An unsupervised approach for sentiment analysis via financial texts

Cong Chi Pham[1], Bay Van Nguyen[1], Huy Quoc Nguyen[1*]

[1]Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[*]Corresponding author: huy.nq@ou.edu.vn

ARTICLE INFO

DOI:10.46223/HCMCOUJS.tech.en.15.2.3684.2025

Received: August 23rd, 2024

Revised: September 28th, 2024

Accepted: October 17th, 2024

Keywords:

autoencoder; deep clustering; natural language processing; transformer; unsupervised sentiment analysis

ABSTRACT

The rapidly increasing volume of textual data has made manual labeling extremely costly and time-consuming. To address this limitation, researchers have gradually focused on unsupervised learning techniques that enable models to classify text without relying on labeled data. Among these, deep clustering has garnered significant interest. However, most existing deep clustering methods are primarily designed for computer vision tasks. In this paper, we propose modifications to two of the most powerful deep clustering methods, including DEKM and DeepCluster, by integrating transformer algorithms in the Natural Language Processing (NLP) domain, enabling these methods to handle textual data. With the proposed methods, we achieved the best results on the test set of the Financial Phrase Bank (FPB) dataset with an accuracy of 57.71% and on the test set of the Twitter Financial News (TFN) dataset with an accuracy of 65.58%. Although these results are still lower than those of traditional supervised deep learning methods, we have demonstrated that the performance of our proposed methods can be further improved when trained with more data. This highlights the promising potential of deep clustering methods for natural language processing tasks. Especially when addressing tasks where the data is either unlabeled or lacks sufficient labeling.

1. Introduction

In recent years, the vast amount of financial text data available from news articles, financial reports, social media, and other sources has presented both a challenge and an opportunity for business people. Analyzing this wealth of unstructured data can yield valuable insights into market trends, sentiment, and economic indicators, which are crucial for making financial decisions. Nowadays, the evaluation and categorization of financial texts are primarily based on the feelings and assessments of experts. However, this method is one-sided, which is unsuitable for sentiment analysis. Therefore, researchers have applied deep learning methods to analyze better the context and relationships of the words in the financial texts. Traditional sentiment analysis methods mainly rely on supervised learning, which requires much-labeled data. Specifically, deep learning models use labeled data to evaluate the predicted results, thereby optimizing the weights of deep learning models and improving the efficiency of sentiment analysis systems. However, manually labeling a massive text volume is time-consuming and requires significant skilled labor. This opens a potential avenue for unsupervised learning approaches.

Deep clustering is an emerging field within unsupervised deep learning (Ren et al., 2024), which can offer a promising solution to this challenge. By leveraging deep neural networks' powerful feature extraction capabilities and integrating them with clustering algorithms, such as K-Means, deep clustering enables the automatic categorizing of financial texts based on their inherent similarities. The main idea behind deep clustering is the simultaneous training of feature extraction and clustering, where each process mutually reinforces the other. We employ autoencoder, a neural network for unsupervised learning, to integrate with deep clustering algorithms. The autoencoder learns the features of the text by attempting to reconstruct the original data based on the extracted features. This efficient self-learning way allows the essential features of the text to be fully extracted without the need for labeled data.

Inspired by the study of Deep Embedded K-Means Clustering (DEKM) (Guo et al., 2021), a deep clustering method for image classification, we replace their Autoencoder backbone with a Natural Language Processing (NLP) transformer since the vanilla DEKM is based on CNN. This makes our customized DEKM suitable for processing with text. On the other hand, we also apply a transformer-based feature extractor for DeepCluster, another robust deep clustering method presented by Caron et al. (2018), to tackle this unsupervised task. In this paper, we proposed a deep clustering approach for sentiment analysis via financial texts based on a transformer-based DEKM and DeepCluster. We deploy and compare several well-known transformers to find the most optimal backbone. We evaluate the proposed approaches using the Financial Phrase Bank (FPB) and Twitter Financial News (TFN) datasets. These datasets contain phrases and sentences related to financial information, primarily sourced from financial reports, news articles, and business documents. Economic experts have annotated the datasets, categorizing sentences into sentiment groups such as positive, negative, or neutral. Moreover, we only use the labels to evaluate the performance of the proposed deep clustering methods. The labels were not used in any training process of deep clustering methods.

The contribution of our work can be summarized as follows:

- Introduced unsupervised learning approaches for NLP tasks, including transformer-based DEKM and DeepCluster.
- Intergrated a transformer-based Autoencoder in DEKM and a transformer-based feature extractor in DeepCluster for deep clustering via text.

2. Theoretical basis

2.1. Sentiment analysis

Sentiment analysis is one of the everyday NLP tasks that aims to classify the polarity of a given text by extracting opinions, attitudes, emotions, and sentiments from the text. Essentially, sentiment analysis is a small branch of the text classification field with three popular classes in most cases, including positive, negative, and neutral. Sentiment analysis has broad applications across various domains, such as marketing, customer service, and finance, particularly in assessing market sentiment from financial reports, new articles, and social media sources.

With the development of deep learning, recent sentiment analysis methods are primarily based on the transformer, specifically the encoder of the transformer. For instance, BERT is one of the most well-known transformers designed to tackle most NLP tasks, such as machine translation, text classification, and part-of-speech tagging (Devlin et al., 2018). RoBERTa is also a robust NLP transformer (Liu et al., 2019). It is an optimal version of BERT with several training data sizes 10 times that of BERT. With the massive training data and hyperparameter

optimization, RoBERTa claimed to be outperformed BERT on all well-known benchmarks in the sentiment analysis field. LlamBERT is another upgraded version of BERT, which aims for low-cost data annotation, providing greater cost-effectiveness but preserving the outperformed accuracy like BERT and RoBERTa (Csanády et al., 2024). T5 is also a noteworthy transformer model. Raffel et al. (2020) presented that the T5 model has an encoder-decoder-based architecture similar to the original transformer model (Vaswani et al., 2017). This architecture makes T5 highly versatile and capable of handling various tasks, including translation, summarization, and question-answering. The pre-trained data for T5 is also much larger than that of BERT. The encoder of T5 can be used similarly to BERT and RoBERTa for solving text sentiment analysis tasks. Furthermore, experimental results of T5 on the SST-2 dataset (a well-known dataset in the sentiment analysis field) (Socher et al., 2013) have shown superior performance compared to both BERT and RoBERTa. However, since T5 requires a lot of training data to optimize its performance, we will not use this model for our experiments.

However, these supervised approaches require a massive number of labeled data, especially since the T5 model requires a substantial amount of training data and available labels. Moreover, the results of most approaches are based on the computation and prediction of deep learning models, so the similarities between texts of the same class cannot be explained and visualized. On the other hand, deep clustering, specifically K-Means clustering, can categorize unsupervised data and visualize it, observing the commonality between data with the same class.

2.2. Deep clustering

Clustering is one of the most well-known unsupervised data analysis methods. The main goal of clustering is to ensure that data points within the same cluster are more similar than those in other clusters. Unlike classification, clustering does not require labeled data, making it helpful in exploring wild, unconstrained data, such as financial texts. To leverage that advantage, researchers gradually optimize clustering algorithms. Eventually, deep clustering started to gain increasing attention. However, recent deep clustering methods have primarily been developed for computer vision tasks. Among the available deep clustering methods, we find DEKM and DeepCluster are the two approaches that are suitable for text data.

DEKM and DeepCluster have two different approaches to unsupervised learning. DEKM utilizes an Autoencoder to embed the input data into a lower-dimensional space. K-Means then cluster the embedded representations to predict the labels. DEKM then applied an orthonormal matrix to transform the representation and optimize the clustering performance. DeepCluster leverages a pre-trained feature extractor to extract the crucial elements of the input data; K-Means then cluster extracted features to predict the pseudo-labels. The authors of DeepCluster subsequently train the classification model with these pseudo-labels. After several iterations of the training process, the authors expect the pseudo-labels of each iteration to be the same with a negligible change, which means that the feature extractor is learning the input data pattern.

Both methods have their unique advantages in deep clustering. However, these methods were first built for computer vision tasks. Therefore, we replaced the original autoencoder and feature extractor with NLP transformers to make these methods suitable for our task.

3. Data and research methods

3.1. Transformer-based autoencoder

An autoencoder is an encoder-decoder-based neural network usually utilized for unsupervised tasks. The architecture of an autoencoder is similar to that of a transformer, where

the encoder extracts the feature vector from input data. At the same time, the decoder tries to reconstruct the input data from the output of the encoder. Therefore, the implementation of a transformer-based Autoencoder is easy to deploy. For the encoder, we apply several pre-trained encoder-based transformers, such as BERT, Roberta, and FinBERT (Araci, 2019). We utilize the same model as the encoder, but with decoder mode turned on for the decoder.

At first, the tokenized input sequence is fed into the encoder to extract the low-dimensional representations. We decided to use the global average of the last hidden state of the encoder instead of the class token (Beyer et al., 2022). The decoder generates a new sequence based on the output of the encoder. The target sequence is set the same as the input sequence, so the generated sequence is expected to be as close as possible to the input sequence. The autoencoder is trained to minimize the reconstruction loss. Therefore, we use Mean Squared Error (MSE) to calculate the loss between the generated and target sequences. The loss function of training the Autoencoder is formulated as follows:

$$\min L = \sum_{i=1}^{n} \left\| x_i - g\left(f(x_i)\right) \right\|^2 \quad (1)$$

Where x_i is the i-th input sequence, $f(.)$ is the encoder, $g(.)$ is the decoder.

As the embedded representations need a low dimensional space compared to the input data, we project the representation to the dimension of 64. By training the model with the loss function (1), the encoder of the Autoencoder can learn how to extract the most crucial features from the input data. The decoder then uses these features to reconstruct the original data, eliminating the need for labeled data throughout the learning process. After optimizing the Autoencoder, we use the encoder of the Autoencoder to extract the representations from the input data and then deploy DEKM to cluster the obtained representations. It is important to note that our transformer-based Autoencoder integrates exclusively with the DEKM method and not with DeepCluster. Unlike DEKM, DeepCluster does not leverage an autoencoder to extract features from the input data but uses another approach: the feature extractor. We will discuss this issue further in Section 3.3.

3.2. Deep-embedded K-Means clustering

Deep Embedded K-Means Clustering (DEKM) was proposed by Guo et al. (2021). This is a deep clustering method that alternately optimizes representation learning and clustering. The process has three main steps: (1) generate an embedding space using Autoencoder, (2) detect clusters in embedding space using K-Means clustering, and (3) optimize the representation to increase cluster-structure information.

In the first step, DEKM applies Autoencoder to learn low-dimensional representations of the input data. After training, the Autoencoder extracts the low-dimensional representations extracted from the input data using only the encoder. K-Means is then applied to cluster the extracted representations and predict each representation's label using the clusters' centroid. Subsequently, DEKM optimizes the representations by transforming the embedding space using an orthonormal matrix, which reveals the cluster structure. Optimizing makes the data points closer to their cluster centroid, but only in the last dimension of the transformed space. This greedily reduces entropy and increases the cluster-structure information.

Experimental results show that DEKM outperformed preceding deep clustering methods, including K-Means, DEC (Xie et al., 2016), DCEC (Guo, Liu, et al., 2017), IDEC (Guo, Gao, et al., 2017), DCN (Yang et al., 2017), DKM (Fard et al., 2020) on various datasets. Inspired by this work, we proposed applying DEKM to our task.

3.3. DeepCluster

DeepCluster was presented by Caron et al. (2018) in 2019. It is an unsupervised learning approach for training convolutional neural networks on large-scale image datasets without labels. The main idea of DeepCluster is to iteratively cluster the features produced by the deep learning model and use the resulting cluster assignments as pseudo-labels to update the model's parameters.

Specifically, DeepCluster deploys a deep learning model as a feature extractor for input data. K-Means then cluster the extracted features to compute cluster centroids. Subsequently, the authors use the cluster assignments as pseudo-labels to train and optimize the parameters of the feature extractor.

DeepCluster allows the feature extractor to progressively learn more discriminative features without requiring manual annotations. For the performance, DeepCluster outperforms preceding unsupervised methods on standard benchmarks, demonstrating its effectiveness for learning general-purpose visual representations in a fully unsupervised way. However, the limitation of DeepCluster is that it requires a large-scale dataset, such as ImageNet or COCO. Moreover, the original version of DeepCluster uses AlexNet or VGG as a feature extractor, an image processing method, so we replaced them with an NLP transformer suitable for our task.

3.4. Data preprocessing

In this paper, we apply the Financial Phrase Bank (FPB) and Twitter Financial News (TFN) datasets to evaluate the proposed methods. The FPB dataset contains 4,846 sentences from financial news articles (Malo et al., 2013). The dataset sentences are labeled for sentiment analysis tasks in neutral, positive, and hostile categories. We divided this dataset into three subsets: train set with 70% sentences, validation set with 15% sentences, and test set with 15% sentences of the dataset.

We also utilize the TFN dataset to evaluate the proposed methods. The TFN dataset consists of 11,931 texts collected from Twitter social media related to the financial domain, which have been labeled for sentiment analysis by experts, including the categories bearish, bullish, and neutral. The original dataset was pre-split by its creators into two subsets: a training set with 80% of the data and a test set with 20% of the data. For the convenience of training and monitoring the performance of the proposed methods, we further divide the training set into two smaller subsets: a training set with 75% of the original training data and a validation set containing the remaining data.

Preprocessing has a crucial role in extracting features from text data. Preprocessing not only cleans the text and denoises it but also helps remove redundant information that is not important to the main context of the input sequence. In this paper, we apply several preprocessing techniques, including normalization, noise reduction, and tokenization. The length of all texts is truncated to 128 for uniformity and convenience for training.

4. Result and discussion

4.1. Results

For DEKM, we first train the transformer-based Autoencoder with the loss function (1) in Section 3.1 to optimize the low-dimensional representations. The training process is executed throughout 20 epochs. Subsequently, we use the encoder of the pre-trained Autoencoder to extract the representations for clustering. The cluster optimization is done after 20 epochs. For DeepCluster, we set BERT as the feature extractor and then use K-Means to predict the pseudo-labels. The training process of DeepCluster is performed with 200 epochs.

We use an AdamW optimizer with a learning rate 2e-5 for the experiment setup. The hyperparameters are chosen after several experiments and based on (Choe et al., 2023). The number of clusters in K-Means is set to 3, corresponding to the number of categories commonly seen in sentiment analysis tasks. All experiments are implemented on a computer with 256GB RAM and Nvidia RTX A5000 24GB GPU. Tables 1 and 2 show the experimental results on the FPB and TFN datasets.

Table 1

Experimental Results of several Deep Clustering Methods on FPB Dataset (%)

Methods	Backbone	Validation	Test
Supervised	BERT	81.82	82.78
DeepCluster	BERT	**59.92**	**57.71**
DEKM	BERT	52.86	52.62
DEKM	FinBERT	51.84	53.58
DEKM	RoBERTa	46.00	53.86

Source. Data analysis result of the research

Table 2

Experimental Results of several Deep Clustering Methods on TFN Dataset (%)

Methods	Backbone	Validation	Test
Supervised	BERT	77.01	77.09
DeepCluster	BERT	**63.82**	**65.58**
DEKM	BERT	44.50	43.93
DEKM	FinBERT	44.29	44.64
DEKM	RoBERTa	60.19	59.88

Source. Data analysis result of the research

Table 3

Comparison of DEKM Performances on TFN Dataset with and without Data Preprocessing (%)

Methods	Backbone	Preprocessed		Non-preprocessed	
		Validation	Test	Validation	Test
DEKM	BERT	44.50	43.93	41.54	42.71
DEKM	FinBERT	44.29	44.64	43.15	41.46
DEKM	RoBERTa	60.19	59.88	55.99	57.62

Source. Data analysis result of the research

4.2. Discussion

Based on Table 1, on the FPB dataset, the DeepCluster algorithm combined with the BERT model yields the best results among the deep clustering methods, with an accuracy of 59.92% on the validation set and 57.71% on the test set. However, these results are easily surpassed when compared to the performance of the BERT model under supervised training

(with an accuracy of 82.78% on the FPB test set). The primary reason for this is that the supervised training process takes advantage of the ground truth labels in the data, allowing for better optimization of the model parameters. This is essentially the most effective approach for classification tasks. On the other hand, as mentioned in Section 1, supervised training requires pre-labeling the data, which is both time-consuming and costly. The training processes of the DeepCluster and DEKM methods in our experiments are entirely independent of any assistance from the ground truth labels of the data, as they only perform unsupervised learning based on the internal information within the data. Therefore, a massive training dataset is required for these unsupervised learning methods to operate at their optimal potential (at least one million data points, similar to the ImageNet dataset used in the original DeepCluster). This requirement is also emphasized by the authors of the DeepCluster algorithm in their paper. Due to the limited number of samples in the FPB dataset (only around 4,000 data points), it cannot meet the requirements of deep clustering methods, leading to suboptimal accuracy in the experimental models.

We supplemented the evaluation with the TFN dataset to demonstrate that our proposed methods would perform better with larger datasets. The TFN dataset contains twice as many samples as the FPB dataset, and since it is also a financial text dataset, it is highly suitable for the sentiment analysis task. According to Table 2, it is evident that the performance of the DeepCluster algorithm combined with the BERT model has significantly improved due to the larger data size of the TFN dataset. Specifically, this method achieved an accuracy of 63.82% on the validation set and 65.58% on the test set of the TFN dataset. With these results, we believe that the accuracy of the DeepCluster method could be further optimized if trained with more data without requiring ground truth labels. This performance demonstrates the potential of deep clustering methods in addressing sentiment analysis tasks specifically and other functions in Natural Language Processing (NLP). Additionally, while the DEKM algorithm presents a novel and theoretically convincing approach, it did not perform well in practice. Specifically, when applying DEKM with the RoBERTa model, the highest accuracy obtained was 53.86% on the FPB test set and 59.88% on the TFN test set.

Moreover, we conducted a comparative analysis to highlight the importance of data preprocessing. Unlike the FPB dataset, the TFN dataset contains more complex texts and noisy information. After applying preprocessing to the TFN dataset, the performance of the proposed methods improved, with accuracy increasing by approximately 3% to 5% (Table 3).

In summary, based on the experimental results of the proposed methods, we believe that deep clustering holds promising potential for text sentiment analysis tasks and Natural Language Processing (NLP) in general. Specifically, our customized DeepCluster approach, utilizing a transformer-based feature extractor, demonstrates this potential. With this method, text classification no longer needs to rely on human effort but can fully leverage the computational strengths of machines. Compared to traditional supervised deep learning methods, DeepCluster does not require labeled data for performance optimization; it simply benefits from learning from as much data as possible, reducing the costs associated with resources, labor, and time required for manual labeling. Given the current abundance of text data, particularly in the financial domain, vast amounts of open data are available online, which can be automatically collected to train more optimized deep clustering models.

5. Conclusions & recommendations

In this paper, we propose an unsupervised approach for sentiment analysis via financial text. Despite not requiring labeled data for training, deep clustering methods can still maintain promising performance comparable to traditional supervised deep learning approaches. The key advantage here is that deep clustering methods do not need labeled data to optimize the model's parameters, which significantly reduces the costs, labor, and time associated with manual labeling. DeepCluster and DEKM are two representative algorithms in the field of deep clustering. Initially designed for computer vision tasks, we optimized these methods for sentiment analysis and natural language processing by integrating them with powerful transformer algorithms. Experimental results show that DeepCluster, when combined with the BERT model, achieves the highest accuracy among the deep clustering methods. Although this result does not surpass traditional supervised deep learning methods, we have demonstrated that it can be significantly improved by increasing the training data size, as evidenced by our evaluation of DeepCluster on the FPB and TFN datasets.

In summary, DeepCluster and other deep clustering methods hold great potential for addressing natural language processing tasks where labeled data is scarce or nonexistent. In the future, we plan to collect more data to augment further the training set for DeepCluster and improve its accuracy. Additionally, we will experiment with two enhanced versions of DeepCluster, namely DeeperCluster and DeepCluster v2 (SwAV) (Caron et al., 2020), to compare and evaluate their performance and find the most optimal algorithm.

NO CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflict of interest.

References

Araci, D. (2019). *FinBERT: Financial sentiment analysis with pre-trained language models*. https://doi.org/10.48550/arxiv.1908.10063

Beyer, L., Zhai, X., & Kolesnikov, A. (2022). *Better plain ViT baselines for ImageNet-1k*. https://doi.org/10.48550/arxiv.2205.01580

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Lecture notes in computer science* (pp. 139-156). Springer. https://doi.org/10.1007/978-3-030-01264-9_9

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). *Unsupervised learning of visual features by contrasting cluster assignments*. https://doi.org/10.48550/arxiv.2006.09882

Choe, J., Noh, K., Kim, N., Ahn, S., & Jung, W. (2023). *Exploring the impact of corpus diversity on financial pretrained language models.* https://doi.org/10.48550/arxiv.2310.13312

Csanády, B., Muzsai, L., Vedres, P., Nádasdy, Z., & Lukács, A. (2024). *LlamBERT: Large-scale low-cost data annotation in NLP*. https://doi.org/10.48550/arxiv.2403.15938

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* https://doi.org/10.48550/arxiv. 1810.04805

Fard, M. M., Thonet, T., & Gaussier, E. (2020). Deep k-Means: Jointly clustering with k-Means and learning representations. *Pattern Recognition Letters, 138*, 185-192. https://doi.org/10.1016/j.patrec.2020.07.028

Guo, W., Lin, K., & Ye, W. (2021). Deep embedded K-Means clustering. *2021 International Conference on Data Mining Workshops (ICDMW)*. https://doi.org/10.1109/icdmw53433 .2021.00090

Guo, X., Gao, L., Liu, X., & Yin, J. (2017). *Improved deep embedded clustering with local structure preservation*. https://doi.org/10.24963/ijcai.2017/243

Guo, X., Liu, X., Zhu, E., & Yin, J. (2017). Deep clustering with convolutional autoencoders. In *Lecture notes in computer science* (pp. 373-382). https://doi.org/10.1007/978-3-319-70096-0_39

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *ROBERTA: A robustly optimized BERT pretraining approach.* https://doi.org/10.48550/arxiv.1907.11692

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782-796. https://doi.org/10.1002/asi.23062

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1-67.

Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., Yu, P. S., & He, L. (2024). Deep clustering: A Comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21. https://doi.org/10.1109/tnnls.2024.3403155

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631-1642). Association for Computational Linguistics. https://aclanthology.org/D13-1170

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. https://doi.org/10.48550/arxiv.1706 .03762

Xie, J., Girshick, R., & Farhadi, A. (2016). *Unsupervised deep embedding for clustering analysis*. https://proceedings.mlr.press/v48/xieb16.html

Yang, B., Fu, X., Sidiropoulos, N. D., & Hong, M. (2017). *Towards K-means-friendly spaces: Simultaneous deep learning and clustering*. https://proceedings.mlr.press/v70/yang 17b.html