[CL-AFF Shared Task] Multi-label Text Classification Using an Emotion Embedding Model

Jiwung Hyun[1], Byung-Chull Bae[2], and Yun-Gyung Cheong[1] ✉

[1] College of Computing, Sungkyunkwan University, Suwon-si, South Korea
{kabbi159, aimecca}@skku.edu
[2] School of Games, Hongik University, Sejong-si, South Korea
byuc@hongik.ac.kr

Abstract. In this paper, we propose a deep learning model-based approach that combines a language embedding model and an emotion embedding model in the classification of text for the CL-AFF Shared Task 2020. The task aims to predict the disclosure and supportiveness labels of the comments (to the posts) in the OffMyChest dataset which consists of a small labeled dataset and a large unlabeled dataset. We investigate the effectiveness of the BERT, Glove, and Emotional Glove embedding models, to represent the text for label prediction. We also propose to use the original posts in the dataset as contextual information. We evaluated our approach and report the results.

Keywords: Semisupervised learning · Emotion embedding · BERT.

1 Introduction

This paper describes our approach to solve the CL-AFF (Computational Linguistics - Affect Understanding) Shared Task 2020. In the CL-AFF 2020 task, the OffMyChest conversation dataset is introduced to help understand the role of emotion in conversations (see [3] for details). The dataset consists of top posts and the comments to these posts collected from the CasualConversations and the OffMyChest communities on Reddit. A small portion of the comments are labeled as informational disclosure, emotional disclosure, and supportiveness, where supportiveness is further characterized as general, informational, and emotional. As a result, a comment text is annotated with a total of 6 labels, where a single comment can have multiple labels. Therefore, this task involves multi-label classification problems.

Text classification in natural language processing has traditionally employed classification algorithms in machine learning such as support vector machines, Bayesian classifiers, decision trees, etc. A variety of features in text are given as the input of the classifiers, which are crucial to traditional text classification algorithms. For the representation of text, word frequency-based approaches (e.g., bag-of-words feature) or sequence-based approaches (e.g., N-grams feature) have been commonly used.

 In: N. Chhaya, K. Jaidka, J. Healey, L. H. Ungar, A. Sinha (eds.): Proceedings of the 3rd Workshop of Affective Content Analysis, New York, USA, 07-FEB-2020, published at http://ceur-ws.org

Text classification using neural models are comprised of embedding models and classification models. Along with the success of word embedding models such as Word2vec [6] and Glove [7] in text classification, advances of various deep learning algorithms have lead to more complex embedding models, such as contextual language model, also known as ELMo [8] and BERT [1]. Deep neural networks are used not only to extract features from text but also to construct classifiers. Kim [4], for example, presented great performances in text classification by applying 1D CNN on sentence classification problems.

In this paper, focusing on emotional words in the data, we propose a deep learning model-based text classification approach using an embedding model, which has the combination of a language model and an emotion embedding model. We particularly focus on emotional words from OffMyChest dataset to improve learning emotional labels (emotional disclosure, emotional support). We combine labeled and unlabeled comments data for our semi-supervised method to help supervised learning. And then the sentence features extracted from the embedding model are given as TextCNN[4] input for text classification.

Furthermore, to improve the classification performance, we apply EDA (Easy Data Augmentation)[10] on small labeled data in the training step. This paper presents our baseline and variables in our experiment, as well as the evaluation models for binary classification of disclosure and supportiveness (Task 1).

2 The OffMyChest Conversation Dataset

OffMyChest conversation dataset is comprised of three sets - labeled training set, unlabeled training set, and test set: unlabeled training data set includes unlabeled posts and comments; labeled training set and unlabeled test set include about 10,000 labeled sentences and 3,000 unlabeled sentences respectively from the top commented posts. Table 1 lists several excerpts sampled from the labeled data.

Text	id	ED	ID	S	GS	IS	ES
Hope you have a nice day	91px39	0	0	1	1	0	0
My wife came in when I was around half way through this and asked why I was all choked up and watery eyed, so we read it together and now we're both crying.	91px39	1	1	0	0	0	0
I am crying a lot of happy tears right now.	91px39	1	0	0	0	0	0
He's my father in every sense of the word but name, I still call him by his first name but only because we are both used to it and he doesn't mind a bit.	946qw9	1	0	0	0	0	0
Stepdad will be the one walking me down the aisle when I get married.	946qw9	1	1	0	0	0	0
dThat's wonderful. :) My step-dad has been around for 30 yrs now.	946qw9	1	1	1	1	0	1

Table 1: Example Sentences sampled from the labeled dataset, where ***ED*** denotes **E**motional Disclosure, ***ID*** denotes **I**nformational **D**isclosure, ***S*** denotes **S**upport, ***GS*** denotes **G**eneral **S**upport, ***IS*** denotes **I**nformational **S**upport, and ***ES*** denotes **E**motional **S**upport.

Category	**Number of label 1**	**Percentage in the category**
Emotional Disclosure	3,948	31%
Informational Disclosure	4,891	38%
Support	3,226	25%
General Support	680	5%
Informational Support	1,250	10%
Emotional Support	1,006	8%
None (all label is 0)	4,157	32%
All	12,860	

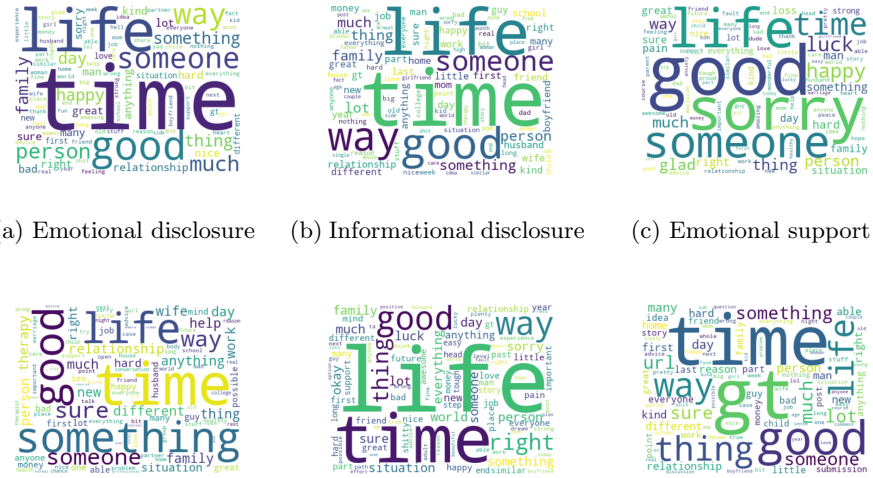
Table 2: A simple statistical analysis of the labeled training data

Words	**No. of Occurrences in Emotional Disclosure (N=3,948)**	**No. of Occurrences in Emotional Support (N=1,006)**	**Normalized Frequency**
know	203	66	117.02
really	227	55	112.17
get	213	57	110.61
hope	99	82	106.59
sorry	73	88	105.97
people	219	45	100.20
you	82	65	85.382
even	139	32	67.017
right	83	36	56.808
things	111	28	55.949

Table 3: Top 10 words in the emotional categories, ranked by the frequency the occurrences normalized by the number of labels in each category. N denotes the number of 1 labels in the category.

To understand the data, we examine the number of labeled data in each category (Table 1). As seen in the table, the number of the data labeled as class 1 in the 'Support' category groups (e.g., support, general support, informational support, emotional support) are far less than the number of data labeled as class 0. For instance, the data labeled as 1 in the 'General Support' category occupy only 5% of the a the data. This means that the data exhibits class imbalance problems, severely in the labels of 'general support', 'informational support', 'emotional support'. Although these categories seem to be the sub-types of the 'Support' category, we treat them independently for the classification because the sum of all their instances with label 1 does not match with the number of instances in the 'Support' category (N=3,226).

Furthermore, we investigate the word usage in the emotional categories. First, we extract the top 100 most frequently used words in each emotional category. Then, we remove the words that also appear frequently in the non-emotional categories (e.g., informational disclosure, support, general support, informational support). Finally, we normalize the raw count by the number of labels of the



(a) Emotional disclosure (b) Informational disclosure (c) Emotional support

(d) Informational support (e) general support (f) no label

Fig. 1: Word clouds for each label group

corresponding category and sum up the two numbers. Table 4 shows the top 10 words ranked by the normalized frequencies. Therefore, these words can be used to characterize the emotional categories.

Finally, Figure 1 shows the word cloud for each category, visualizing the frequently used words in the categories. While only nouns are generally used to construct word clouds, we also use adjectives as they can represent emotion. The word clouds show that many words (e.g., life, time, good) overlap across different categories. There are some words unique to each group: for example, 'way' in the informational disclosure category, and 'sorry' in the emotional support category.

3 Approach

This section details our approach which employs pre-trained embedding models to generate the vectors representing the text. These vectors serve as the input for the textCNN [4] model for multi-label classification. We also investigate if the use of the post as contextual information can enhance the label prediction performance Figure 2 illustrates the overall architecture of our system.

3.1 Word Embedding Models

As our word embedding model, we utilize the pre-trained BERT [1] and an emotional embedding scheme [9]. The original posts of comments are generally lengthy. Therefore, we summarize them into 3∼5 sentences using LexRank [2] before text processing. As the first text pre-processing step, the sentences are

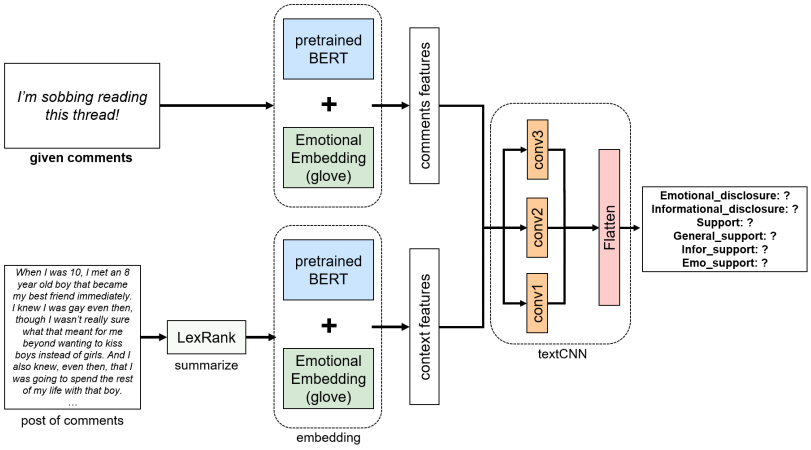


Fig. 2: Overall System Architecture

tokenized using the BERT tokenizer. We set the max sequence length of tokens as 64, because the average sequence length of comments data is 19 and only 1% of the total training data exceeds 64 tokens. As a result, the feature vector of a single comment consists of the 64 tokens representing the comment itself and the 64 tokens representing its corresponding post.

Additionally, we use an emotional embedding model that incorporates emotional information into a word embedding model such as Word2vec [6], Glove [7], etc. Emotional embedding refers to a new vector space by fitting emotional information into pre-trained word vectors. Constraint set constructed by all word/emotion relations are used for training. It is learned to get closer to the pairs of words that are in positive relation. We use pre-trained emotional embedding Glove [9]. As the vocabulary in BERT is different from that of Glove, we set zero embedding when the token tokenized by BERT is not present in the Glove vocabulary. Features from the pre-trained BERT and emotional embedding Glove are concatenated. As a result, the features vector for a comment has (62, 1068) shape, representing 62 tokens of 1068 dimensions (768 BERT dimension + 300 emotional Glove dimension 300). When the post is used as a context, the input of our CNN model has (124, 1068) shape.

3.2 TextCNN

Our textCNN model uses one-dimension convolutions with the filter size of 256 and the kernel size of 3, 4, 5. After max pooling on convolutions result, the features are concatenated and flatten to be connected with a fully-connected layer as the final layer for the classification. The output dimension is 6, and the sigmoid function is used as the activation function. The output represents prediction probability of each class; the probability of greater than or equal to

0.5 is labeled as 1, and otherwise labeled as 0. We adopt the Adam optimizer with learning rate 1e-4 and epsilon 1e-8 in this study. The binary cross entropy loss function is used to train our model.

3.3 Data Augmentation for Training

Via informal experimentations, we discovered that the prediction performance of *'General_support'* is very low. We attribute this to the small number of the label 1 data. On the other hand, the numbers of the *'Info_support'* and the *'Emo_support'* labels are 1250 and 1006 respectively. Those labels show half the performance of the rest of labels. To address this class imbalance issue, we apply EDA (Easy Data Augmentation) [10] to the general support, informational support, and the emotional support categories. EDA uses synonym replacement, random insertion, random swap, and random deletion for augmentation. We augment 9 sentences for each sentence of the support group categories during the training stage in the system run.

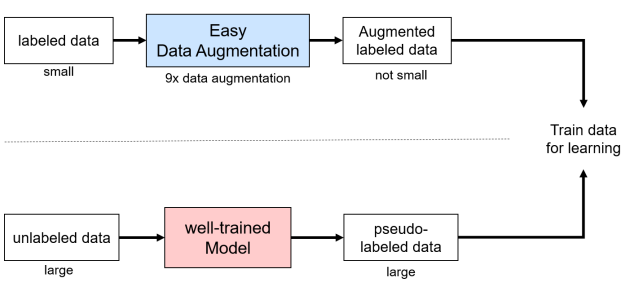


Fig. 3: Our Data Augmentation and Semi-supervised Method

3.4 Semi-supervised Learning Method

The OffMyChest dataset also provides unlabeled comment data, which contain over 420,000 sentences. To make use of the data, we assign pseudo-labels to the unlabeled comments data using the best classification model. Then we re-train the model with the labeled data along with the pseudo-labeled data to improve the classification performance following the semi-supervised method in [5].

When training, we want the model to learn from the labeled data more than the pseudo-labels as the pseudo-labels can be incorrect. Thus, we build the initial model trained with the augmented labeled data on 10 epochs. Then, we re-train the model with the pseudo-labeled data on 3 epochs. While we applied the semi-supervised learning for the system run submission, the experiment results presented below are obtained without the semi-supervised learning scheme because our limited computing facility cannot allow the 10-fold cross validation.

4 Evaluation

To evaluate the proposed approach in this paper, we use 10-fold cross validation on the labeled training data. In this experimentation, EDA and semi-supervised learning were not used due to limited computing resources for 10-fold cross validation. The different conditions examined in our experiments are described as below.

Glove [7]: use of the pre-trained Glove trained with Twitter data. That was our first approach to solve this problem, however not used exclude this experiment. Since the data we use in this study were collected from the Internet community Reddit, we expected that the pre-trained Glove model may show good performance. The pre-trained model uses a 200 dimension vector with 27B tokens.

Emotional Glove [9]: emotional embedding model using an approach combining emotional information with Glove embedding. This model uses a 300 dimension vector with 6B tokens.

BERT [1]: using the state-of-the-art language model to generate the feature vector. Embedding dimension is 768, and max sequence length is 64.

Context: utilizing the posts as a contextual information to test if it can enhance the prediction performance.

4.1 Results

Table 4 reports the accuracy of label prediction. Overall, the BERT embedding model without using the post shows the best performance in accuracy for three categories. The combination of BERT with Emotional Glove results in the best performance in the emotional disclosure category (without context) and the general support category (with context). The combination of BERT with Glove results outperforms the other models in the information support category. The performance in accuracy seems promising in the support group categories, ranging from 0.814(support) to 0.938(general support). Yet, their F1 scores show different findings.

Table 5 shows the F1 scores of our model in each category. Overall, the *'BERT + Emotional Glove'* model and the *'BERT + Glove'* model show the best performance. The performance of the support subgroups (i.e., General Support, Informational Support, and Emotional Support) are poor, as low as 0.05 (for the general support label prediction), which are not sufficient for its practical use. Meanwhile, its corresponding accuracy is 0.934. This means that accuracy is not a good metric when the class is imbalanced. Precision and recall performances of the selected model are described in the following Table 6.

Methods	Emotional Disclosure	Information Disclosure	Support	General Support	Information Support	Emotional Support	Micro Average
Glove (Baseline)	0.662	0.645	0.767	0.935	0.882	0.915	0.807
Emotional Glove	0.672	0.666	0.773	0.937	0.891	0.919	0.816
BERT	0.693	**0.673**	**0.814**	0.935	0.897	**0.924**	**0.829**
BERT + context	0.682	0.664	0.785	0.937	0.893	0.915	0.819
BERT + Glove	0.684	0.663	0.808	0.934	**0.898**	0.922	0.825
BERT + Emotional Glove	**0.696**	0.662	0.811	0.935	0.896	0.922	0.827
BERT + Emotional Glove + context	0.677	0.664	0.789	**0.938**	0.892	0.915	0.819

Table 4: Accuracy of our models

Methods	Emotional Disclosure	Information Disclosure	Support	General Support	Information Support	Emotional Support	Micro Average
Glove (Baseline)	0.383	0.470	0.464	0.033	0.161	0.163	0.422
Emotional Glove	0.323	0.505	0.422	0.021	0.089	0.137	0.406
BERT	0.438	0.534	0.544	0.049	0.222	0.215	0.494
BERT + context	0.270	0.448	0.500	0.023	0.189	0.206	0.412
BERT + Glove	**0.464**	**0.544**	0.525	**0.050**	0.228	0.222	**0.503**
BERT + Emotional Glove	0.445	0.526	**0.557**	0.045	**0.251**	**0.237**	0.499
BERT + Emotional Glove + context	0.350	0.467	0.485	0.006	0.216	0.163	0.429

Table 5: F1 score of our models

Methods	Emotional Disclosure		Information Disclosure		Support		General Support		Information Support		Emotional Support		Micro Average	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT + Emotional Glove	0.507	0.396	0.575	0.484	0.591	0.527	0.051	0.040	0.285	0.224	0.264	0.215	0.575	0.441
BERT + Emotional Glove + context	0.430	0.295	0.577	0.392	0.546	0.437	0.007	0.005	0.250	0.190	0.192	0.142	0.555	0.349

Table 6: Precision and recall of our selected model

4.2 Discussions

The evaluation results indicate the followings:

- As for the evaluation measures, accuracy is a poor metric to evaluate the proposed approach due to class imbalance problems. Therefore, we propose to use an F1 score, a harmonic mean of precision and recall, instead.
- For the word embedding model, the BERT models perform better than the Glove models.

- The combination of BERT and Glove enhances the classification performance.
- The use of emotional Glove improves the classification performance for the categories where classes are imbalanced (i.e., support, information support, emotional support).
- Opposed to our initial assumption, the use of the original post as a context did not contribute to enhancing the prediction performance. We postulate that it is because our method of concatenating one comment to one context (i.e., a post associated with the comment) fails to give proper weight to contexts, as the comments data are presumably labeled without considering the contexts.

For the CL-Aff Shared task 2020 competition, we used *'BERT + Emotional Glove'* models as the system run model, as it reports the best F1 scores without considering context. In the system run we applied EDA (Easy Data Augmentation) and semi-supervised learning as described in Section 3.3 and 3.4. Our submission contains the labels generated using 4 different settings from the combination of ['with context' and 'without context'] and ['with pseudo-label' and 'without pseudo-label'].

5 Conclusion

This paper describes our approach that combines the word and emotion embedding models to predict 'Disclosure' and 'Supportiveness' in OffMyChest dataset. Three language embedding models - BERT, Glove, and Emotional Glove - were compared, and BERT showed a better performance (about 10%) for the label prediction than Glove. Our evaluation results also indicate that the combination of embedding models can improve the performance. It is particularly noted that Emotional Glove better represents the text than Glove when the class is imbalanced. We adopted the original posts along with their associated comments to increase the prediction performance. However, the result shows that using the post makes no contribution to increasing the model's classification performance. In the future, we plan to investigate how to convey the context of a post in an efficient manner rather than just concatenating to increase prediction performance.

Acknowledgement

This research was partially supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2016R1D1A1B03933002). This work was partially supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2019R1A2C1006316). This work was also partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2017R1A2B4010499).

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
2. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res. **22**(1), 457–479 (Dec 2004), http://dl.acm.org/citation.cfm?id=1622487.1622501
3. Jaidka, K., Singh, I., Jiahui, L., Chhaya, N., Ungar, L.: A report of the CL-Aff OffMyChest Shared Task at Affective Content Workshop @ AAAI. In: Proceedings of the 3rd Workshop on Affective Content Analysis @ AAAI (AffCon2020). New York, New York (February 2020)
4. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1181, https://www.aclweb.org/anthology/D14-1181
5. Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. ICML 2013 Workshop : Challenges in Representation Learning (WREPL) (07 2013)
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, Curran Associates Inc., USA (2013), http://dl.acm.org/citation.cfm?id=2999792.2999959
7. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162
8. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proc. of NAACL (2018)
9. Seyeditabari, A., Tabari, N., Gholizadeh, S., Zadrozny, W.: Emotional embeddings: Refining word embeddings to capture emotional content of words. CoRR **abs/1906.00112** (2019), http://arxiv.org/abs/1906.00112
10. Wei, J.W., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. CoRR **abs/1901.11196** (2019), http://arxiv.org/abs/1901.11196