form reduction is stemming, but the final result of stemming is stemming, which may not be a complete word. Therefore, word form reduction is more valuable for research and application, so in this article, we choose word form reduction.

D. BERT Embedding

BERT_{base} model is adopt in this paper, and BERT model [19] is one of the most popular pre-trained language model armed with Transformer, which uses bidirectional transformers to pre-train a large corpus, and fine-tunes the pre-trained model on other tasks. We obtain sentence embedding by BERT pre-trained model at first, then fine-tune the pre-trained BERT model for multi-label classification task, described in Section III.F.

First of all, input embedding is constructed by combination of the token embedding, position embedding and segment embedding. Then L transformer layers are stacked to refine the features. BERT model adapted in this paper contains an encoder with 12 transformer layers, i.e. L is 12, and the hidden size of 768. The Output is the final hidden vector of the special [CLS] token which contains all the information of the whole sentence. The embedding dimension is 768, and we get a 768 dimensional vector as the representation of input after BERT pre-training.

E. Modified TF-IDF

TF-IDF is one of the most commonly used weighting metrics for measuring the relationship of words to documents. It is widely used for word feature extraction. But TF-IDF is a unsupervised method which does not utilize the known class information of training dataset while weighting a term, so the computed weight cannot fully reflect the term's importance for classification task. For example, the weight of topic term "food" is same for the text associate with "food" or "price", namely it does not have the class distinguish ability. Apparently it is not reasonable, so we proposed a modified TF-IDF in a supervised method manner for multi-label classification.

In the training phase, training data $\mathcal{D}_{tr}$ is group by labels divided into k subsets, each subset is a corpus denoted by C_j, $1 \leq j \leq K$. Similar to traditional TF-IDF, $tf(w_i, C_j)$ describes the frequency of word w_i in corpus C_j, because the key words for a class often appear frequently in corresponding corpus and they should be assigned greater weights. To eliminate the influence of common words, $idf(w_i, C_j)$ is denoted. For a word w_i the less other corpora containing it are, the higher its representativeness of the corpus C_j is.

For each corpus C_j, the calculation formula of modified TF-IDF is as follows.

$$tf(w_i, C_j) = \frac{n_{w_i,C_j}}{\sum_q n_{w_q,C_j}} \tag{1}$$

where n_{w_i,C_j} is the number of word w_i occurrences in the corpus C_j, $\sum_q n_{w_q,C_j}$ is the number of all words in corpus C_j.

$$idf(w_i, C_j) = \log \frac{\mid k \mid}{\mid \{j : w_i \in C_j\} \mid +1} \tag{2}$$

where $\mid k \mid$ is the total number of corpora, $\mid \{j : w_i \in C_j\} \mid$ represents the number of corpora containing the word w_i. Then we can calculate the TF-IDF weight of word w_i in corpus C_j as in

$$W(w_i, C_j) = tf(w_i, C_j) \times idf(w_i, C_j) \tag{3}$$

In the modified TF-IDF, each word that appears in the training dataset corresponds to k TF-IDF values, and one value corresponds to the corpus associated with one label. For a sentence s, we sum the TF-IDF values of all words for each corpus as the feature of this sentence, and the dimension is k. In the testing phase, if the word does't appear in training dataset, its TF-IDF is defined 0.

F. Fine-Tuning

We add a fully connected layer with sigmoid activation function to finetune the pre-trained BERT model for multi-label classification task. As show in Fig. 1, we concatenate the two features obtained by BERT pre-training model and supervised TF-IDF as the input. Further, the loss function used is a combination of binary cross entropy, because loss function is designed to evaluate all the probability of categories individually rather than as compared to other categories. Hence the use of binary cross entropy rather than cross entropy when defining loss. The model's output are the probabilities of all classes, we hence set a threshold value to get a multi-label output for the input sentence.

IV. Experiment

A. Dataset

This dataset collects online customer reviews of restaurants. The reviews in the data set are within a week of 2006, with reviews from over 50,000 restaurants by different customers in the New York area [23]. Among them, the data set includes 5531 restaurants, written by 32284 different users, with a total of 52264 reviews. The data set also includes other relevant structured information, restaurant id, customer id, restaurant type, customer rating, etc.

We randomly sampled 10,000 sentences, including 2612 restaurants. In this experiment, sentences are used as the basic unit of sentiment analysis, and each sentence is used as a sample. In [23], author calculate the kappa coefficient (K) which measures pairwise agreement among a set of annotators making category judgments. Due to the Kappa value of ambiguous Anecdotes category was moderate (0.51), we don't consider this label in this paper. Thus, the aspect set A is $\{$Food, Price, Ambience, Staff, Miscellaneous$\}$, and m is 5; the sentiment polarity set P is $\{$Positive, Negative, Neutral, Conflict$\}$, and n is 4; the number of labels k is 20, labels are encoded using a one-hot encoding scheme.

B. Experiment Setup

In experiments, the ratio of training dataset to test is 7:3. To verify the validity of supervised TF-IDF, we choose five multi-label classification methods as classifier(BR, CC, LP, MLkNN, BRkNN-a), and compare the effects of two word embedding methods, supervised TF-IDF and traditional TF-IDF. For multi-label classification, the base classifier of BR, CC, LP is SVM; parameter K for MLkNN and BRkNN-a is 20. For BERT model, hyper-parameters' setup is following, the number of epochs is 27; the batch size is 32; the maximum length allows the maximum number of words in one sentence to be 128; the threshold value the last fully connected layer is 0.28; the learning rate is $2 \times 10^{-}5$, used in the optimization algorithm for updating the parameters in both models.

C. Evaluation Metrics

Four performance measures for multi-label classification are employed in experiments. Include example based metrics, subset accuracy, the hamming loss, and two label based metrics, the macro-average F1 score, the micro-average F1 score. These measures can be expressed as follows:

$$Accuracy = \frac{1}{p} \sum_{i=1}^{p} [\![h(x_i) = Y_i]\!] \tag{4}$$

where p is the number of examples, $h(\cdot)$ is a multi-label classification. The subset accuracy evaluates the fraction of correctly classified examples.

$$HammingLoss = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{k} \, |h(x_i) \Delta Y_i| \tag{5}$$

where Δ stands for the symmetric difference between two sets. The hamming loss is defined as the fraction of incorrectly classified labels.

Label based metrics are based on confusion matrix characterizing the binary classification performance. For the j-th label y_j:

$$TP_j = |\{x_i \mid y_j \in Y_i, \wedge y_j \in h(x_i), 1 \leq i \leq p\}|$$

$$FP_j = |\{x_i \mid y_j \notin Y_i, \wedge y_j \in h(x_i), 1 \le i \le p\}|$$

$$TN_j = |\{x_i \mid y_j \notin Y_i, \wedge y_j \notin h(x_i), 1 \le i \le p\}|$$

$$FN_j = |\{x_i \mid y_j \in Y_i, \wedge y_j \notin h(x_i), 1 \le i \le p\}|$$

$$F1_{mac/mic} = 2 \cdot \frac{precision_{mac/mic} \cdot recall_{mac/mic}}{precision_{mac/mic} + recall_{mac/mic}} \quad (6)$$

where mac is short for macro, and mic is short for micro. Macro/micro F1 score is calculated by corresponding recall and precision, higher value of macro/micro F1 score indicates better multi-label classification results. Macro/micro recall and precision are as follows.

$$precision_{mac} = \frac{1}{k}\sum_{j=1}^{k}\frac{TP_j}{TP_j + FP_j} \quad (7)$$

$$precision_{mic} = \frac{\sum_{j=1}^{k} TP_j}{\sum_{j=1}^{k} TP_j + FP_j} \quad (8)$$

$$recall_{mac} = \frac{1}{k}\sum_{j=1}^{k}\frac{TP_j}{TP_j + FN_j} \quad (9)$$

$$recall_{mic} = \frac{\sum_{j=1}^{k} TP_j}{\sum_{j=1}^{k} TP_j + FN_j} \quad (10)$$

D. Experimental Results

a) Experimental result for supervised TF-IDF: A supervised TF-IDF is proposed in this paper, which calculate the TF-IDF score of the sentences in each classes as sentence representation. In Table I, we show the classification performance of features by supervised TF-IDF and unsupervised TF-IDF on different multi-label algorithms. On the whole, it is obvious that the performance of supervised TF-IDF outperforms than traditional TF-IDF. Experimental result demonstrate the supervised TF-IDF for ABSA problem. Traditional TF-IDF is too spare to classify effectively. Thus a sentence is represented by the sum of all the words in it for each classes in proposed supervised TF-IDF. On the other hand, the performance of LP with SVM base classifier is significantly higher than that of the other methods.

b) Experimental result for BERT with supervised TF-IDF: The result of BERT with supervised TF-IDF for multi-label ABSA is shown in Table II, where our proposed BERT models with supervised TF-IDF indeed present better performances than BERT model. Also, the result of BERT with supervised TF-IDF is better than supervised TF-IDF model, namely our proposed model indeed present the most superior performances. Such results prove that the proposed models are capable of multi-label ABSA.

TABLE II
EXPERIMENT RESULTS OF BERT WITH SUPERVISED TF-IDF

	BERT	BERT with supervised TF-IDF
Accuracy	0.6284	**0.6400**
Hamming Loss	0.0336	**0.0330**
$F1_{mic}$	0.7324	**0.7359**
$F1_{mac}$	0.4793	**0.4936**

V. CONCLUSION

In this paper, we propose a multi-label classification model based on BERT with modified TF-IDF to improve the analytical capabilities in ABSA. In particular, we put forward a modified TF-IDF method by calculating the TF-IDF score in each classes, which indicates the importance of sentences in each classes. The modified TF-IDF method learned the information containing in labels by a supervised method manner, and the comparison between the experimental results of unsupervised TF-IDF and supervised TF-IDF demonstrates the excellent performance of the method we proposed in classification problems. Meanwhile we combine pre-training model BERT with proposed modified TF-IDF in order to learn abundant feature information, and we evaluated our framework on a real-world tasks, achieving slight-improved performance with baseline BERT.

REFERENCES

[1] R. Liang and J. Wang, "A linguistic intuitionistic cloud decision support model with sentiment analysis for product selection in e-commerce," *Int. J. Fuzzy Syst.*, vol. 21, no. 3, pp. 963–977, 2019. [Online]. Available: https://doi.org/10.1007/s40815-019-00606-0

[2] J. Tao and X. Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," *J. Big Data*, vol. 7, no. 1, p. 1, 2020. [Online]. Available: https://doi.org/10.1186/s40537-019-0278-0

[3] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23 522–23 530, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2969854

[4] F. Xu, Z. Pan, and R. Xia, "E-commerce product review sentiment classification based on a nave bayes continuous learning framework," *Information Processing & Management*, vol. 57, no. 5, p. 102221, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457319308040

[5] F. Zhou, R. J. Jiao, X. J. Yang, and B. Y. Lei, "Augmenting feature model through customer preference mining by hybrid sentiment analysis," *Expert Syst. Appl.*, vol. 89, pp. 306–317, 2017. [Online]. Available: https://doi.org/10.1016/j.eswa.2017.07.021

[6] B. Liu and L. Zhang, *A Survey of Opinion Mining and Sentiment Analysis*. Boston, MA: Springer US, 2012, pp. 415–463.

[7] A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and SVM classifier," *World Wide Web*, vol. 20, no. 2, pp. 135–154, 2017. [Online]. Available: https://doi.org/10.1007/s11280-015-0381-x

[8] V. Pérez-Rosas, C. Banea, and R. Mihalcea, "Learning sentiment lexicons in spanish," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, 2012, pp. 3077–3081. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/summaries/1081.html

[9] D. Rao and D. Ravichandran, "Semi-supervised polarity lexicon induction," in *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, 2009, pp. 675–682. [Online]. Available: https://www.aclweb.org/anthology/E09-1077/

TABLE I
EXPERIMENTAL RESULTS OF SUPERVISED TF-IDF

Classifier	Unsupervised TF-IDF				Supervised TF-IDF			
	Accuracy	Hamming Loss	$F1_{mic}$	$F1_{mac}$	Accuracy	Hamming Loss	$F1_{mic}$	$F1_{mac}$
MLkNN	0.0587	0.0706	0.1178	0.0253	**0.5335**	**0.0376**	**0.6557**	**0.3432**
BRkNN-a	0.1657	0.0869	0.1844	0.0182	**0.5214**	**0.0380**	**0.6406**	**0.2504**
BR+SVM	0.0802	0.0631	0.1515	0.0342	**0.5138**	**0.0354**	**0.6602**	**0.3023**
CC+SVM	0.1295	0.0921	0.1553	0.0492	**0.5824**	**0.0400**	**0.6538**	**0.3685**
LP+SVM	0.2495	0.0776	0.3019	0.0592	**0.6025**	**0.0382**	**0.6715**	**0.3409**

[10] A. M. Zubair, K. Aurangzeb, A. Shakeel, Q. Maria, K. I. Ali, and Z. Quan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," *Plos One*, vol. 12, no. 2, p. e0171649, 2017.

[11] M. Rezwanul, A. Ali, and A. Rahman, "Sentiment analysis on twitter data using knn and svm," *International Journal of Advanced Computer ence and Applications*, vol. 8, no. 6, 2017.

[12] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 5876–5883. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16541

[13] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews," *J. Comput. Sci.*, vol. 27, pp. 386–393, 2018. [Online]. Available: https://doi.org/10.1016/j.jocs.2017.11.006

[14] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 2017, pp. 452–461. [Online]. Available: https://doi.org/10.18653/v1/d17-1047

[15] J. Feng, S. Cai, and X. Ma, "Enhanced sentiment labeling and implicit aspect identification by integration of deep convolution neural network and sequential algorithm," *Cluster Computing*, vol. 22, no. Supplement, pp. 5839–5857, 2019. [Online]. Available: https://doi.org/10.1007/s10586-017-1626-5

[16] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *CoRR*, vol. abs/2003.08271, 2020. [Online]. Available: https://arxiv.org/abs/2003.08271

[17] I. M. e. a. Peters M, Neumann M, "Deep contextualized word representations," 2018.

[18] S. T. S. I. Radford A, Narasimhan K, *Improving language understanding by generative pre-training*, 2018.

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[20] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 380–385. [Online]. Available: https://doi.org/10.18653/v1/n19-1035

[21] D. Zeng, Y. Dai, F. Li, J. Wang, and A. K. Sangaiah, "Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 3971–3980, 2019. [Online]. Available: https://doi.org/10.3233/JIFS-169958

[22] J. Tao, L. Zhou, and C. Feeney, "I understand what you are saying: Leveraging deep learning techniques for aspect based sentiment analysis," in *52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019*, 2019, pp. 1–10. [Online]. Available: http://hdl.handle.net/10125/59487

[23] G. Ganu, N. Elhadad, and A. Marian, "Beyond the stars: Improving rating predictions using review text content," in *12th International Workshop on the Web and Databases, WebDB 2009, Providence, Rhode Island, USA, June 28, 2009*, 2009. [Online]. Available: http://webdb09.cse.buffalo.edu/papers/Paper9/WebDB.pdf