SuperConText: Supervised Contrastive Learning Framework for Textual representations
Youness Moukafih, Nada Sbihi, Mounir Ghogho, Kamel Smaïli

To cite this version:
Youness Moukafih, Nada Sbihi, Mounir Ghogho, Kamel Smaïli. SuperConText: Supervised Contrastive Learning Framework for Textual representations. IEEE Access, In press, 10.1109/ACCESS.2023.3241490. hal-03964804

HAL Id: hal-03964804
https://hal.science/hal-03964804
Submitted on 31 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
SuperConText: Supervised Contrastive Learning Framework for Textual representations

YOUNESS MOUKAFIH¹,², NADA SBIHI¹, MOUNIR GHOGHO¹, (FELLOW, IEEE), KAMEL SMAILİ²
¹TICLab, College of Engineering and Architecture, Université Internationale de Rabat, Morocco
²Loria, Campus Scientifique, Vandoeuvre Lés-Nancy, France
Corresponding author: First A. Author (e-mail: youness.moukafih@univ-lorraine.fr).

ABSTRACT In the last decade, Deep neural networks (DNNs) have been proven to outperform conventional machine learning models in supervised learning tasks. Most of these models are typically optimized by minimizing the well-known Cross-Entropy objective function. The latter, however, has a number of drawbacks, including poor margins and instability. Taking inspiration from the recent self-supervised Contrastive representation learning approaches, we introduce Supervised Contrasitve learning framework for Textual representations (SuperConText) to address those issues. We pretrain a neural network by minimizing a novel fully-supervised contrastive loss. The goal is to increase both inter-class separability and intra-class compactness of the embeddings in the latent space. Examples belonging to the same class are regarded as positive pairs, while examples belonging to different classes are considered negatives. Further, we propose a simple yet effective method for selecting hard negatives during the training phase. In extensive series of experiments, we study the impact of a number of parameters on the quality of the learned representations (e.g. the batch size). Simulation results show that the proposed solution outperforms several competing approaches on various large-scale text classification benchmarks without requiring specialized architectures, data augmentations, memory banks, or additional unsupervised data. For instance, we achieved top-1 accuracy of 61.94% on the Amazon-F dataset, which is 3.54% above the best result obtained when using the cross-entropy with the same model architecture.

INDEX TERMS Deep learning; Contrastive Learning; Text Classification; Hard Negative Examples

I. INTRODUCTION

Over the past few years, deep neural networks (DNNs) have achieved state-of-the-art results surpassing conventional machine learning algorithms in a variety of applications across many disciplines [1] [3] [4]. The success of deep learning is usually attributed to their ability to automatically learn multiple levels of representations in an end-to-end manner. Most of these models are usually optimized using the well-known Cross-Entropy objective function. Indeed, the concept of cross-entropy is straightforward and intuitive: every class is assigned a vector of a target (usually 1-hot). Despite its popularity, however, the cross-entropy – the KL-divergence between one-hot vectors of labels and the distribution of the model’s output logits – suffers from major robustness issues. For example, training a deep neural network by the cross entropy loss is vulnerable to adversarial attacks [2]. Several works have demonstrated, theoretically, that training with cross-entropy loss can cause the representations to spread sparsely over the representation space during training [16]. Additionally, introducing noisy data seems to reduce the performance substantially, due to the fact that the loss considers that all the training labels are true, and neglects the fuzziness of noisy labels [5]. To overcome these issues, many successful alternatives were proposed to address the reference label distribution problems through label smoothing [6], [7], Mixup [9], and knowledge distillation [10]. Recently, contrastive representation learning, which was shown to be related to the estimation of mutual information, has led to major advances in self-supervised learning. Contrastive learning was also shown to achieve state-of-the-art performance on many large-scale benchmark datasets.

Contrastive learning is a particular form of a Siamese neural network, which consists of two or more identical sub-networks, each producing a vector representation of its re-
spective input. Modern Siamese network approaches use data augmentation techniques to generate an augmented view of the same data instance across each sub-network and pull together the learned representations in the embedding space. Maximizing the similarities between each pair of augmented views can result in a trivial solution where all representations are equal to other; this is referred to as collapsing problem. Several approaches have proposed to solve the collapsing problem, one of which is contrastive learning. This latter prevents the undesirable trivial solution by contrasting between positive (similar) examples and many negative (dissimilar) examples. These methods explicitly aim at training a neural network to learn embeddings by pulling together the representations of augmented views, while pushing away the representations of augmented views of different data instances (negative examples), often by using noise-contrastive estimation. The most common strategy is to uniformly sample from the training dataset using examples either from the current batch or from a memory bank. However, it has been observed empirically that contrastive learning methods still suffer from dimensional collapse [38].

In this paper, we propose a supervised contrastive learning framework for text multi-class classification tasks. Inspired by the recent success of joint embedding approaches for learning representations in a self-supervised setting [11] [12], we develop a framework that learns sentence embeddings by maximizing the agreement between the representations of a cluster of instances belonging to the same class using a novel fully-supervised contrastive loss that guides a neural network to better separate the classes. We address the problem of text-multi-class classification applications. We consider many positives per anchor, unlike previous works on self-supervised contrastive learning which use only a single positive example and many negatives. In other words, instead of using augmented views of the same anchor as done in self-supervised contrastive learning (which is not obvious in textual data), we leverage label information to consider many positives and many negatives for each anchor. In figure 1, we show how we select positive/negative examples for each class. The use of many positives and many negatives in our framework allows the encoder function to better maximize the intra-class compactness and the inter-class separability (learns useful and generalizable features) than the standard framework which relies on the cross-entropy loss. Figure 2 shows tSNE plots of the learned representations of the model trained on SST-2 dataset with our loss against those learnt by cross entropy. The increased intra-class compactness and inter-class separability naturally lead to a better text classifier in the fine-tuning stage.

A number of studies have demonstrated that hard negatives (i.e. ones that are hard to distinguish from positives) are important for learning more powerful representations in contrastive learning. Therefore, a number of works have proposed novel sampling strategies. In [13], the authors use hard negative mixing to synthesize new examples from the available hard negatives, while in [14], the authors sample...
negatives from a ring around each positive (i.e. negatives that are neither too close nor too far from the positive example). In this paper, we propose a novel tunable feature-based sampling strategy for selecting hard negative examples based on the similarities resulting from the pretrained model, and show that our approach improves the performance of the learned representations on downstream classification tasks.

Further, several research studies have shown that the number of negative examples is crucial for learning high-quality representation in self-supervised contrastive learning. Accordingly, recent contrastive learning approaches use large batch sizes, or keep large memory banks. For instance, in [25], the authors proposed Momentum Contrast (MoCo) that uses a queue with features of the last few batches, while in [31], a memory of the whole training data is utilized. It is however shown that increasing the memory/batch size does not always give better results. In this work, we conduct extensive experiments to analyse and investigate whether the conclusions that have been drawn from the most successful approaches for self-supervised contrastive learning are still valid when applied to textual data in a supervised contrastive learning setting.

The experimental results show that our Framework consistently outperforms the commonly-used supervised learning framework based on cross-entropy loss on many publicly available benchmark datasets. For instance, we achieve an accuracy of 61.94% and 95.45% using our framework on Amazon-F and Yelp-P respectively, while the scores obtained by the cross-entropy are 58.40% and 92.12% using the same neural architecture. The experimental results show that our proposed method benefits from large positive examples. However, we find that beyond a certain threshold, increasing the number of positive examples does not improve the quality of the representations. In contrast, the method produces high-quality representations when the number of negative instances is large. In addition, simulations show that the sampling strategy is crucial for learning more generalizable sentence representations for downstream tasks on several benchmark datasets. Further experiments on Moroccan and Algerian dialects demonstrate that our method also works well for low-resource languages.

We summarize our contributions as follows:

- We propose a novel supervised learning framework for pre-training text representation by leveraging the contrastive learning paradigm.
- We compare the proposed approach with the standard cross-entropy loss-based method using several large-scale text classification datasets.
- We empirically study the effect of the number of positive/negative examples on the performance of the proposed supervised contrastive learning method.
- We propose a novel tunable feature-based sampling strategy for selecting hard negative examples that further improves performance.

II. RELATED WORK

A. CROSS ENTROPY LOSS

Cross Entropy (CE) is the de facto choice for the loss function in classification tasks. This prominence is due to many reasons. First, CE has good theoretical grounding in information theory, which makes it useful for theoretical analysis of systems [15]. Second, CE loss has been proven to rival many loss functions in large datasets [45]. However, it suffers from major robustness issues. Indeed, CE suffers from adversarial robustness, as was shown in [16] which demonstrated empirically that training with a CE loss can cause the representations to spread sparsely over the representation space during training. Additionally, introducing noisy data seems to degrade performance substantially [19] which is due to the fact that the cross entropy loss supposes that all the training labels are true, and neglects the fuzziness of noisy labels [17], [19]. Classification models are theoretically evaluated by their ability to separate classes in the representation space.
Separability is also of practical use since large margins can make models robust to small perturbations of the input space, and hence more robust to noise. In [37], [42], the authors showed that CE does not maximize the separating margins between classes, and proposed an alternative that solves this problem. This phenomenon can be attributed to the leniency of the penalties of the cross entropy when close to the ground truth label (i.e. CE is eager for the model to be right), and can lead to poor generalization.

**B. SELF-SUPERVISED REPRESENTATION LEARNING**

One of the most prominent lines of research that ventures out of unsupervised representation learning is Self-Supervised Learning (SSL). This paradigm uses pretext tasks, which use intrinsic properties of the data in order to evaluate representations [46]. Contrastive SSL is a SSL training technique that tries to discriminate between two types of examples, a) positive examples and b) negative examples, given an anchor, often by using noise-contrastive estimation [49]. In contrastive SSL (e.g., [11], [36], [50]), the loss takes the following form:

\[ L^{\text{self}} = - \sum_{i \in \mathcal{I}} \log \frac{\exp(v_i, v_{p(i)}/\tau)}{\sum_{n \in \mathcal{N}(i)} \exp(v_i, v_n/\tau)} \]

(1)

where \( v_i \) and \( v_{p(i)} \) are views of the same data element, \( \mathcal{I} \equiv \{1, \cdots, 2N\} \) with \( N \) being the number of data elements in the batch, \( \mathcal{N}(i) \equiv \mathcal{I} \setminus \{i\} \), and \( \tau \) is a temperature parameter.

This SSL paradigm has been explored extensively for image representation learning [11], [21] and graph representation learning [22]. These methods sample negative and positive examples based on a certain principle of semantic similarity in the data space. These methods sample negative examples using three main strategies: a) cross-scale based strategies (e.g. Computer vision [51]), b) augmentation based strategies (e.g. graph [34], text [39]), and c) hybrid strategies (cross-scale and augmentation based strategies). Cross-scale-based methods contrast using representations of intermediate layers. That is, given a batch of training examples, the intermediate representations of a data instance are positive examples and the representations of other data instances in that batch are considered to be negative examples. Augmentation based methods contrast using augmented versions of the input data. That is, augmentations of a data instance are considered to be positive examples, while augmentations of different data instances are considered to be negative examples. Data augmentation strategies are not always straightforward, especially in the cases of graphs and text, and are thus still being investigated. Hybrid methods contrast intermediate representations of augmented data [52]. Multiple works have stressed the importance of sampling; [23] showed than sampling harder negative examples is more beneficial for model performance.

However, these methods are known to suffer from large computational costs [53], since the computation of the loss requires multiple forward passes in order to get embeddings for the negative examples. This motivated the development of a new set of methods referred to as non-contrastive. BYOL [24] is a pioneering work in this line of research, which was later followed by many works (e.g. SwaV, Barlow Twins, SEER [26]–[28]). These works assert that negative examples regularize the models to prevent them from being naive, and they try to replace their role by explicit regularization using certain constraints that prevent the models from learning trivial representations.

**C. SUPERVISED CONTRASTIVE REPRESENTATION LEARNING**

Recently, many works have extended the self-supervised contrastive learning approach to the fully-supervised setting by leveraging label information for learning representations. In the computer vision field, [29] propose SupCon, an objective function for the task of image classification that bridges the gap between self-supervised learning and fully supervised learning. SupCon was shown to outperform SimCLR [11], Max-Margin [33] and cross-entropy on several benchmark datasets such as ImageNet. The SupCon objective function is defined by:

\[ L^{\text{sup}} = \sum_{i \in \mathcal{I}} - \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(v_i, v_p/\tau)}{\sum_{n \in \mathcal{N}(i)} \exp(v_i, v_n/\tau)} \]

(2)

where \( \mathcal{I} \equiv \{1, \cdots, N\} \) with \( N \) being the batch size, \( \mathcal{P}(i) \) is the set of indices of all data elements in the batch that belong to the same class as data element \( v_i \), \( \mathcal{N}(i) \equiv \mathcal{I} \setminus \{i\} \), and |.| denotes the cardinality operator.

In [55], the authors proposed a novel objective function for fine-tuning transformer-based language models which consists of a weighted sum of the cross-entropy loss and the above-mentioned supervised contrastive loss: \( (1 - \lambda) L_{CE} + \lambda L^{\text{sup}} \). This approach was shown to outperform a strong RoBERTa-large baseline on the GLUE benchmark dataset in few-shot learning settings. Many works have extended SupCon and proposed new variants. In [56], for instance, the authors evaluated three variants of pixel-wise label-based contrastive loss to pre-train a semantic segmentation model.

**D. NEGATIVE MINING IN CONTRASTIVE REPRESENTATION LEARNING**

In self-supervised contrastive learning, negative sampling has been shown to be very useful for learning good representations. Several strategies have been proposed to build negatives examples for visual presentations [11], [13], [14], [25]. In most of these works, the aim is to maximize the distance between the representation of a given anchor and those of negative examples that are difficult to discriminate against. In [14], for instance, the authors propose a method which consists of picking two percentiles \( w_k \) and \( w_l \) (\( l \in [0, 100]\)) and considering \( h_{n_c} \) as a negative example for a representation of a query \( h_q \) if and only if \( h_q^T h_{n_c} \) is within the \( w_k \)-th to the
$w_i$-th percentile of all $h_n \in Q^{-}_q$ where $Q^{-}_q$ denotes a set of negative examples. This enables to easily build hard negative examples (i.e., negatives that are hard to distinguish from the current sample) which are beneficial in learning powerful representations.

III. METHODOLOGY

A. REPRESENTATION LEARNING FRAMEWORK

We propose a supervised contrastive learning framework for textual representations. In our framework, we introduce a novel fully-supervised contrastive loss that we call SuperLoss. Our loss is optimized by training an encoder function, a neural network, to maximize the agreement between the normalized representations of a cluster of points with the same class label, while simultaneously pushing away clusters of samples from different classes.

The training takes as input $q$ batches of data, each of which is composed of sentences with the same target, where $q$ is the number of classes. All batches are forward propagated through a neural network to obtain a high-dimensional $l_2$ normalized embedding. The proposed framework is designed to maximize the agreement of vector representations of points belonging to the same class and contrast them with those of the other classes (see figure 1). To use the pretrained encoder function where the sub-index $w$ associates with the highest level representation of the encoder where $N_k$ is the batch size and $d$ is the dimension of the embedding vector.

2) Inter-class and Intra-class Distances

We first define notations and describe the proposed framework for classification tasks that will be essential for the analysis. Let $D = \{(x_i,y_i)\}_{i=1}^q$, be the available dataset, where $x_i$ represents the $i^{th}$ sentence and $y_i$ is its label. Let $S_k = \{(x_i,y_i)|y_i = k\}_{i=1}^N$, $B_k$, we denote as $w$, the sub-index $w$ refers to the weights of the encoders to be learnt. Let $H_k = f_w(B_k) \in \mathbb{R}^{N_k \times d}$ be the highest level representation of the encoder where $N_k$ is the batch size and $d$ is the dimension of the embedding vector. The $j^{th}$ row of $H_k$ is the transpose of the embedding vector associated with the $j^{th}$ sentence of $B_k$, which we denote as $h_j$, i.e.

$$H_k = \begin{bmatrix} h_1^T \\ h_2^T \\ \vdots \\ h_N^T \end{bmatrix} \in \mathbb{R}^{N \times d}.$$  

As shown in algorithm 1, in each training step our framework starts with sampling at random a batch of sentences from each class. Then we feed all of them to the encoder $f_w$ separately. Finally, our objective function takes as input the normalized representations matrices produced by the encoder.

3) Objective function Formulation

Now that we have all the mathematical notions needed, we proceed with the formulation of the objective function of the proposed contrastive learning framework. First, we calculate the dot product between the representation of each sentence in a class batch with those of all other sentences within the same batch:

$$G_{pos}^k = H_k H_k^\top \in \mathbb{R}^{N_k \times N_k}.$$  

Matrix $G_{pos}^k$ contains the similarities between sentences belonging to the same class $k$. The aim is to maximize these
similarities (intra-class similarity).

Then, we calculate the similarities between the representation of sentences belonging to different classes (intra-class similarity):

\[
G^{k}_{\text{neg}} := \mathcal{N}_{k} \cap_{j \neq k, 1 \leq j \leq q} H_{k}H_{j}^{\top} \in \mathbb{R}^{N_{k} \times N_{k}}
\]  

(5)

where \(\sum_{j \neq k, 1 \leq j \leq q} N_{j}\).

Next, we propose to average each matrix along the column axis after applying the exponential function:

\[
v^{\text{pos}} = \mathcal{N}_{k} \cap_{k \in [1,q]} \text{avg} \left[ \exp \left( \frac{G^{k}_{\text{pos}}}{\tau} \right) \right] \in \mathbb{R}^{N}
\]  

(6)

\[
v^{\text{neg}} = \mathcal{N}_{k} \cap_{k \in [1,q]} \text{avg} \left[ \exp \left( \frac{G^{k}_{\text{neg}}}{\tau} \right) \right] \in \mathbb{R}^{N}
\]  

(7)

where \(\tau \in \mathbb{R}^{+}\) is a scalar temperature parameter and \(N = \sum_{k=1}^{q} N_{k}\).

The proposed loss function is defined as follows:

\[
\text{SuperLoss} = -\frac{1}{N} \sum_{\ell=1}^{N} \log \left( \frac{v^{\text{pos}}[\ell]}{v^{\text{pos}}[\ell] + v^{\text{neg}}[\ell]} \right)
\]  

(8)

where \(v^{\text{pos}}[\ell]\) and \(v^{\text{neg}}[\ell]\) are the \(\ell^{th}\) elements of \(v^{\text{pos}}\) and \(v^{\text{neg}}\) respectively.

Here, the encoder’s weights will be learnt so as to maximize the elements of \(v^{\text{pos}}\) (clusters of points belonging to the same class are pulled together in the latent space) and minimize those of \(v^{\text{neg}}\), which will result in pushing representations of elements that do not belong to the same class apart from each other.

IV. EXPERIMENTS

In this section, we compare the proposed method with other techniques for sentence representation. First, we describe the datasets used in this paper, then we provide details of the architecture and training process of the proposed method.

A. DATASET AND TRAINING DETAILS

We evaluated the effectiveness of the proposed framework on sentence classification tasks by measuring accuracy on 6 benchmark datasets namely, SST-2, Yelp-P, yelp-F, Amazon-P, Amazon-F, and IMDb. Furthermore, the framework is also tested for representation learning on low-resource language setting datasets namely, MSAC, ASAC. We summarize each dataset based on their main task, domain, number of training examples, and number of classes in Table 1.

B. TRAINING DETAILS

Our framework allows various choices of the network architecture without any constraints. Here, we opt for simplicity and adopt the BiLSTM neural network architecture to compare different objective functions.

For ASAC and MSAC datasets, we used the following settings: we train our framework for 15 epochs using Adam [32] optimized with a learning rate of 0.003. However, for these datasets, we use an encoder with 1 hidden layer only
datasets, and those obtained by previous methods which are used for the classification task. We report the obtained results of the linear logistic regression model to solve the multiclass sentence classification. More precisely, we use the learnt representations to train a linear evaluation of the learned sentence representations. Following common protocol, to test our method, we opt for a batch size up to 1000. We run all experiments on 1 GPU server.

For the evaluation performance, we evaluated our approach for transfer learning in two different settings: (1) the classifier is trained on top of the frozen representations (transfer learning); (2) we train the classifier (projection head), where we allow all weights to be adjusted during training (fine-tuned). Simulations showed that the learned representations by the proposed objective function are better for the downstream tasks without adjusting them which means that our framework is capable of capturing robust features that better separate the classes. Table 3 illustrates the obtained results of both strategies. In this paper, we provide the results that we obtained with the transfer learning strategy.

Table 4 shows the obtained results using our objective function on the previously described datasets, as well as those based on the CE, Triplet-loss, N-pair-loss, as well as SupCon [29]. The results are given in terms of the accuracy score measured on the same balanced test set.

Following common practice in contrastive learning, we first study the importance of adding a projection head that maps representations to new space where the supervised contrastive loss is applied. Similar to [11], [44], we tested three different MLP architectures: (1) identity mapping; (2) linear projection \( z = g(h) = W^{(1)} h \in \mathbb{R}^{512} \); (3) non-linear projection with one additional hidden layer as used by several previous approaches \( z = g(h) = W^{(2)} ReLU(W^{(1)} h) \in \mathbb{R}^{512} \). Similar to what was found in previous works, we observe that a non-linear architecture is better than both the linear and the identity functions for the projection head network (See table 2). Note that the projection head network is used only in the contrastive training phase; it is discarded in the fine-tuning and inference phases.

### Table 1: Statistics of datasets used for evaluation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Train</th>
<th>#Dev</th>
<th>#Test</th>
<th>#Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>60k</td>
<td>3.5k</td>
<td>3.5k</td>
<td>2</td>
</tr>
<tr>
<td>Yelp-P</td>
<td>600k</td>
<td>50k</td>
<td>50k</td>
<td>2</td>
</tr>
<tr>
<td>Yelp-F</td>
<td>600</td>
<td>50k</td>
<td>50k</td>
<td>5</td>
</tr>
<tr>
<td>Amazon-P</td>
<td>3M</td>
<td>600k</td>
<td>400k</td>
<td>2</td>
</tr>
<tr>
<td>Amazon-F</td>
<td>2.5M</td>
<td>500k</td>
<td>650k</td>
<td>5</td>
</tr>
<tr>
<td>IMDB</td>
<td>20k</td>
<td>5k</td>
<td>25k</td>
<td>2</td>
</tr>
<tr>
<td>MSAC</td>
<td>1.6k</td>
<td>0.2k</td>
<td>0.2k</td>
<td>2</td>
</tr>
<tr>
<td>ASAC</td>
<td>6.8k</td>
<td>0.8k</td>
<td>0.8k</td>
<td>2</td>
</tr>
</tbody>
</table>
those obtained with other objective functions. The results are given in terms of the accuracy score measured on the same balanced test set. It is seen that in 87% of the cases, our framework achieves better performance; the gain in performance is significant. Indeed, SuperLoss leads to, for instance, a 2.87% improvement of accuracy on SST-2, 3.33% improvement on Yelp-P, 3.34% improvement on Yelp-F, 3.93% improvement on Amazon-F, 3.93% improvement on Amazon-P, 5.95% improvement on MSAC compared to CE loss, respectively. The large performance gap for MSAC dataset demonstrates that cross-entropy struggles with separating the classes when dealing with small datasets. Furthermore, the results for MSAC and ASAC prove that our framework is very promising for under-resourced languages. Moreover, our experiments showed that CE overfits the MSAC dataset very quickly, with a training accuracy of 96.33% and only 72% accuracy on test. The overfitting problem cannot be explained by the large number of parameters of the model, since our objective function also uses the same model architecture (i.e., the same number of parameters as CE). Indeed, the problem can be explained by the fact that CE learns very poor margins between the two classes.

V. ABLATION STUDY

We investigate here the effects of different parameters on performance. All experiments have been conducted using Yelp-F dataset. We run each experiment with 10 different seeds, and report the average test accuracy.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Projection model</th>
<th>Identity</th>
<th>Linear</th>
<th>Non-Linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>93.40</td>
<td>93.60</td>
<td>94.15</td>
<td></td>
</tr>
<tr>
<td>Yelp-P</td>
<td>95.13</td>
<td>95.31</td>
<td>95.45</td>
<td></td>
</tr>
<tr>
<td>Yelp-F</td>
<td>63.74</td>
<td>64.02</td>
<td>64.89</td>
<td></td>
</tr>
<tr>
<td>Amazon-P</td>
<td>93.91</td>
<td>94.07</td>
<td>94.71</td>
<td></td>
</tr>
<tr>
<td>Amazon-F</td>
<td>60.01</td>
<td>60.13</td>
<td>61.94</td>
<td></td>
</tr>
<tr>
<td>IMDB</td>
<td>84.33</td>
<td>84.45</td>
<td>86.82</td>
<td></td>
</tr>
<tr>
<td>MSAC</td>
<td>76.16</td>
<td>76.89</td>
<td>80.10</td>
<td></td>
</tr>
<tr>
<td>ASAC</td>
<td>81.54</td>
<td>81.87</td>
<td>82.63</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2: Linear evaluation of representations with different projection heads g(·) (Accuracy). The representation h (before projection) is 512-dimensional (%).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>learning strategy</th>
<th>Transfer-learning</th>
<th>Fine-tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>94.15</td>
<td>94.42</td>
<td></td>
</tr>
<tr>
<td>Yelp-P</td>
<td>95.53</td>
<td>95.45</td>
<td></td>
</tr>
<tr>
<td>Yelp-F</td>
<td>64.89</td>
<td>64.23</td>
<td></td>
</tr>
<tr>
<td>Amazon-P</td>
<td>94.71</td>
<td>94.58</td>
<td></td>
</tr>
<tr>
<td>Amazon-F</td>
<td>61.94</td>
<td>61.11</td>
<td></td>
</tr>
<tr>
<td>IMDB</td>
<td>86.82</td>
<td>85.14</td>
<td></td>
</tr>
<tr>
<td>MSAC</td>
<td>80.10</td>
<td>78.21</td>
<td></td>
</tr>
<tr>
<td>ASAC</td>
<td>82.63</td>
<td>82.16</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 3: Comparison of transfer learning and fine-tuning performance (Accuracy).

A. IMPACT OF THE MODEL ARCHITECTURE

We consider several BiLSTM encoder-based architectures with growing capacity, particularly the number of hidden layers and the number of hidden units of the model. In Figure 4a depicts how changing the model’s architecture affects the quality of the learnt representations on the downstream task. We found that increasing the number of hidden layers works better for the proposed framework. In this work, we used three layers due to the GPU memory constraint. Our experiments show that SuperLoss surpasses the cross-entropy loss using the same model architecture for all configurations. Figure 4b shows the obtained accuracy score for different numbers of hidden units ({100, 200, 300, 512, 768}). It is clear that by increasing the dimension of the hidden layers the model works better, though the gain beyond 512 is small.

B. TRAINING WITH LARGE BATCH SIZE

Here, we show empirically the impact of batch size on the quality of the models’ representations trained with the same number of epochs (60 epochs). Figure 4c shows the accuracy of a linear classifier trained upon the learned 512-dimensional representations while varying the batch size. Similar to self-supervised contrastive learning, we found that training the model on larger batch sizes have a significant (high-quality representation) advantage over the smaller ones. Note that with the CE loss, the highest scores are obtained for a batch size of 500; larger batch sizes decreased the accuracy of the downstream classification task. In contrast, for our objective function, larger batch sizes provide more negative examples, thus improving the results. In this ablation study, we evaluated the model’s representations with batch sizes of {500, 750, 1000}. However, we believe that by increasing the batch size further, the model can learn higher quality features that can be useful on downstream task.

C. IMPACT OF THE TEMPERATURE

Figure 4d shows the impact of scalar temperature parameter on Top-1 accuracy performance of our framework. Empirical observations show that smaller temperature benefits training more than higher ones (lower temperature increases the influence of examples that are harder to separate). However, very low temperatures are harder to train due to numerical instability. Thus, an appropriate temperature can help the model learn from hard negatives. The empirical behavior of the effect of the temperature parameter is in line with the observations made in previous work related to self-supervised/fully-supervised contrastive learning.
D. EFFECT OF THE NUMBER OF POSITIVE/NEGATIVES EXAMPLES

In recent contrastive learning approaches, the number of negative examples has been shown to be a key component for learning high-quality representations. The majority of these methods sample negatives from a very large batches or a memory bank to increase the number of negative examples beyond the batch size and have reported significant performance gains with increasing batch sizes. In this research, we study the impact of the number of the negative/positive examples. To the best of our knowledge, this work is the first to consider the effect of the number of positive and negative examples on the fully-supervised contrastive learning setting. Similar to self-supervised contrastive learning, we found that increasing the number of negative examples is beneficial for learning representations in our framework. Our experiments show that a high number of negatives helps our loss function to encourage the encoder to find features that can better separate the representations of different classes in the latent space. In this paper, we use 900 (see figure 5c) negative examples for each class due to the GPU memory constraint. However, we believe that increasing further the number of negative examples will produce better results. In figure 5a, we report the top-1 accuracy performance of the downstream classification task for different values of the number of positive examples. In each simulation, we fixed the number of positive instances for a given class and the negative examples are uniformly sampled from the remaining classes (e.g. 300 positive points and 175 negative points for each remaining classes). We observe that training our supervised objective function with a high number of positive examples leads to good representations. However, simulations show that at beyond a certain threshold, increasing the number of positives decreases the accuracy of the downstream task.

VI. NEGATIVE SAMPLING STRATEGY

Contrastive learning is recently proposed to learn feature embeddings in a self-supervised manner. The latter relies on the positive and negative instances. As revealed by recent studies, negative examples are crucial in learning robust representations. Accordingly, different strategies have been proposed to sample negatives that are hard to distinguish for a given anchor in the latent space [11], [13], [14], [25]. In this paper, we propose a simple yet effective strategy for selecting hard negative examples for supervised contrastive learning for text representations. In the proposed framework, in each iteration, we maximize the distance of the average similarity of a given anchor with all instances from the remaining classes which means that all negative examples are considered as hard negative. To overcome this, we first train the model for a number of epochs (20 epochs on YELP-F dataset) using all negative instances within the batches, then we modify the training strategy by maximizing the distance of anchors with those pairs that have a similarity higher than a fine-tuned threshold. By doing so, our loss will guide the encoder function to produce representations by considering...
only truly hard negatives. In our strategy, we simply select 
these examples by first ordering the similarities for a given 
an anchor, then we select the most similar instances (hard 
negatives). Following this new strategy, the simulations show 
that, indeed, the distances between anchors and the negative 
examples become higher compared to those obtained using 
the previous learning strategy. We also noticed that for a 
given anchor, the selected hard negative examples are those 
from the closest class (for class ‘Very Positive’ the majority 
of the negative examples are from the class ‘Positive’ which 
is the most similar class to that of the anchor). We fine-tune 
the similarity threshold using the validation set by selecting 
the top most similar examples to the anchor. Figure 5c 
shows the different top-1 accuracy obtained as a function 
of similarity threshold. We evaluate our negative sampling 
strategy on several benchmark datasets. Experimental results 
show that the strategy is beneficial.

In table 5, we report the SuperLoss* which refers to 
the results obtained using the proposed negative sampling 
strategy. In bold, we report the best performance. As it can 
be seen in the table, SuperLoss* outperforms SuperLoss 
in most cases which means that the proposed strategy leads 
the encoder to learn better representations (relevant features 
for distinguishing the classes in the latent space).

### VII. CONCLUSION

We proposed SuperConText, a new framework for learning 
text representations using a novel supervised contrastive loss. 
SuperConText encourages an encoder function to learn rep-
resentations by maximizing the average agreement between 
the representation of an anchor and those of N positive 
pairs, determined as elements belonging to the same class 
as the anchor, while distancing the anchor’s representation 
with those of negative examples. Simulations show that the 
proposed framework outperforms several methods based on 
other objective functions on various benchmark datasets. We 
have conducted a number of experiments to understand the 
effects of both negative and positive examples on the quality 
of learned representations. We further introduced a simple yet 
effective negative sampling strategy to enhance the quality of 
the representations. The experimental results show that the 
proposed strategy improves performance in most cases.

### REFERENCES

[1] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. 
conference on computer vision and pattern recognition 2016, 770–778.
[2] Carlini, Nicholas and Wagner, David. Towards evaluating the robustness of 
[4] Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio, Yoshua. Neural ma-
chine translation by jointly learning to align and translate. arXiv preprint 
and Artzi, Yoav. Revisiting few-sample BERT fine-tuning. arXiv preprint 
[6] Szegedy, Christian and Vanhoucke, Vincent and Ioffe, Sergey and Shlens, 
Jon and Wojna, Zbigniew. Rethinking the inception architecture for com-
puter vision. Proceedings of the IEEE conference on computer vision and 
pattern recognition 2016, 2818–2826.
[9] Zhang, Hongyi and Cisse, Moustapha and Dauphin, Yann N and Lopez-
Paz, David. mixup: Beyond empirical risk minimization. arXiv preprint 
[10] Hinton, Geoffrey and Vinyals, Oriol and Dean, Jeff. Distilling the knowl-
Geoffrey. A simple framework for contrastive learning of visual represen-
[12] Weng, Bowen and Xiong, Huaying and Zhao, Lin and Liang, Ying-in and Zhang, Wei. Momentum Q-learning with finite-sample con-

<table>
<thead>
<tr>
<th>Dataset / Loss</th>
<th>SuperLoss</th>
<th>SuperLoss*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>94.15</td>
<td>94.62</td>
</tr>
<tr>
<td>Yelp-P</td>
<td>95.45</td>
<td>95.89</td>
</tr>
<tr>
<td>Yelp-F</td>
<td>94.71</td>
<td>94.93</td>
</tr>
<tr>
<td>Amazon-p</td>
<td>61.94</td>
<td>62.76</td>
</tr>
<tr>
<td>Amazon-F</td>
<td>61.94</td>
<td>62.88</td>
</tr>
<tr>
<td>IMDB</td>
<td>86.82</td>
<td>86.94</td>
</tr>
<tr>
<td>MSAC</td>
<td>80.10</td>
<td>81.32</td>
</tr>
<tr>
<td>ASAC</td>
<td>82.63</td>
<td>83.45</td>
</tr>
</tbody>
</table>

TABLE 5: Performance Results (%)
YOUNESS MOUKAFIH has received the master’s degree in big data analytics and smart systems from Sidi Mohammed Ben Abdellah University, Fez, Morocco, in 2018. He is currently pursuing the joint Ph.D. degree with University of Lorraine and International University of Rabat. His research interests include machine learning, natural language processing, computer vision and graph representation learning.

NADA SBIHI is an Assistant Professor at ESIN engineer school and member of TIClab at International University of Rabat. She obtained her engineering degree from the National Institute of Posts and Telecommunications (INPT) in 2006, in computer science, networks and systems. After three years of experience as a Business Intelligence engineer at Sofrecom, she continued her studies at Pierre and Marie Currie University (Paris 6), where she obtained a master’s degree and then a doctorate. His doctoral studies were carried out at INRIA under the supervision of Dr James ROBERTS. Before joining the UIR faculty, she worked at TIClab as a post-doctoral fellow under the supervision of Professor Mounir GHOGHO. His research themes are: Big Data and artificial intelligence and their applications in various fields: intelligent transport, social networks, content-oriented networks, urban pollution and health.

KAMEL SMAILI is Professor of Computer Science at University of Lorraine. He is an expert in natural language processing. He carries out his research activity at LORIA, a UL laboratory. He obtained an engineering degree in computer science from the University (USTHB), Algiers, Algeria in 1986, a doctorate from the University Henri Poincaré in 1991 and an HDR from the University of Nancy 2 in 2001. Between 1999 and 2001, he was seconded to the CNRS. In 2014, he created the SmarT research team (https://smart.LORIA.fr/) at LORIA. The main activity of his team is processing natural language using statistical and neural paradigms and more generally the development of deep machine learning methods for NLP. He has supervised 19 doctoral students, 17 of whom have defended and two are at the end of their thesis. Every year, he reviews ANR and foreign project files (NeuroInsight 2022). Since 2015, he co-directs an associated international laboratory (LIA CNRS) named DataNet. In this laboratory, are involved two laboratories of the University of Lorraine and several Moroccan universities including the UIR. He has published more than 150 articles in international journals and conferences1. He coordinated several international projects, the last one ended in November 2019, the Chist-Era AMIS (Access Multilingual Information opinions) project. He leads a new project TRADEF (ASTRID ANR) that started in January 2023. He has organized and chaired international conferences (ICNLSSP 2017 and ICALP 2019) and he is a member of several international conference program committees such as: ICASSP, Interspeech, LREC, etc. He has been invited to give lectures in Japan, France, Spain, Poland, Algeria, Tunisia and Morocco. He has been a member of more than 70 thesis and HDR juries in France, Germany, Spain, India, Ireland, Algeria, Morocco and Tunisia.

MOUNIR GHOGHO (Fellow, IEEE) has received the M.Sc. degree in 1993 and the PhD degree in 1997 from the National Polytechnic Institute of Toulouse, France. He was an EPSRC Research Fellow with the University of Strathclyde (Scotland), from Sept 1997 to Nov 2001. In Dec 2001, he joined the school of Electronic and Electrical Engineering at the University of Leeds (England), where he was promoted to full Professor in 2008. While still affiliated with the University of Leeds, in 2010 he joined the International University of Rabat (Morocco) where he is currently Dean of the College of Doctoral Studies and Director of TIClab (ICT Research Laboratory). He is a Fellow of IEEE, a recipient of the 2013 IBM Faculty Award, and a recipient of the 2000 UK Royal Academy of Engineering Research Fellowship. He is the co-founder and co-director of the CNRS-associated International Research lab DataNet, in the field of Big Data and artificial intelligence. His research interests are in Machine Learning, Signal Processing and Wireless Communication, on which he has published over 300 papers in journals and conferences. He has coordinated around 20 research projects and supervised over 30 PhD students in the UK and Morocco. In the past, he served as an associate editor of many journals including the IEEE Signal Processing Magazine and the IEEE Transactions on Signal Processing.

1 https://members.LORIA.fr/KSmaili/publications/