Semi-Supervised Learning based on Auto-generated Lexicon using XAI in Sentiment Analysis

Hohyun Hwang, Younghoon Lee
Seoul National University of Science and Technology, Korea, South (KR)
hhhwang94@ds.seoultech.ac.kr, yhoon.lee@seoultech.ac.kr

Abstract

In this study, we proposed a novel Lexicon-based pseudo-labeling method utilizing explainable AI(XAI) approach. Existing approach have a fundamental limitation in their robustness because poor classifier leads to inaccurate soft-labeling, and it lead to poor classifier repetitively. Meanwhile, we generate the lexicon consists of sentiment word based on the explainability score. Then we calculate the confidence of unlabeled data with lexicon and add them into labeled dataset for the robust pseudo-labeling approach. Our proposed method has three contributions. First, the proposed methodology automatically generates a lexicon based on XAI and performs independent pseudo-labeling, thereby guaranteeing higher performance and robustness compared to the existing one. Second, since lexicon-based pseudo-labeling is performed without re-learning in most of models, time efficiency is considerably increased, and third, the generated high-quality lexicon can be available for sentiment analysis of data from similar domains. The effectiveness and efficiency of our proposed method were verified through quantitative comparison with the existing pseudo-labeling method and qualitative review of the generated lexicon.

1 Introduction

Sentiment analysis is employed to identify the sentiment orientation and measure the emotional strength (Khan et al., 2016; Khan & Lee, 2019; Silva et al., 2016). To better understand information generated by online user and to take advantage of it, sentiment analysis is becoming major topic in text mining field in last two decades (Duan et al., 2020; Nagarajan & Gandhi, 2019; Valdivia et al., 2017). Previous studies of sentiment analysis can be roughly categorized into two different groups: 1) lexicon-based approaches and 2) machine learning-based approaches (Khan et al., 2019; Khoo & Johnkhan, 2018).

The lexicon-based approaches efficiently calculate the sentiment score of sentence of document since it does not need to train the classification model in advance. However, The lexicon-based approaches depends on the availability of a sentiment lexicon which is collection of manually pre-created sentiment words lexicon and its sentiment polarity (Huang et al., 2020; Taj et al., 2019; Alqaryouti et al., 2019).

Meanwhile, the machine-learning based approaches require a training set consists of labeled data (e.g. positive, negative or neutral). To address the challenge caused by limited labeled data, which is the usual case in practice, semi-supervised learning have attracted more attention recently (Han et al., 2020; Zhang et al., 2020; Lee et al., 2019). The semi-supervised learning can be divided into several categories such as consistency regularization approaches, entropy minimization approaches and augmentation based approaches and pseudo-labeling approaches.

Among those approaches, the pseudo-labeling approaches are one of the most intuitive and widely used semi-supervised learning in sentiment analysis (Xu & Tan, 2019; Wu et al., 2019; Chen et al., 2020). The pseudo-labeling approaches tried to train the sentiment classification model with small number of labeled data and add unlabeled data with high-confidence of sentiment score, calculated by

Proceedings of Recent Advances in Natural Language Processing, pages 593–600
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_067

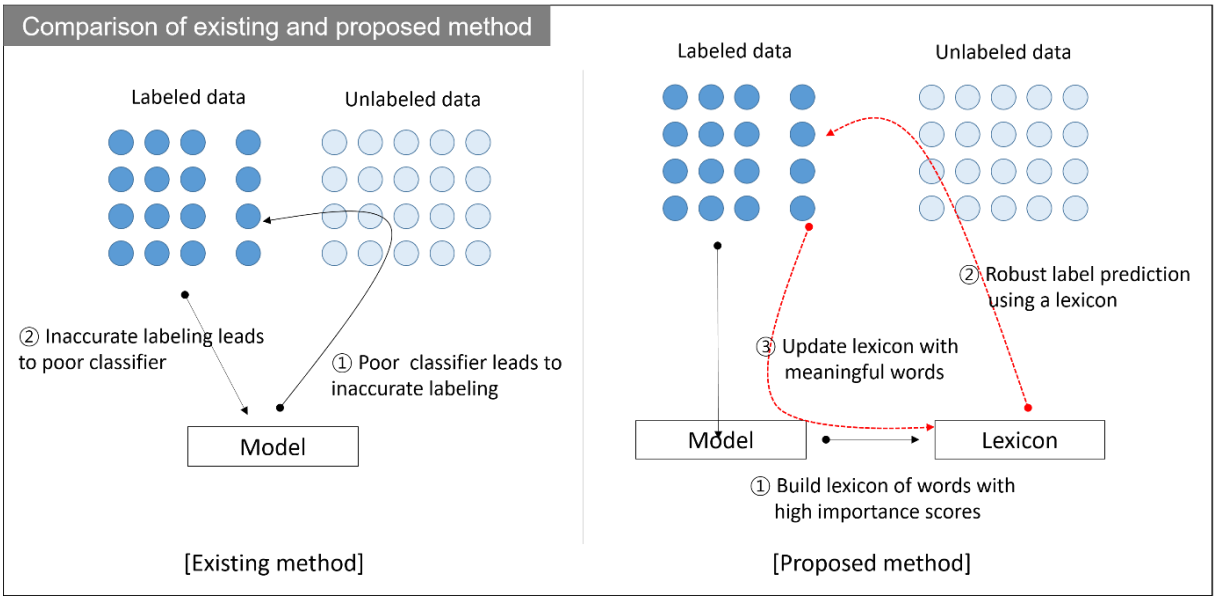


Figure 1 : Limitation of existing pseudo-labeling approach

trained model, into the labeled dataset in each learning cycle.

However, existing approaches have a fundamental limitation in their robustness because soft labeling task and classification task completely depends on each other. That is, poor classifier leads to inaccurate soft-labeling, and it lead to poor classifier repetitively (Van Engelen & Hoos, 2020; Devgan et al., 2020). The left illustration in Figure 1. illustrate the limitation of existing approaches of pseudo-labeling. Thus, this study proposes the robust pseudo-labeling approaches by combining heterogeneous frameworks of sentiment analysis. And the performance of the proposed method will be justified by comparing the two changes in accuracy with graphs.

2 Literature Review

2.1 Studies on semi-supervised learning in sentiment analysis

As a fore-mentioned, the semi-supervised learning can be divided into several categories such as consistency regularization approaches, entropy minimization approaches, augmentation based approaches and pseudo-labeling approaches. The principle of consistency regularization underlines that the model predictions should be less sensitive to the extra perturbation imposed on the input samples (Yu et al., 2020). The entropy minimization approaches encourage the model to output confident predictions on unlabeled data, and the augmentation based approaches are methods of generating various augmented data and using it for learning (Tu & Yang, 2019).

Among those approaches, the pseudo-labeling approaches such as self-training (pseudo-labeling) or co-training is one of the most intuitive and widely used semi-supervised learning in sentiment analysis. In self-training, the most confident unlabeled data with their predicted label, are selected to add to the training set. (Baugh, 2013) employ the self-training for increasing the size of the feature space and (Becker et al., 2013) adapt a static polarity lexicon along with self-training to increase the number of labeled dataset. (Haimovitch et al., 2012) makes use of self-training for large-scale reviews of polarity prediction and (Wang et al., 2016) apply the self-training into text sentiment classification to improve the quality of the training text. (Hajmohammadi et al., 2016) utilized semi-supervised self-training approaches to incorporate unlabelled sentiment documents from the target language in order to improve the performance of cross-lingual methods.

And co-training assumed that feature space can be divided into two different views. Two different classifiers are trained with the labeled data, and then applied to the unlabeled data to add them into trained set with confidence level of prediction. (Yu et al., 2014) focuses on revisiting co-training in depth and discusses several co-training strategies for sentiment analysis following a loose assumption and (Zhang et al., 2014) applies co-

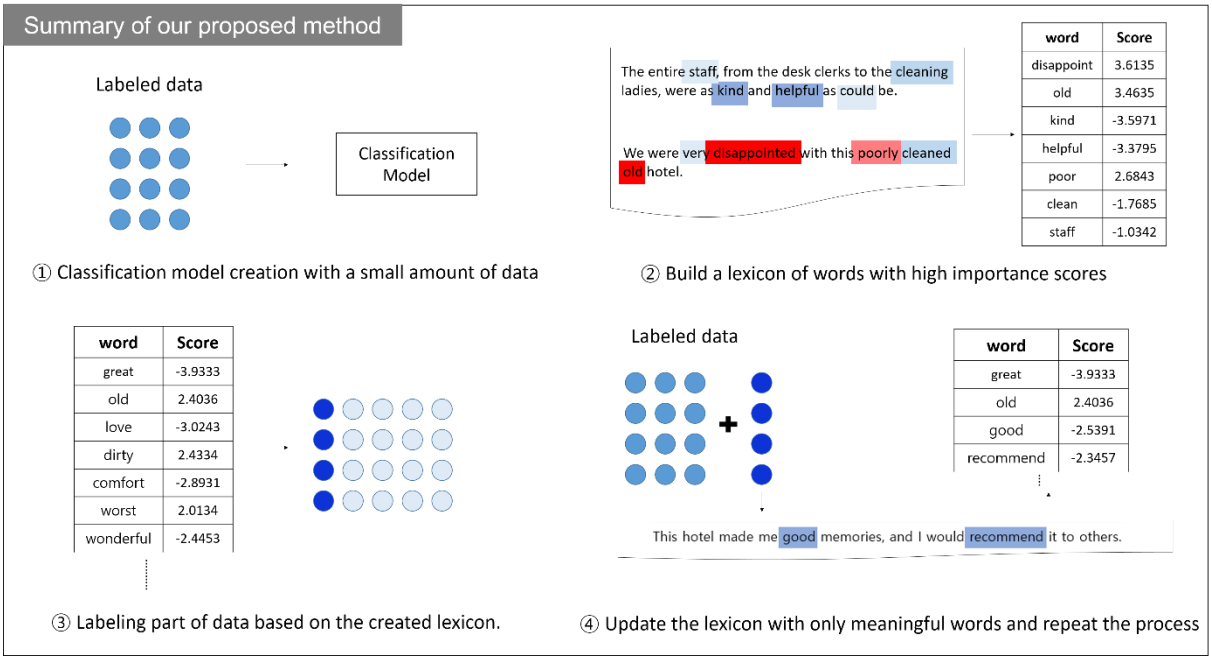


Figure 2 : Summary of our proposed method

training to select the most reliable instances according to the two criteria of high confidence and nearest neighbor for boosting the classifier, also exploit the most informative instances with human annotation for improve the classification performance. (Wang et al., 2014) implemented co-training on multiple component learner of different types to allow performance of their respective advantages and (Xia et al., 2015) propose a dual-view co-training algorithm based on dual-view document representation for semi-supervised sentiment classification. (Catal & Nangir, 2017) investigate the potential benefit of multiple classifier systems concept on Turkish sentiment classification problem with co-training approach and (Li et al., 2019) proposed semi-supervised learning approach based on the hybrid mechanism of self-leaning for textual sentiment classification.

However, the high confidence is not necessarily correct with aforementioned approaches. Label error will be transferred and accumulated in the training and labeling process, and it lead the unstable semi-supervised learning process without robustness. Thus, we proposed robust semi-supervised learning approach by combining heterogeneous framework based on auto-generated lexicon.

2.2 Studies on explainable model

In order to generate the lexicon utilized for soft-labeling automatically. We utilized numerous explainable models such a Local Interpretable Model-Agnostic explanations (LIME), SHapley Additive exPlanations (SHAP), Layer-wise Relevance Propagation (LRP) and Gradient-weighted Class Activation Mapping (Grad-CAM) respectively or in and ensemble. And, we also utilize linear model-agnostic XAI methods such as Logistic regression (LR) and support vector machine (SVM) in our experiments.

The key intuition behind LIME is that it is much easier to approximate a black-box model by a simple model locally (in the neighborhood of the prediction we want to explain), as opposed to trying to approximate a model globally. This is done by weighting the perturbed images by their similarity to the instance we want to explain (Hu et al., 2018; Lee et al., 2020).

SHAP is a method to explain individual predictions. SHAP is based on the game theoretically optimal Shapley Values. SHAP values for each feature represent the change in the expected model prediction when conditioning on that feature. For each feature, SHAP value explains the contribution to explain the difference between the average model prediction and the actual

prediction of the instance (Adadi & Berrada, 2018; Lundberg & Lee, 2017).

Grad-CAM uses the gradients of any target prediction flowing into the certain convolutional layer in CNN model to produce a coarse localization map highlighting the important regions in the image for predicting the class of the image (Lee et al., 2020; Selvaraju et al., 2017). We modified the Grad-GAM algorithm to be applied for textual sentiment classification.

LRP is a method to compute scores for image pixels and image regions denoting the impact of the particular image region on the prediction of the classifier for one particular test image (Binder et al., 2016). We also modified the LRP algorithm to be applied for textual sentiment classification.

3 Method

As aforementioned, instead of existing approaches which calculate the confidence of sentiment score for unlabeled dataset by same classifier, we calculated the confidence of sentiment score for unlabeled dataset by auto-generated lexicon. That is, a lexicon is automatically generated through the explainability score calculated while learning the classifier with labeled data, and pseudo-labels are assigned to the unlabeled dataset based on the generated lexicon. The summary of our proposed method is illustrated in Figure 2.

In detail, the learning process in the work of creating lexicon is as follows. First, a binary classification model is trained using data with positive and negative la bels. And the importance score of each word is grasped through the coefficients derived from each model. An initial lexicon is created based on this importance score.

Second, pseudo-labels are additionally assigned to N unlabeled data using the generated lexicon, and the previous process is repeated using N additional data, and the lexicon is updated based on this result.

When creating a lexicon, each word's importance score is assigned as the average of the word's scores each time the dictionary is updated.

By repeating the process, the lexicon is updated and the process of assigning pseudo-labels to unlabeled data is completed. These processes are defined formally in the Algorithm 1.

Algorithm 1 Creating Sentiment Lexicon

1: Obtain a small set of L of labeled examples
2: Obtain a large set of U of unlabeled examples
3: **for** N iterations **do**
4: **for** each explainable classifier C_i **do**
5: Learn classifier C_i from L
6: **end for**
7: Update lexicon D from ensemble of C_i
8: Choose confidently predicted example E from U based on normalized D
9: E is removed from U and added (with their given labels) to L
10: **end for**

3.1 Details for creating lexicon

As mentioned in the previous section, several criteria were used in updating the lexicon in the proposed method. This section describes the details applied to update the lexicon. And the basic parameter settings used in each methodology are defined in the Table 1. The setting of each parameter was selected based on experience in various experiments.

Explainable Method	Embedding Method	Hyper Parameter		
		initial	update	prediction
LR	TF-IDF	$\lambda = 0.1$	δ=0.2	-
SVM	TF-IDF	$\lambda = 0.1$	δ=0.2	-
LIME	TF-IDF	$\alpha = 20$	$\beta = 10$	$\gamma > 0.8$
SHAP	TF-IDF	$\alpha = 20$	$\beta = 10$	$\gamma > 0.8$
Grad-CAM	Word2vec	-	-	θ<0.25 or θ>0.75
LRP	Word2vec	-	-	θ<0.25 or θ>0.75

Table 1 : Entire parameter setting

3.1.1 Linear model-agnostic approaches

LR and SVM calculate the importance of words using the coefficient values of the classification model. The two models vectorized data and trained the model using Term Frequency – Inverse Document Frequency (TF-IDF).

Calculate the importance score of each word and build a dictionary using only words with a score of λ (= 0.1) or higher. And in the process of updating dictionaries, only words with an importance score of δ (=0.2) or higher are used.

When using a Support Vector Machine, a lexicon was created through a binary classifier through Support Vector Classifier. Similar to Logistic Regression, words were generated by calculating

regression coefficients for each word learned in the model.

3.1.2 XAI-based approaches

Unlike linear model agnostic approaches, in XAI-based approaches, scores are assigned to words per sentence. That is, in the process of constructing a lexicon, a score is calculated and updated one by one.

In this study, the importance of words was calculated for each sentence by applying LIME and SHAP to the model trained by Logistic Regression, and the importance of words was calculated for each sentence through Grad-CAM for the model trained with CNN and LRP for the model trained with LSTM. We proceeded to calculate the importance. LIME and SHAP used TF-IDF matrix to vectorize sentences, and CNN and LSTM training data used Word2Vec to embed sentences.

To construct a meaningful lexicon, not all sentences are used for lexicon construction, but only sentences with a predicted value of 0.8 or higher are used, and only the top 20 words of importance score are used in each sentence. In the process of updating a dictionary, the dictionary is updated using only the top 10 words in the sentence.

The process of building a lexicon through Grad-CAM and LRP, use only sentences with sigmoid values greater than 0.75, or less than 0.25, close to zero and one.

Lastly, considering the characteristics of Grad-CAM, which does not show directionality, the frequency of each word in the positive lexicon and the negative lexicon is compared and set as a positive word or negative word.

4 Experiment

4.1 Data description

In this study, experiments were conducted using 7 open datasets. The data used were composed of various domains such as movies, accommodations, games, shopping, airlines, and clothing. The Table 2 below summarizes the description of the dataset.

4.2 Experiment setup

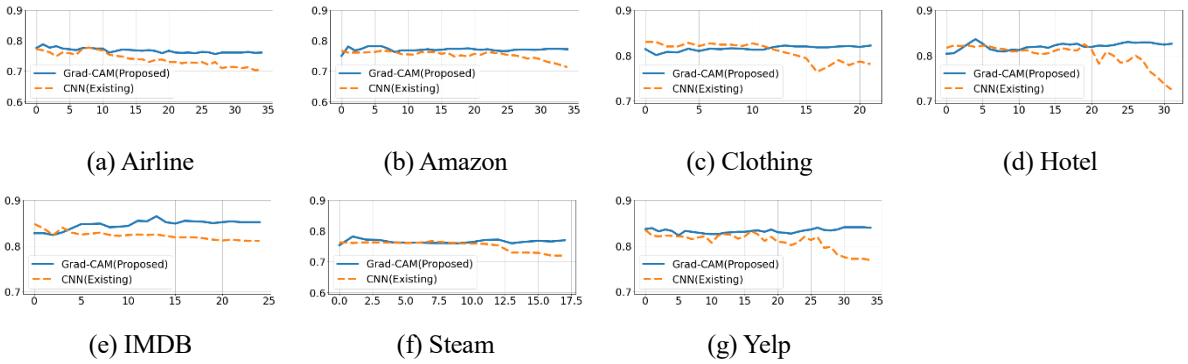
In the experiment, we basically verify that the proposed method shows higher performance and robustness than the existing pseudo-labeling method in each dataset. For the performance comparison in the same experimental setting, the same baseline architecture was used, and accordingly, a one-to-one comparison was performed as follows: 1) LR based existing pseudo-labeling approach vs. LR based proposed method, 2) SVM based existing pseudo-labeling approach vs. SVM based proposed method, 3) LR based existing pseudo-labeling approach vs. LIME based proposed method, 4) LR based existing pseudo-labeling approach vs. SHAP based proposed method, 5) CNN based existing pseudo-labeling approach vs. Grad-CAM based proposed method, and 6) LSTM based existing pseudo-labeling approach vs. LRP based proposed method.

The experimental setup of the proposed method is as follows. First, the initial emotion lexicon is constructed using 1000 positive and 1000 negative sentences. Then, based on the lexicon, pseudo-labeling is repeatedly performed by 1000 pieces, and the emotional lexicon update is performed again using the data. In addition, 1000 pieces that were not used for learning were set as test data, and the change in accuracy of pseudo-labeling based on the lexicon was measured.

The experiment of the existing pseudo-labeling methodology was carried out as follows. As in the experiment of the proposed methodology, a classifier is created using 2000 data (1000 positive sentences and 1000 negative sentences), and prediction is performed with an additional 1000 data units. Among them, pseudo-labeling was performed on 100 data, which are the top 10% of

Dataset	Num. of instance	Pos.	Neg.	Maximum length of reviews	Average length of reviews	Num. of vocabs
Airline review	74,623	37,352	37,271	385	23	1,146
Amazon review	400,000	200,000	200,000	86	30	1,887
Clothing review	23,486	19,314	4,172	54	24	1,046
Hotel review	38,932	26,521	12,411	606	71	2,301
IMDB review	25,000	12,500	12,500	776	101	4,366
Steam review	17,494	9,968	7,526	900	64	2,213
Yelp review	38,000	19,000	19,000	381	56	2,451

Table 2 : Summary for dataset



(a) Airline (b) Amazon (c) Clothing (d) Hotel

(e) IMDB (f) Steam (g) Yelp

Figure 3 : Comparison of results of existing pseudo-labeling(CNN) vs Proposed method (Grad-CAM)

the predicted values, and the change in accuracy of this data was measured.

4.3 Experiment result

Throughout the model, the figures of accuracy for the data were shown similarly. Initially, the accuracy of the automatic generation lexicon was lower compared to the general semi-supervised learning method, but over time, the accuracy of the semi-supervised learning decreased and the accuracy of the proposed method was maintained or increased.

In this study, we conducted an experiment comparing the accuracy of the automatic generation-based method using six methods and the existing method using seven data. In this section, we present Figure 3. only graphs comparing the accuracy of methods that conducted Pseudo-labeling based on CNNs and the accuracy of proposed methods that utilize automatic generative lexicons based on Grad-CAM. In the case of Grad-CAM, it can be seen that the proposed method shows better performance than the existing method at all times. In particular, in the case of the existing methodology, it can be confirmed that the classification performance rapidly decreases after the initial classification performance is poor.

5 Conclusion

In this study, a novel Lexicon-based pseudo-labeling method utilizing XAI approach was proposed that improved the limitations of the existing pseudo-labeling method. The existing approaches have a fundamental limitation in their robustness because soft labeling task and classification task completely depends on each other.

However, the proposed methodology automatically generates a lexicon based on XAI and performs independent pseudo-labeling, thereby guaranteeing higher performance and robustness compared to the existing one. In addition to robustness, since dictionary-based pseudo-labeling is performed without re-learning, time efficiency is considerably increased, the generated high-quality lexicon can be available for sentiment analysis of data from similar domains.

The quantitative excellence of the proposed method was verified through a one-to-one performance comparison with the existing method, and the effectiveness and efficiency of the proposed method were qualitatively verified by reviewing the generated lexicon.

Future research may extend the scope of XAI based lexicon construction in a more general point of view. As shown in the experimental results, there are differences in lexicons for each domain, and a study to construct a general-domain lexicon by integrating them is presented as a future work. Moreover, such studies can be expected to aid in the widespread application of the proposed semi-supervised learning in various tasks arising within the natural language processing domain.

Acknowledegments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2020R1F1A1067914).

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6, 52138–52160.

Alqaryouti, O., Siyam, N., Monem, A. A., & Shaalan, K. (2019). Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics*, 2210–8327.

Baugh, W. (June). bwbaugh: Hierarchical sentiment analysis with partial self-training. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 539-542). Association for Computational Linguistics.

Becker, L., Erhart, G., Skiba, D., & Matula, V. (2013). Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, (pp. 333-340). Association for Computational Linguistics.

Binder, A., Bach, S., Montavon, G., Müller, K. R., & Samek, W. (2016). Layer-wise relevance propagation for deep neural network architectures. In K. Kim, & N. Joukov (Eds.), *Information science and applications (ICISA) 2016* (pp. 913-922). Springer.

Catal, C., & Nangir, M. (2017). A sentiment classification model based on multiple classifiers. *Applied Soft Computing, 50,* 135–141..

Chen, J., Feng, J., Sun, X., & Liu, Y. (2020). Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts. *Symmetry*, 12(1), 8.

da Silva, N. F. F., Coletta, L. F., Hruschka, E. R., & Hruschka Jr, E. R. (2016). Using unsupervised information to improve semi-supervised tweet sentiment classification. *Information Sciences, 355*, 348–365.

Devgan, M., Malik, G., & Sharma, D. K. (2020). Semi-Supervised Learning. *Machine Learning and Big Data: Concepts, Algorithms, Tools and Applications*, 251–280.

Duan, J., Luo, B., & Zeng, J. (2020). Semi-supervised learning with generative model for sentiment classification of stock messages. *Expert Systems with Applications, 158,* 113540.

Haimovitch, Y., Crammer, K., & Mannor, S. (2012, November). More is better: Large scale partially-supervised sentiment classification. In *Asian Conference on Machine Learning* (pp. 175-190). PMLR

Hajmohammadi, M. S., Ibrahim, R., Selamat, A., & Fujita, H. (2015). Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Information sciences, 317*, 67–77.

Han, Y., Liu, Y., & Jin, Z. (2020). Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. *Neural Computing and Applications, 32*(9), 5117–5129.

Hu, L., Chen, J., Nair, V., & Sudjianto, A. (2018). Locally interpretable models and effects based on supervised partitioning (lime-sup). *arXiv preprint arXiv:1806.00663.*

Huang, M., Xie, H., Rao, Y., Liu, Y., Poon, L., & Wang, F. (2020). Lexicon-based sentiment convolutional neural networks for online review analysis. *IEEE Transactions on Affective Computing*

Khan, F. H., Qamar, U., & Bashir, S. (2016). Sentimi : Introducing point-wise mutual information with sentiwordnet to improve sentiment polarity detection. *Applied Soft Computing*, 39, 140–153.

Khan, J., Alam, A., Hussain, J., & Lee, Y. (2019). Enswf: effective features extraction and selection in conjunction with ensemble learning methods for document sentiment classification. *Applied Intelligence, 49*(8), 3123–3145.

Khan, J., & Lee, Y. (2019). Lessa: A unified framework based on lexicons and semi-supervised learning approaches for textual sentiment classification. *Applied Sciences*, 9(24), 5562.

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science, 44*(4), 491–511.

Lee, V. L. S., Gan, K. H., Tan, T. P., & Abdullah, R. (2019). Semi-supervised learning for sentiment classification using small number of labeled data. Procedia Computer Science, 161, 577–584.

Lee, Y., Park, J., & Cho, S. (2020). Extraction and prioritization of product attributes using an explainable neural network. *Pattern Analysis and Applications, 23*(4), 1767–1777.

Li, Y., Lv, Y., Wang, S., Liang, J., Li, J., & Li, X. (2019). Cooperative hybrid semi-supervised learning for text sentiment classification. *Symmetry*, 11(2), 133.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *In Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768-4777).

Nagarajan, S. M., & Gandhi, U. D. (2019). Classifying streaming of Twitter data based on sentiment analysis using hybridization. *Neural Computing and Applications, 31*(5), 1425–1433.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626).

Taj, S., Shaikh, B. B., & Meghji, A. F. (2019, January). Sentiment analysis of news articles: A lexicon based approach. In 2019 2nd International Conference on Computing, *Mathematics and Engineering Technologies (iCoMET)* (pp. 1-5). IEEE.

Tu, E., & Yang, J. (2019). A review of semi supervised learning theories and recent advances. *arXiv preprint arXiv:1905.11590*

Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment analysis in tripadvisor. *IEEE Intelligent Systems, 32*(4), 72–77.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning, 109*(2), 373–440.

Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision support systems, 57*, 77–93.

Wang, Z., Feng, Y., Qi, T., Yang, X., & Zhang, J. J. (2016). Adaptive multi-view feature selection for human motion retrieval. *Signal Processing, 120*, 691–701.

Wu, C., Wu, F., Wu, S., Yuan, Z., Liu, J., & Huang, Y. (2019). Semi-supervised dimensional sentiment analysis with variational autoencoder. *Knowledge-Based Systems*, 165, 30–39.

Xia, R., Wang, C., Dai, X., & Li, T. (2015). Co-training for semi-supervised sentiment classification based on dual-view bags-of-words representation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1054-1063). Association for Computational Linguistics.

Xu, W., & Tan, Y. (2019). Semi-supervised target-oriented sentiment classification. *Neurocomputing, 337*, 120–128.

Yu, K., Ma, H., Lin, T. R., & Li, X. (2020). A consistency regularization based semi-supervised learning approach for intelligent fault diagnosis of rolling bearing. *Measurement, 165*, 107987.

Yu, N. (2014). Exploring co-training strategies for opinion detection. *Journal of the Association for Information Science and Technology, 65*(10), 2098–2110.

Zhang, D., Li, S., Zhu, Q., & Zhou, G. (2020). Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning. *IEEE Access, 8*, 22945–22954.

Zhang, Y., Wen, J., Wang, X., & Jiang, Z. (2014). Semi-supervised learning combining co-training with active learning. *Expert Systems with Applications, 41*(5), 2372–2378.