# CRL+: A Novel Semi-Supervised Deep Active Contrastive Representation Learning-Based Text Classification Model for Insurance Data

Amir Namavar Jahromi [1], Ebrahim Pourjafari [2,*], Hadis Karimipour [1], Amit Satpathy [2], and Lovell Hodge [2]

[1] Schulich School of Engineering, University of Calgary, Alberta, Canada;
Email: amir.namavarjahromi@ucalgary.ca (A.N.J.), hadis.karimipour@ucalgary.ca (H.K.)
[2] Munich Re Canada, Canada; Email: ASatpathy@munichre.ca (A.S.), LHodge@munichre.ca (L.H.)
*Correspondence: EPourjafari@munichre.ca (E.P.)

***Abstract*—Financial sector and especially the insurance industry collect vast volumes of text on a daily basis and through multiple channels (their agents, customer care centers, emails, social networks, and web in general). The information collected includes policies, expert and health reports, claims and complaints, results of surveys, and relevant social media posts. It is difficult to effectively extract labels, classify, and interpret the essential information from such varied and unstructured material. Therefore, the Insurance Industry is among the ones that can benefit from applying technologies for the intelligent analysis of free text through Natural Language Processing (NLP). In this paper, CRL+, a novel text classification model combining Contrastive Representation Learning (CRL) and Active Learning is proposed to handle the challenge of using semi-supervised learning for text classification. In this method, supervised (CRL) is used to train a RoBERTa transformer model to encode the textual data into a contrastive representation space and then classify using a classification layer. This CRL-based transformer model is used as the base model in the proposed Active Learning mechanism to classify all the data in an iterative manner. The proposed model is evaluated using unstructured obituary data with objective to determine the cause of the death from the data. This model is compared with the CRL model and an Active Learning model with the RoBERTa base model. The experiment shows that the proposed method can outperform both methods for this specific task.**

***Keywords*—natural language processing, contrastive representation learning, active learning, text classification, transformers, CRL+**

## I. Introduction

Text classification is a classical problem in Natural Language Processing (NLP) that aims to assign a label to textual units like words, sentences, paragraphs, or documents [1]. Text classification has a wide range of applications. It is used in question answering, spam detection, sentiment analysis, news categorization, user intent detection, and many others. NLP applications are becoming more popular daily due to the advances in various computational linguistics and the abundance of training data from websites, personal communications (emails, text messages), social media, tickets, insurance claims, user reviews, and questions/answers from customer services. The insurance industry collects a large amount of textual data as part of their day to day processes. The vast majority of this textual data is read and interpreted by human agents resulting in higher costs and slower processes. This creates an opportunity to automate some text processing tasks and make the insurance processes faster and cheaper. Traditional NLP techniques will assist insurance companies to personalize insurance products and better respond to existing and future clients' needs. More accurate underwriting models and constant learning from new data will be reflected in premiums that are often conservatively overestimated due to a lack of non-traditional underwriting data. Using NLP solutions in the insurance industry yields better customer satisfaction and profit. The use cases include claims classification, optimizing payment processes, monitoring policy changes, personalized product offerings, improved risk assessment, enhanced fraud detection, and business process automation [2].

Documents and textual data are rich sources of information that can be used to solve various problems. However, extracting information from this type of data requires more complex techniques due to the unstructured nature of textual data, which can be time consuming and challenging [1]. Text classification is an important step in text processing. This can be done manually by domain experts or automatically. Recently, due to the increasing number of textual documents, automatic text classification has become popular and important. Automatic text classification approaches are grouped into rule-based and data-driven methods. Rule-based methods categorize text samples using a set of pre-defined rules that requires deep expert knowledge, which is context-based and hard to acquire. On the other hand, data-driven methods use

Manuscript received March 6, 2023; revised April 17, 2023; accepted May 24, 2023; published October 13, 2023.

Machine Learning (ML) approaches to find patterns to classify textual samples [2].

However, most of the ML-based text classifiers are built from labeled training samples. Manual labeling of a large set of training documents is time-consuming. In the past few years, researchers investigated various forms of semi-supervised learning to reduce the burden of manual labeling by using a small labeled set for every class and a large unlabeled set for classifier building. Semi-supervised learning is a hybrid approach that combines supervised and unsupervised learning elements to train models with a small amount of labeled data, and a large amount of unlabeled data [3, 4]. Moreover, word embedding techniques were used to represent the words of a text using representation learning methods in a way that words that have the same meaning have a similar representation [5]. Similar to word embeddings, distributed representations for sentences can also be learned in an unsupervised fashion by optimizing some auxiliary objectives, such as the reconstruction loss of an autoencoder [4]. Such unsupervised learning results in sentence encoders that can map sentences with similar semantic and syntactic properties to similar fixed-size vector representations.

In this paper, a novel text classification model, CRL+, that combines Contrastive Representation Learning (CRL) and Active Learning is proposed to handle the challenge of using semi-supervised textual data for text classification in the insurance industry. In this method, supervised CRL will be used to train a RoBERTa transformer model to encode the textual data into the representational vector and then classify using a classification layer. This CRL-based transformer model will be used as the base model of a modified Active Learning mechanism to classify all the data in an iterative manner. The remainder of this paper is organized as follows: Section II provides a summary of the related works in literature. Section III provides some background about the methods used in this paper. Section IV describes the proposed approach for classifying semi-supervised textual data. Section V explains the experimental setup, including the dataset and evaluation metrics. Section VI shows the experimental results. Finally, Section VII concludes the paper.

## II. Related Works

It is estimated that around 80% of all information is unstructured [6], with text being one of the most common unstructured data types. The unstructured data structure is irregular or incomplete, and there is no predefined data model. Compared to structured data, this data is still difficult to retrieve, analyze and store [7]. This is where text classification with ML comes in. ML-based techniques can automatically classify all manner of relevant text, from legal documents, social media, surveys, and more, in a fast and cost-effective way [8, 9]. Some of the most popular machine learning algorithms for creating text classification models include the Naïve Bayes (NB) [10] family of algorithms, Support Vector Machine (SVM) [11], and Deep Neural Network (DNN).

Recent research shows that it is effective to cast many NLP tasks as text classification by allowing DNN to take a pair of texts as input [12–14]. Compared to traditional ML, DNN algorithms need more training data. However, they do not have a threshold for learning from training data like traditional machine learning algorithms, such as SVM and NB. The two main DNN architectures for text classification are Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). RNNs are trained to recognize patterns across time, whereas CNNs learn to recognize patterns across space [15]. RNNs work well for the NLP tasks such as Question Answering (QA), where the comprehension of long-range semantics is required. In contrast, CNNs work well where detecting local and position-invariant patterns are essential [1, 16]. Thus, RNNs have become one of NLP's most popular model architectures. Long Short-Term Memory (LSTM) is a popular architecture, which addresses the gradient vanishing problems in RNNs by introducing a memory cell to remember values over arbitrary time Intervals [17, 18]. There have been works improving RNNs and LSTM models for NLP applications by capturing richer information, such as tree structures of natural language, long-span word relations in text, and document topics [19, 20]. Character-level CNNs have also been explored for text classification [21, 22]. Studies investigate the impact of word embeddings and CNN architectures on model performance. Conneau and Schwenk *et al.* [23] presented a Very Deep Convolutional Neural Network (VDCNN) model for text processing. It operates directly at the character level and uses only small convolutions and pooling operations. This study shows that the performance of VDCNN improves in deeper models. DNN algorithms, like word2vec [24] or Glove [25], are also used to obtain better vector representations for words and improve the accuracy of classifiers trained with traditional machine learning algorithms. Recently, transformer models were introduced to improve the performance of LSTM models. Unlike LSTM, transformer models can be fully parallelized on GPUs, which makes the training step faster. Moreover, transformers can memorize longer sentences better than the LSTMs due to their attention mechanism. As one of the most popular transformer models, the Bidirectional Encoder Representations from Transformers (BERT) model was introduced by Google in 2019 [12]. BERT model outperforms its predecessors, word2vec and ELMo [26], exceeding state-of-the-art by a margin in multiple natural language understanding tasks. Section III-*E* will discuss the transformers in more detail.

## III. Background

In this chapter, background materials related to this paper will be introduced.

### *A. Semi-Supervised Learning*

Semi-supervised learning is a type of ML in which only a portion of the training samples are labeled. Most of the real-world problems in the fintech and insurance industries are semi-supervised. Therefore, handling this type of problem has recently gained momentum [2, 27]. Active Learning and CRL are two popular solutions for these problems.

### B. Active Learning

Active Learning is a mechanism to label unlabeled data iteratively using an ML model. This mechanism increases the performance of ML models on datasets with limited labeled samples. In this mechanism, first, an ML model is trained using the labeled data. Then, the model is applied to unlabeled data to classify them. Based on the classification result, some samples are selected to pass to an expert person to label them manually and add them to the labeled data. This process is repeated until the model meets predefined performance criterion [27].

### C. Contrastive Representation Learning

CRL is a ML technique used for unsupervised and semi-supervised problems as a pre-training to enhance the performance of ML models. This model is used to train a representation space in which similar sample points stay close, while dissimilar ones are far apart. CRL was first developed as a self-supervised method for unsupervised image classification problems (or as an unsupervised pre-training for supervised problems) [28]. In this method, a self-supervised pre-training was done on the data using augmentation techniques to map the samples from the original space to the contrastive space. Then, Khosa *et al.* [29] proposed a supervised version of CRL and showed its advantages compared to the self-supervised version on image classification problems. In this method, instead of considering the anchor sample and its augmented one as the positive samples, the samples with similar labels are also considered as positives. Moreover, several NLP versions of CRL were proposed that include both self-supervised and supervised versions of these models [30, 31]. In this paper, a supervised CRL model, introduced in [32], is used as the base model to pre-train the RoBERTa model using the Active Learning mechanism.

### D. RNN

RNN is a type of neural network that predicts new situations based on previous ones. RNNs can handle sequential problems like NLP. Several RNN models are proposed in the literature. One of the most popular RNN models is LSTM, which suffers less from the vanishing gradient problem compared to RNN [33].

### E. Transformers

One of the computational bottlenecks of training RNNs on GPUs is the sequential processing of text. Transformers [34] overcome this limitation by applying self-attention to compute an attention score in parallel for every word in a sentence or document an “attention score” to model each word’s influence on another. Due to this feature, transformers allow for much more parallelization than CNNs and RNNs, which makes it possible to efficiently train huge models on large amounts of data on GPUs.

Since 2018, we have seen the rise of a set of largescale transformer-based Pre-trained Language Models (PLMs). Compared to earlier contextualized embedding models based on CNNs [35] or LSTMs [26], transformer-based PLMs use much deeper network architectures (e.g., 48-layer transformers [36]), and are pre-trained on much larger amounts of text corpora to learn contextual text representations by predicting words conditioned on their context. These PLMs are fine-tuned using task specific labels and have created a new state of the art in many downstream NLP tasks, including text classification. Although pre-training is unsupervised (or self-supervised), fine-tuning is supervised learning. A recent survey by Qiu *et al*. [37] categorizes popular PLMs by their representation types, model architectures, pretraining tasks, and downstream tasks.

PLMs can be grouped into two categories, autoregressive and autoencoding PLMs. One of the earliest autoregressive PLMs is OpenGPT [36, 38], a unidirectional model that predicts a text sequence word by word from left to right (or right to left), with each word prediction depending on previous predictions. It consists of 12 layers of transformer blocks, each consisting of a masked multi-head attention module, followed by a layer normalization and a position-wise feed-forward layer. OpenGPT can be adapted to NLP applications such as text classification by adding task-specific linear classifiers and fine-tuning using task-specific labels.

BERT is one of the most widely used autoencoding PLMs [12]. Unlike OpenGPT, which predicts words based on previous predictions, BERT is trained using the Masked Language Modeling (MLM) task that randomly masks some tokens in a text sequence and then independently covers the masked tokens by conditioning on the encoding vectors obtained by a bidirectional transformer. There have been numerous works on improving BERT. RoBERTa [39] is more robust than BERT, mainly because its pre-training method focuses on MLM with changing mask tokens per epoch, and is trained using much more training data. ALBERT [40] lowers the memory consumption of the BERT model and increases its training speed. DistillBERT [41] utilizes knowledge distillation during pre-training to reduce the size of BERT by 40%, while retaining 99% of its original capabilities and making the inference 60% faster. SpanBERT [42] extends BERT to better represent and predict text spans. Electra [43] uses a more sample efficient pre-training task than MLM, called replaced token detection. Instead of masking the input, it corrupts it by replacing some tokens with plausible alternatives sampled from a small generator network. ERNIE [44, 45] incorporates domain knowledge from external knowledge bases, such as named entities, for model pre-training. ALUM [46] introduces an adversarial loss for model pretraining that improves the model’s generalization to new tasks and robustness to adversarial attacks. BERT and its variants have been fine-tuned for various NLP tasks, including QA [47], text classification [13], and Natural Language Inference (NLI) [14, 48].

## IV. Proposed Method

This paper combines supervised CRL with an Active Learning mechanism to enhance the performance of these methods to classify textual data. In this paper, a modified version of SupCL-Seq [32] is used as the CRL model.

SupCL-Seq extends the self-supervised Contrastive Learning for textual data to a supervised setting. In the developed model, several dropouts are used to make augmented samples from the anchor by changing sample embeddings. Then, the anchor sample, its augmented samples, and other samples with the same label in the dataset are used to train their presentation using the contrastive loss function. The representation learning task consists of an encoder part of a transformer (RoBERTa) to make the augmented samples and train the CRL part of the model. Fig. 1 shows the training of the developed model. Then, a classification layer is added to the trained representation learning model to do the final classification. Finally, all the CRL encoder parameters are frozen, and the classification model is trained.

Eq. (1) shows the supervised contrastive loss function.

$$\mathcal{L}_i^{sup} = \sum_{i\in I} \frac{-1}{|p(i)|} \sum_{p\in p(i)} log \frac{e^{cosine(\tilde{x}_i,\tilde{x}_p)/\tau}}{\sum_{b\in B(i)} e^{cosine(\tilde{x}_i,\tilde{x}_b)/\tau}} \quad (1)$$

where $p(i)$ contains positive samples (samples with a similar label to the anchor), $B(i)$ has negative samples (samples with a different label from the anchor), $\tau$ is the scaling term, and *cosine* $(a, b)$ is the similarity function between $a$ and $b$ (see Eq. (2)).

$$cosine(A,B) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2)$$

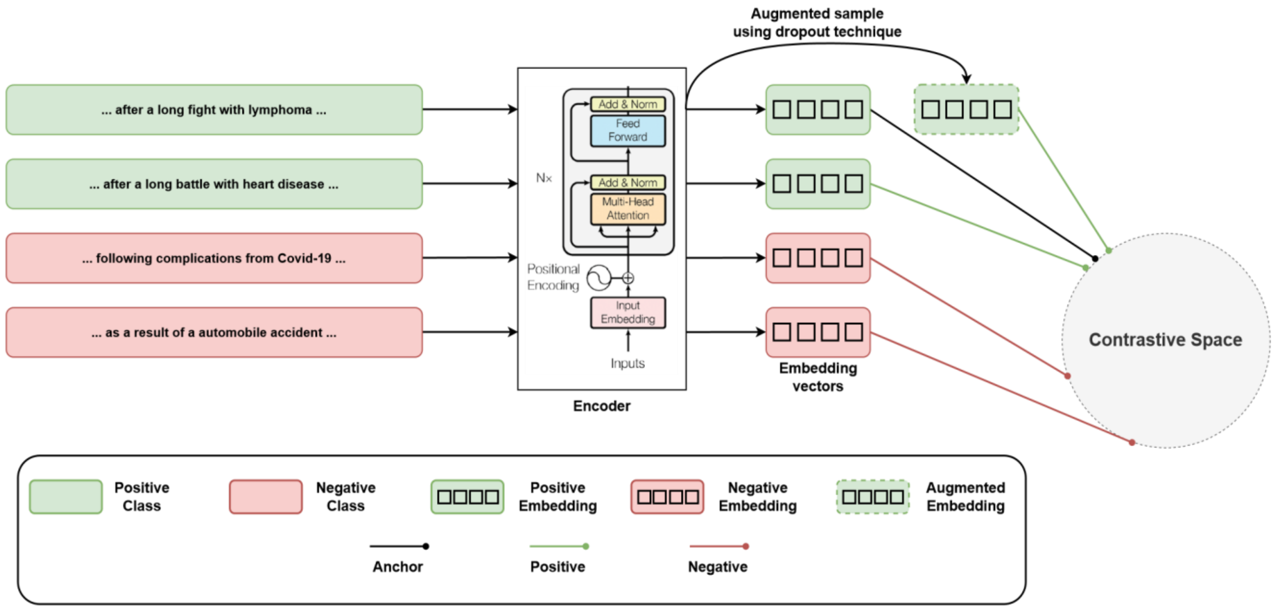


Figure 1. The used Supervised CRL model in this work. Samples with similar labels are considered positive samples, and all other samples are considered negative. The augmentation is done using the dropout technique in the encoder.

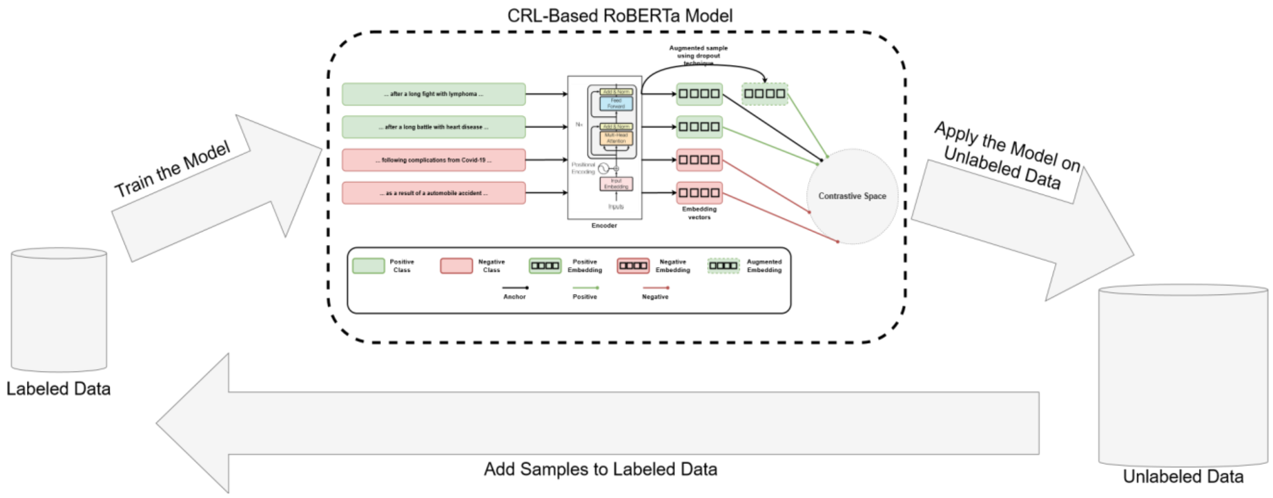


Figure 2. The proposed model consists of the supervised CRL-based RoBERTa model as the base model in the Active Learning mechanism. It starts with a small labeled data and tries to label all the samples in several iterations using the CRL-based RoBERTa base model.

This method is used to pre-train the encoder of the RoBERTa transformer to encode the textual data into a contrastive representation before passing it to the classification layer.

To enhance the performance of this model on semi-supervised applications, this model is used as the base model in the Active Learning mechanism. Fig. 2 shows the proposed model.

To make the proposed method free from human intervention, the Active Learning mechanism is changed in a way to remove the expert person from it. In the proposed method, in each iteration of the Active Learning mechanism, the classified samples with reasonable confidence (based on the SoftMax layer) are added to the labeled samples and used to train the CRL-based RoBERTa for the next iteration. This process continues until the model gets a pre-defined performance or a specific number of iterations is passed.

## V. Experimental Setup

### A. Dataset

To evaluate the proposed method for labeling and classification of the insurance-based textual data, a dataset consisting of obituary texts is used to predict the cause of death. However, among more than 2,500,000 samples, only 3% of them were labeled. Table I shows the number of labeled data for each class in the mentioned dataset.

TABLE I. Number of the Labeled Samples for Each Class Label in the Obituary Dataset

| Class Label | Number of Labeled Samples |
|---|---|
| Neoplasms (Cancers) | 33,104 |
| Circulatory System | 3,477 |
| Accidents | 11,942 |
| Respiratory System | 3,427 |
| Nervous System | 1,991 |
| Suicides | 3 |
| Digestive System | 543 |
| COVID-19 | 5,007 |

### B. Evaluation Metrics

The basic ML metrics are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which represent the number of samples correctly classified as positive, correctly classified as negative, wrongly classified as positive, and wrongly classified as negative, respectively. Using these basic metrics, more complex metrics, including Accuracy, Precision, Recall, and F-measure, are defined and used to quantify the performance of ML algorithms.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$F-measure = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{6}$$

- Accuracy indicates the number of correctly classified samples over the entire dataset (see Eq. (3)).
- Precision indicates the number of samples classified correctly as each class label over total samples classified for that class (see Eq. (4)).
- Recall indicates the number of samples classified as each class label correctly over the total instances of the dataset for that class label (see Eq. (5)).
- F-measure is the harmonic value of precision and recall (see Eq. (6)).

## VI. Results

The proposed combination of the supervised CRL, RoBERTa transformer, and Active Learning is evaluated on an obituary dataset introduced in Section V. Table II shows the performance of the proposed method. This table illustrates that the proposed method can accurately classify textual insurance data for all class labels. Moreover, the proposed method is compared with the RoBERTa transformer with supervised CRL and Active Learning with the RoBERTa base model.

TABLE II. Results of the Proposed Method on the Obituary Dataset. The Overall Row Shows the Performance of the Model over all the Classes. The Other Rows Show the One-Versus-All Version of the Proposed Method for Each Class Label

| Class Label | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Overall | 95.68 | 95.25 | 95.68 | 95.46 |
| Neoplasms (Cancers) | 96.72 | 98.50 | 96.72 | 97.60 |
| Covid-19 | 99.00 | 92.25 | 96.30 | 94.23 |
| Circulatory System | 99.11 | 93.28 | 91.20 | 92.23 |
| Accidents | 98.26 | 96.35 | 95.00 | 95.67 |
| Respiratory System | 96.17 | 59.44 | 87.70 | 66.95 |
| Nervous System | 99.70 | 95.59 | 95.20 | 95.57 |
| Suicides | 100 | 100 | 100 | 100 |
| Digestive System | 99.20 | 58.65 | 71.60 | 64.50 |

As illustrated in Table III, the proposed method outperformed both these models. This table shows that the supervised CRL model significantly outperformed the Active Learning model. However, adding Active Learning to this model empowers it to detect the cause of death more accurately.

TABLE III. Comparing the Performance of the Proposed Method (CRL+) with CRL and Active Learning Models

| Model | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| CRL+ | 95.68 | 95.25 | 95.68 | 95.46 |
| CRL | 92.77 | 92.62 | 92.77 | 92.69 |
| Active Learning | 75.28 | 90.78 | 75.28 | 82.31 |

## VII. Conclusion

Insurance companies gathered enormous amounts of textual data through different channels. This information can help insurance companies to perform highly complex advanced analytics using data mining and machine learning. In this paper, CRL+, a semi-supervised active contrastive representation learning model is proposed to map the semi-supervised data into a contrastive space. This learned representation can be used for classification or regression models and also sequence-to-sequence applications. The proposed method is evaluated using an obituary dataset to classify the cause of death based on the document's content and compared with a modified Active Learning and contrastive representation learning methods using RoBERTa base models. The experiments show that the proposed method outperforms the others in all metrics. The proposed method could be improved by using a two-step pre-processing and adding self-supervised contrastive learning. Moreover, the Active Learning part of the algorithm could be modified to make it more efficient.

## CONFLICT OF INTEREST

The authors declare no conflicts of interest that could potentially influence the results and findings reported in this paper. All authors affirm that their professional affiliations have not resulted in any bias on the conclusion of this study.

## AUTHOR CONTRIBUTIONS

Amir Namvar Jahromi is the main contributor to the paper. Amir Namvar Jahromi and Ebrahim Pourjafari conducted the research and composed the paper, Hadis Karimipour and Lovell Hodge supervised the research and revised the paper, Amit Satpathy prepared the training dataset, supervised the research and revised the paper; all authors had approved the final version.

## REFERENCES


[1] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1– 40, 2021.

[2] A. Ly, B. Uthayasooriyar, and T. Wang, "A survey on Natural Language Processing (NLP) and applications in insurance," arXiv preprint, arXiv: 2010.00462, 2020.

[3] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," Tech. Rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[5] Z. Yin and Y. Shen, "On the dimensionality of word embedding," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[6] Deep Talk. (2021). 80% of the world's data is unstructured. [Online]. Available: https://deep-talk.medium.com/80-of-theworlds-data-is-unstructured-7278e2ba6b73

[7] S. Gupta and S. K. Gupta, "Natural language processing in mining unstructured data from software repositories: A review," *Sādhanā*, vol. 44, no. 12, pp. 1–17, 2019.

[8] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single layer multi size filters convolutional neural network," *IEEE Access*, vol. 8, pp. 42689–42707, 2020.

[9] H. Liu, P. Burnap, W. Alorainy, and M. L. Williams, "A fuzzy approach to text classification with two-stage training for ambiguous instances," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 227–240, 2019.

[10] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *Proc. AAAI-98 Workshop on Learning for Text Categorization*, 1998, pp. 41–48.

[11] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2001.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, arXiv: 1810.04805, 2018.

[13] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," arXiv preprint, arXiv: 1804.07461, 2018.

[14] X. Liu, P. He, W. Chen, and J. Gao, "Multitask deep neural networks for natural language understanding," arXiv preprint, arXiv: 1901.11504, 2019.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," arXiv preprint, arXiv: 1404.2188, 2014.

[17] Y. Liu, C. Sun, L. Lin, and X. Wang, "Learning natural language inference using bidirectional lstm model and inner-attention," arXiv preprint, arXiv: 1605.09090, 2016.

[18] P. Liu, X. Qiu, and X. Huang, "Modelling interaction of sentence pair with coupled-lstms," arXiv preprint, arXiv: 1605.05573, 2016.

[19] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," arXiv preprint, arXiv: 1503.00075, 2015.

[20] X. Zhu, P. Sobihani, and H. Guo, "Long short-term memory over recursive structures," in *Proc. the 32nd International Conference on Machine Learning*, 2015, pp. 1604–1612.

[21] X. Zhang, J. Zhao, and Y. LeCun, "Character level convolutional networks for text classification," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[22] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proc. the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[23] A. Conneau, H. Schwenk, L. Barrault, and Y. LeCun, "Very deep convolutional networks for text classification," arXiv preprint, arXiv: 1606.01781, 2016.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint, arXiv: 1301.3781, 2013.

[25] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[26] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 2227–2237.

[27] R. Elie, C. Hillairet, F. Hu, and M. Juillard, "An overview of active learning methods for insurance with fairness appreciation," arXiv preprint, arXiv: 2112.09466, 2021.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. International Conference on Machine Learning (PMLR 2020)*, 2020, pp. 1597–1607.

[29] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020.

[30] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "Declutr: Deep contrastive learning for unsupervised textual representations," arXiv preprint, arXiv: 2006.03659, 2020.

[31] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," arXiv preprint, arXiv: 2104.08821, 2021.

[32] H. Sedghamiz, S. Raval, E. Santus, T. Alhanai, and M. Ghassemi, "Supcl-seq: Supervised contrastive learning for downstream optimized sequence representations," arXiv preprint, arXiv: 2109.07424, 2021.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[35] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[37] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[38] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.* (2018). Improving language understanding by generative pre-training. Preprint. [Online]. Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," arXiv preprint, arXiv: 1907.11692, 2019.

[40] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint, arXiv: 1909.11942, 2019.

[41] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint, arXiv: 1910.01108, 2019.

[42] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pretraining by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[43] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," arXiv preprint, arXiv: 2003.10555, 2020.

[44] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "Ernie: Enhanced representation through knowledge integration," arXiv preprint, arXiv:1904.09223, 2019.

[45] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," in *Proc. the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 5, pp. 8968–8975.

[46] X. Liu, H. Cheng, P. He, W. Chen, Y. Wang, H. Poon, and J. Gao, "Adversarial training for large neural language models," arXiv preprint, arXiv: 2004.08994, 2020.

[47] S. Garg, T. Vu, and A. Moschitti, "Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection," in *Proc. the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 5, pp. 7780–7788.

[48] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware BERT for language understanding," in *Proc. the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 5, pp. 9628–9635.