Label Supervised Contrastive Learning for Imbalanced Text Classification in Euclidean and Hyperbolic Embedding Spaces

Anonymous ACL submission

Abstract

Text classification is an important problem with a wide range of applications in NLP. However, naturally occurring data is imbalanced which can induce biases when training classification models. In this work, we introduce a novel contrastive learning (CL) approach to help with imbalanced text classification task. CL has an inherent structure which pushes similar data closer in embedding space and vice versa using data samples anchors. However, in traditional CL methods text embeddings are used as anchors, which are scattered over the embedding space. We propose a CL approach which learns key anchors in the form of label embeddings and uses them as anchors. This allows our approach to bring the embeddings closer to their labels in the embedding space and divide the embedding space between labels in a fairer manner. We also introduce a novel method to improve the interpretability of our approach in a multi-class classification scenario. This approach learns the inter-class relationships during training which provide insight into the model decisions. Since our approach is focused on dividing the embedding space between different labels we also experiment with hyperbolic embeddings since they have been proven successful in embedding hierarchical information. Our proposed method outperforms several state-of-the-art baselines by an average 11% F1. Our interpretable approach highlights key data relationships and our experiments with hyperbolic embeddings give us important insights for future investigations. We will release the implementation of our approach with the publication.

1 Introduction

A common way of approaching the text classification problem is training a model using pre-trained text embeddings as language features (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2018). These embeddings can be fine-tuned using the signals from an objective function to improve

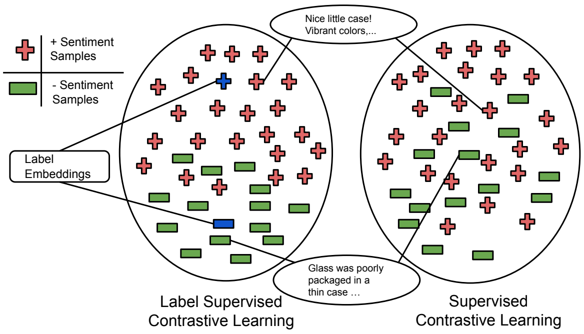

Figure 1: SCL can cause the embeddings for positive and negative sentiment text samples to be dispersed together in the embedding space (right illustration). Our approach in contrast utilizes the embedding space more effectively (left illustration). This is also shown in the form of Euclidean distance between embeddings of text samples of opposite sentiment. Our approach embeds these samples farther away from each than SCL in terms of Euclidean distance: *13.2* vs. *3.2*.

their efficacy for the classification task at hand. However, a common impediment to training a robust classifier is the fact that naturally occurring data is imbalanced. Since classifier predictions reflect the distribution of the training data, they can induce bias. There are many approaches proposed to address this issue, such as oversampling, undersampling, using weighted objective functions or using situation/domain specific methods to improve the robustness of classification models (Chawla et al., 2002a; Tahir et al., 2012). Our work focuses on introducing a novel algorithm to deal with the challenges of imbalanced data.

Recent research shows an increasing use of contrastive learning (CL) to solve different problems in areas of computer vision and NLP (Gao et al., 2021a; Hénaff et al., 2019; Jaiswal et al., 2021). In this work, we explore CL to address the problem of imbalanced text classification. In general, CL uses anchors to embed similar samples closer in the embedding space while pushing dissimilar examples away. Unsupervised CL (Tian et al., 2019) tries to contrast a data sample, called anchor, with every sample in the batch while supervised CL (SCL) (Khosla et al., 2020a) tries to utilize label information and embed samples from the same class as the anchor closer to each other. However, these CL approaches rely on utilizing data embeddings as anchors which are scattered over the embedding space. We hypothesize that label embeddings, representing a label category in the embedding space, can be utilized as key anchors in CL. This allows a model to embed data samples closer to their category representations and results in a model learning better embedding representations for the data. We present an illustration in the figure 1 where we compare how SCL divides embedding space in comparison to our approach utilizing label embeddings in a binary classification task. This shows that our approach is able to achiever better class separation between data belong to different labels. This is highlighted by the fact that distance between a positive and negative text embedding pair is larger when our approach is utilized in comparison to the SCL.

Our proposed approach uses two embedding modules to embed text and labels individually which are fine-tuned using label-supervised CL (LSCL). The text embedding module utilizes a self-attention mechanism to make use of token-level relations while the label embedding module is an embedding layer. Both of these are fine-tuned using our proposed CL objective. In addition, we also experiment with hyperbolic embeddings, where pre-trained model (e.g. BERT), provides representations with hyperbolic structure (Chen et al., 2021). We show that our approach outperforms several SOTA and CL baselines in both Euclidean and hyperbolic spaces. Finally, we also try to improve the interpretability of our model by proposing a modification to our approach which allows it to represent inter-class relationships in an intuitive manner for a multi-class classification task. We structure our contributions in the following manner: section 2 focuses on related work, section 3 highlights the background formulations on CL and section 4 details our approach. Section 5 presents generalization of our model to hyperbolic spaces, section 6 describes our experiment setup followed by the performance analysis in section 7. Finally, we conclude by presenting the limitations and conclusion of our work.

2 Related Work

Data imbalance is a common problem and classification literature has adopted a variety of approaches to deal with the biases it might introduce. One of these ways is oversampling of less frequent data. SMOTE is the first minority oversampling method (Chawla et al., 2002b). Iglesias et al. (2013) presents a hidden markov model which generates data from minority distribution. Other works focus on the use of oversampling on the basis of sample difficulty (Tian et al., 2021). Song et al. (2016) combines the SMOTE technique with a K-Means based undersampling algorithm to try and improve classifier performance on an imabalanced dataset. Some methods undersample the majority class samples to create a balanced data distribution for the training process. Smith et al. (2013); Anand et al. (2010) both present methods which use a notion of sample difficulty to undersample the majority class samples.

Some works rely on weighing the objective function to deal with data imbalance. The idea is to increase the loss contribution for the minority classes during the training. Cao et al. (2019); Chen et al. (2016); Park et al. (2021) each presents a different way of weighing the label-specific loss.

There is a third class of works which tries to introduce novel algorithms focused on the data imbalance problem. These methods avoid inducing biases that might arise because of distribution changes in data. An example is (Gao et al., 2021c) which introduces a convolution based algorithm to handle the class imbalance problem in data. Our work fits in this category as we explore the use of label-supervised CL to address this problem. Another example is Díaz-Vico et al. (2018), which uses cost-sensitive learning to regularize the posterior distributions for a given sample. This relies on domain specific information which can be hard to obtain in realistic scenarios (Krawczyk, 2016).

Lately, contrastive learning is being used in a variety of tasks due to its effective utilization of embedding space. Kang et al. (2021) present KCL which is a variation of SCL algorithm (Khosla et al., 2020b) and explores the use of contrastive learning for learning balanced embedding spaces in the area of computer vision. Lopez-Martin et al. (2022); Zhang et al. (2022) present label-centered variations of CL methods but do not explore the data-imbalance effects or the effect of computational spaces on the model performance.

Hyperbolic spaces are becoming well-known for their superiority in embedding hierarchical information like WordNet graphs (Nickel and Kiela, 2017, 2018). This is because of their natural hierarchical structure. We view the classification task as a sub-class of hierarchical problem where a label embedding represents a category and each data sample is near to its label embedding. This is why we try to assess the performance of our model in the hyperbolic space as well. Another motivation for our work comes from Chen et al. (2021), which show that BERT embeddings contain hyperbolic structure between tokens by probing BERT embedding in hyperbolic spaces.

3 Contrastive Learning Overview

Contrastive learning tries to embed similar samples closer in the embedding space by trying to make the samples closer to their anchors. Formally, CL can be expressed as (Tian et al., 2019; Khosla et al., 2020b):

3.1 Contrastive Learning

We can define $\{(t_1, y_1), (t_2, y_2), ..., (t_N, y_N)\} = D$ as a dataset consisting of a set of text $t_i = \{w_{i1}, w_{i2}, .., w_{is_n}\}$ and label pairs y_i, where s_n is the length of the text sample t_i and w_ij is the token representation corresponding to the j^{th} token in the text sample t_i. Given an embedding representation x_i for the text sample t_i, we can define contrastive learning objective L for mini-batches $B_k \subset D$ of size b_n as:

$$\frac{-1}{b_n} \sum_{x_i \in X_k} log \frac{exp(sim(x_i, x_i^+))}{\sum_{x_j \in \{x_i^+\} \cup A(i)} exp(sim(x_i, x_j))} \tag{1}$$

where sim is a similarity function (usually the dot product), $A(i) = \{x_j | x_j \neq x_i, x_j \in X_k\}$, X_k is set of text representations in the mini-batch B_k and x_i^+ is an augmented representation of the text sample t_i. This objective causes a model to learn embedding for x_i which are closer to its augmentation and pushes it away from other examples in the mini-batch.

4 Proposed Approach

We propose a supervised CL approach which uses label embeddings as anchors and causes the model to learn representations which are closer to their respective label representations or key anchors in the embedding space. An architecture diagram for our approach, Label Supervised Contrastive Learning (LSCL), is presented in the figure 2 and its formulation L_{LSCL} is given as follows:

$$L_{LSCL} = \frac{1}{b_n} \sum_{x_i \in X_k} -log \frac{exp(sim(x_i, l_i))}{\sum_{l_j \in L} exp(sim(x_i, l_j))} \tag{2}$$

where L is the set of all label representations. This approach embeds the text samples closer to their label embeddings in the embedding space. Labels for each text embedding can be predicted by choosing the label whose embedding is closest.

4.0.1 Increasing Interpretability Through Learning Inter-Class Relationships

In a multi-class classification scenario, sometimes label categories are related to each other, e.g. emotions *love* and *joy* are likely to be expressed in similar ways in many cases. In such cases it is hard to interpret how model embedded certain text samples in certain parts of the embedding space. Considering this we modify our approach to learn interpretable inter-class relationships, in form of a weight matrix, so these could be used to highlight the reasoning behind model decisions. This variation L_{LSCL-W} can be formulated as follows:

$$\frac{-1}{b_n} \sum_{x_i \in X_k} log \frac{exp(sim(x_i, l_i))}{\sum_{l_j \in L - l_i} w_{ij} exp(sim(x_i, l_j))} \tag{3}$$

where $w_{ij} \in W^{|L| * |L|}$ is a weight matrix we learn during the training process and $w_{ij} = 1$ when $i = j$. A problem here is that a learning method would just take the weight matrix W to zero. To prevent that, we add a Shanon Entropy (Shannon, 1948) regularization term to the objective which ensure that there is a relative difference in the magnitude of weights so the new objective L'_{LSCL-W} becomes:

$$\begin{aligned} L'_{LSCL-W} &= L_{LSCL-W} + \lambda H(W_i) \\ H(W_i) &= -\sum_{w_{ij} \in W_i} w_{ij} log(w_{ij}) \end{aligned} \tag{4}$$

where w_{ij} is the relation between labels l_i and l_j. The greater the weight the more difficult to separate data belonging to these two labels which is why the model assigns a higher weight to the contrastive weight of these labels. The λ is a term between 0 and 1 to control the contribution of entropy objective. W is not symmetric because of the data imbalance.

4.0.2 How Our Approach Helps with Data Imbalance

Our approach tries to bring the data samples in the closer to their respective labels and push the other

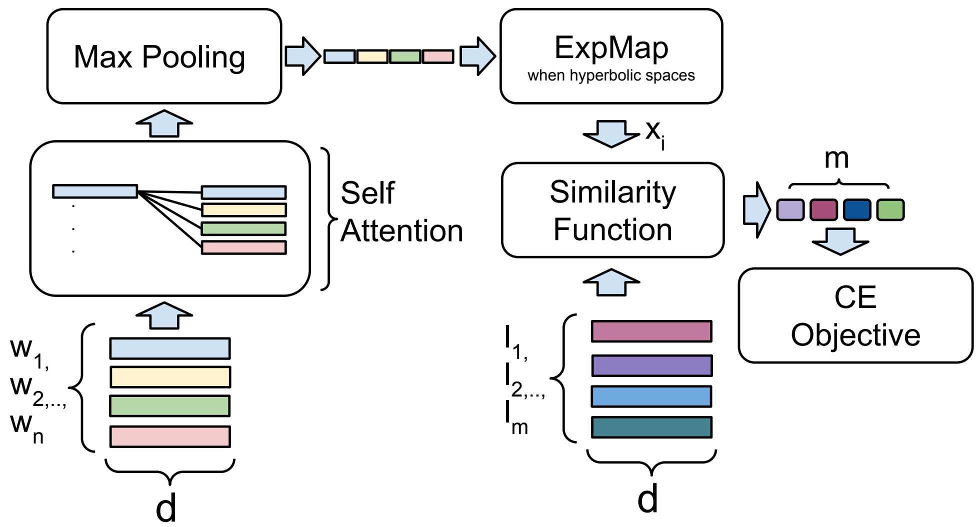

Figure 2: A batch of utterances is passed through a self-attention encoder to obtain text embeddings. These embeddings may be passed through an exponential map function to obtain embeddings in hyperbolic plane. Label embeddings are obtained by passing the input labels through a label embedding layer. These label embeddings are used as anchors in the CL objective which outputs loss signals for fine-tuning both the text encoder and label embedding layer together.

label embeddings away. This creates a push-pull effect for data samples w.r.t to the label embeddings. Both of these effects help improve the model performance. The data samples belonging to the majority class help improve the performance for the minority classes in this way as these samples push the minority label embeddings away as well while trying to get close to the their respective label embeddings.

5 Generalization to Multiple Computational Spaces

Hyperbolic models show great promise for embedding hierarchical or graph structures (Nickel and Kiela, 2017, 2018). Our CL approach treats the classification problem as a hierarchical task by trying to learn the embedding regions for their respective labels. In addition, Chen et al. (2021) shows that pre-trained text embedding contain hyperbolic structure. Due to these reasons we explore the effect of hyperbolic embeddings on our approach and show that these models perform competitively to their Euclidean counterparts and outperform all the baselines.

5.1 Manifold Centric Label Embeddings

We wanted to make use of the information encoded in the pre-trained textual representations and they are usually trained in Euclidean space. Due to this reason, we make use of hyperbolic exponential map to obtain hyperbolic textual embeddings. However, label embeddings need not have any such restriction so we embed the labels in a manifold specific representation space. This entails that hyperbolic versions of our approach embed labels directly in the hypberbolic space so there is no need to use exponential map to obtain label embeddings.

5.2 Notion of Similarity

Contrastive learning uses a measure of similarity to embed similar examples closer to each other in a higher dimensional space. We generalize the notion of similarity between two vectors h and h' across Euclidean and hyperbolic manifolds, in an intuitive manner, as follows:

$$sim_{manifold}(h, h') = -d_{manifold}(h, h') \tag{5}$$

where $d_{manifold}$ represents the manifold specific distance function.

5.2.1 Vector Similarity in Euclidean Space

Following the formulation specified above the similarity function can be defined as:

$$sim_{eucl}(h, h') = -\sum_{i<=d} \sqrt{(h_i - h'_i)^2} \tag{6}$$

5.2.2 Vector Similarity in Hyperbolic Space

For our hyperbolic model variation, we use *Lorentz* formulation of Riemannian Manifolds because (Nickel and Kiela, 2018) suggests that *Loretnz* formulation of hyperbolic space is numerically stable compared to the *Poincare'* formulation. The similarity function for the hyperbolic variation is thus given as:

$$sim_{lorentz}(h, h') = -arcosh(- < h, h' >_L)$$
$$< h, h' >_L = -h_0 h'_0 + \sum_{1<=d} h_i h'_i \tag{7}$$

6 Experiment Setup

We conduct experiments to compare the performance of our approach on two classification tasks with several baselines. In addition, we conduct experiments to compare the performance of our approach between hyperbolic and Euclidean embeddings. We rely on 256 dimensional variation of BERT (Turc et al., 2019) to obtain the seed embeddings for our text encoder.

6.1 Datasets

We rely on two datasets for the purpose of our evaluation: 1) Amazon Reviews Sentiment Classification (Keung et al., 2020) 2) Twitter Emotion Classification dataset[1]. We create a binary sentiment classification task from the former by splitting the the review ratings into positive and negative classes. Reviews with rating greater >=4 are categorized as positive and reviews with rating <=2 are considered negative. We induce a data imbalance of 9:1 for positive and negative classes respectively to obtain an imbalanced dataset containing a total of 15000 reviews.

Twitter emotion dataset is a multi-class data with six emotions: sadness, joy, love, anger, fear, surprise, contains a total of 20000 tweets and is naturally imbalanced. Class ratios for both datasets are given in the tables 1 and 2.

6.2 Model Parameters

Figure 2 shows the architecture of the text encoder we use for CL. We utilize a self-attention layer to embed the text embeddings. When we need to obtain the hyperbolic embeddings we utilize the exponential map operation to project the euclidean embeddings into the hyperbolic space. We seed our text embedding layer with BERT embeddings which improves the training time of the model during fine-tuning with CL. The right side of the architecture diagram shows the label embeddings which are used to computer similarity with the text embeddings. These embeddings are fine-tuned using the LSCL training objective shown in the section 4. We use a prefix of E or H to indicate whether the model utilizes euclidean embeddings or hyperbolic ones respectively.

When using euclidean embeddings we fine-tune our model using the Adam optimizer (Kingma and Ba, 2014) while we use Reimannian SGD[2] to optimize the hyperbolic weights as it relies on the exponential map to update the weights using Reimannian gradients. Inspired from (Gao et al., 2021b), we use a dropout layer (*rate: 0.1*) to obtain the augmented representations when needed. We use a learning rate of 10^{-3} for Adam and a learning rate of 10^{-1} for Reimannian optimizer with a batch size of 64.

6.3 Baselines

We compare our proposed approach with several baselines. We divide the baselines in two groups: 1) SOTA baselines – baselines designed to help with data imbalance in classification task; and 2) CL baselines – baselines utilzing other versions of contrastive.

6.3.1 Baselines for Imbalanced Classification

We use the following baselines to indicate the advantages of using a label-supervised CL approach to deal with the problem of class imbalance in a classification task.

SetConv: Gao et al. (2021c) presents a convolution based method to learn better representations for the minority class samples. It utilizes a minority class representative as anchor to learn kernel weights during the training process.

GILE: Pappas and Henderson (2019) uses joint embeddings obtained using a dimension-wise product of text and label embeddings. Their approach uses a fully-connected layer to score these joint embeddings and makes use of binary cross-entropy objective to train the model.

BertGCN: Lin et al. (2021) treats the textual data as a graph of token and document representations. The graph encodes token-level information using measures like tf-idf and documents using BERT representations. The approach utilizes a graph convolution operation to obtain a vector representation for a given text document.

6.3.2 Contrastive Learning Baselines

We utilize the following CL approaches to highlight the advantages of utilizing our CL approach in a

[1] https://huggingface.co/datasets/emotion

[2] https://github.com/facebookresearch/poincare-embeddings

classification task.

K-Contrastive Learning: Kang et al. (2021) presents KCL, a variation of supervised contrastive learning in the domain of computer vision which learns balanced features spaces. Instead of using batch data samples as positive and negative anchors their approach samples k samples for each class from training data.

Supervised Contrastive Learning: SCL (Khosla et al., 2020a) is a CL approach which tries to contrast data samples from one class with data samples belonging to other classes while trying to bring the data samples from same classes closer to each other. As highlighted by the results presented below, this is a poor choice for imbalanced classification as skew in data distribution will create a bias in favor of majority class data when the model tries to bring samples from same class together.

7 Performance Analysis

We evaluate the performance of our approach on two tasks: binary sentiment classification and multi-class emotion classification. Both tasks highlight different aspects of our approach as a binary classification task with sufficient disparity in labels might be easier than a multi-class classification task which requires a model to learn inter-class relationships. For all our experiments, we measure the overall performance of a model using macro F1 score average because it equally weighs the model performance of the minority classes; hence reflects effect of data imbalance. Our key insights are:

- Our proposed CL approach is able to outperform the baselines in both computational spaces as shown in the tables 1 and 2).
- Euclidean version of our approach achieves the best overall performance as shown in the tables 1 and 2.
- We can improve model-decision interpretability by learning inter-class relationship weights. This is highlighted in the figure 4.
- Visualizing our approach in a 2-dimensional setting shows that hyperbolic version of our approach divides the embedding space fairly in the binary setting. This is highlighted in the figure 3.

7.1 Baseline Performance Comparison

We compare the performance of our approach with several contrastive learning and SOTA baselines as stated in the section 6.3. In short, our approach outperforms the best SOTA baseline by a margin of 7% and 14% in the tasks of binary sentiment classification and multi-class emotion classification, respectively. These results are shown in the tables 1 and 2 respectively. In addition our approach does not sacrifice the majority class performance for a gain in minority class performance. This can be observed in both the binary and multi-class classification settings as our model consistently outperforms all the baselines in both overall and per-class performance, as highlighted in the table 1.

Model	Macro F1	Positive Class F1	Negative Class F1
Class Ratios		0.9	0.1
SOTA Baselines			
SetConv	0.682	0.888	0.476
GILE	0.706	0.951	0.462
BertGCN	0.702	0.948	0.455
Contrastive Learning Baselines			
SCL	0.594	0.95	0.237
KCL(k=5)	0.646	0.944	0.346
Our Approach			
HLSCL	0.72	0.930	0.511
ELSCL	**0.779**	**0.959**	**0.6**

Table 1: This table shows the per class F1 scores achieved by our model and their corresponding macro averages on Amazon Reviews Sentiment classification task. We show the results of both hyperbolic and euclidean models. The bold numbers represent the best performing model.

In the multi-class classification setting, the best performing baseline for the minority emotion *surprise* is BertGCN with a macro F1 of 38%. Our approach utilizing hyperbolic embeddings outperforms BertGCN by 7% in the minor class while achieving better performance in the majority classes – sadness and joy, as shown in the table 2.

Comparing the performance of our approach with CL baselines in the tables 1 and 2, specially SCL, shows the our approach to CL outperforms the other approaches in the task of imbalanced text classification.

7.2 Performance Comparison Among Computational Spaces

As described earlier, our formulation of the classification problem inspires us to test the performance of hyperbolic space embeddings in the tasks of binary and multi-class text classification tasks. In both cases, euclidean embeddings are better at embedding the text samples in the hidden space. However, hyperbolic variant of our approach still

Model	Macro F1	Sadness	Joy	Love	Anger	Fear	Surprise
Class Ratios		0.292	0.335	0.0815	0.135	0.121	0.0357
SOTA Baselines							
SetConv	0.361	0.425	0.469	0.297	0.314	0.378	0.283
GILE	0.401	0.607	0.675	0.242	0.42	0.325	0.138
BertGCN	0.554	0.712	0.778	0.330	0.571	0.55	0.383
Contrastive Learning Baselines							
SCL	0.285	0.555	0.646	0.0523	0.213	0.243	0.0
KCL(k=5)	0.299	0.508	0.63	0.0971	0.219	0.295	0.047
Our Approach							
HLSCL	0.621	0.757	0.774	0.553	0.597	0.595	0.451
ELSCL	**0.695**	**0.793**	**0.836**	**0.611**	**0.704**	**0.637**	**0.591**

Table 2: This table shows the per class and macro F1 scores achieved by our model on the task of emotion classification. We present both the hyperbolic and euclidean versions of our approach. The best performance numbers have been made bold.

outperforms all the baselines. This is evident from the results in the tables 1 and 2. In the case, of binary classification task, the highest performance difference between models in both spaces is minor, approximately 2% macro F1 score, but this difference increases in the case of multi-class sentiment classification task to approximately 8% macro F1. This shows that Euclidean models are better at the task of imbalanced classification even though hyperbolic models are effective classifiers.

7.3 Analyzing Embedding Space

We train our approach in both euclidean and hyperbolic spaces with 2-dimensional embeddings to visualize how our approach divides the embedding space. We find that hyperbolic variation of our approach divides the space more fairly between the minority and majority class in the binary classification case. This is interesting and may require further investigation in future work, as we fail to observe such a result when it comes to the multi-classification task. This could be because of data characteristics or may point to an innate trait of hyperbolic embeddings.

7.4 Interpreting Model Decisions Using Inter-Class Relationships

As described in the section 4, we proposed an approach to make model decisions interpretable by learning the inter-class relationships in the form of weights between 0 and 1. We train a model with the weighted variation of our approach and results highlight that model tries to distance embeddings which belong to similar emotions more than those belonging to different ones. This is apparent by looking at the weights in the figure 4 which shows that relationship weight between the positive labels *love* and *joy* (0.540) is higher in contrast to the weight between opposite ones *joy* and *sadness* (0.186). Similarly, weight between correlated emotions like *anger* and *surprise* (0.447) is higher than between emotions which are not correlated like *anger* and *love* (0.0558). The is interesting as this shows that model is capturing the fact that some emotions even though not similar are correlated. Another interesting insight is that the relationship between non-opposite categories like *anger* and *surprise* or *surprise* and *joy* are comparatively higher. This may point to an interesting characteristic of the data and alludes the fact that text expressing surprise can both be positive or negative. These results highlight that, along with improving interpretability, our approach can be utilized to highlight data specific characteristics and relationships. These may be used in data modeling or adopting data specific approaches for implementing practical solutions.

8 Limitations and Future Work

Our current approach is limited by the architecture of the label embedding layer. In our current implementation the label embeddings are obtained using a simple embedding which is fine-tuned during training along with text embedding module. In future works, we should experiment with more sophisticated ways to obtain label embeddings to check if we can improve our approach further.

Our approach, specially with hyperbolic embeddings, may have applications in hierarchical classification tasks where classes have a hierarchy and relationships between data samples and their classes are more complex. Such a task may be able to better utilize the natural structure of hyperbolic plane more effectively. In addition, our hyperbolic models fall behind in performance to their Euclidean

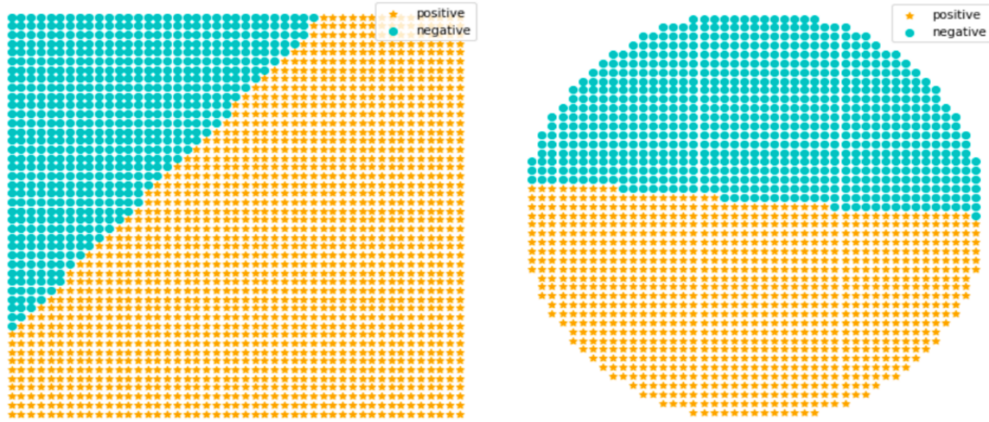

Figure 3: Space division by the 2-dimensional variation of our approach with negative text-samples. The figure shows how our approach divides the embedding space when trained with hyperbolic vs. euclidean embeddings. Rectangular space shows the normalized euclidean space while the circular shows a hyperbolic disk of poincare radius=1.

	Sadness	Joy	Love	Anger	Fear	Surprise
Sadness	--	0.0285	0.1939	0.436	0.1273	0.2144
Joy	0.1855	--	0.1708	0.399	0.0375	0.2072
Love	0.1398	0.5399	--	0.1284	0.1027	0.0893
Anger	0.1142	0.2254	0.0558	--	0.157	0.4475
Fear	0.3165	0.2517	0.1144	0.1889	--	0.1284
Surprise	0.1326	0.5381	0.0569	0.0683	0.204	--

Figure 4: Cells with darker red colors represent that model learns to separate these pairs more.

counterparts so more investigation is needed into how can hyperbolic spaces be used to learn effective classifiers.

Another significant limitation of our approach, lie in the problem formulation. One powerful aspect of CL approaches is that they do not need label information. However, we rely on the presence of label information in the corpus to learn label embeddings. This may not always be possible. In the future, we may be able to combine our approach with traditional CL approaches. This will involve dividing the embedding space during pre-training in the first phase. Using the results from this pre-training, we may be able to obtain key anchors by averaging out the embeddings in a region. These key anchors may then be used in an approach similar to ours to reduce noise in the CL training and better split the embedding space between different distributions in the data.

Finally, weighted variation of our CL objective is successful in quantifying relationship between class pairs. This provides additional insight into how our model is making decisions and improves interpretability. It even helps decipher information which is not obvious without a detailed look at data, like relationship between correlated emotions. However, it does not help in improving the performance. Investigation into how this information can be used to learn better classifiers is another possible venue for future work. Similarly, using this information to design data specific solutions for deployment may offer another avenue for future research.

9 Conclusion

We present a novel CL approach which uses label embeddings as anchors for the task of imbalanced text classification in both the binary and multi-class classification settings. Our approach outperforms several baselines by a margin of 7% in the binary classification task and a margin of 15% in the multi-class classification task. In addition, we extend our approach to hyperbolic spaces, show its effectiveness in the task of imbalanced data classification. We also conduct a study of how our approach utilizes embedding space and show that it may be worth for future investigation that hyperbolic models divide the embedding space in a fairer manner than euclidean couterparts. Finally, we present a interpretable variation of our approach for multi-class classification which helps us draw important conclusions about data relationships.

References

Ashish Anand, Ganesan Pugalenthi, Gary Fogel, and Ponnuthurai Suganthan. 2010. An approach for classification of highly imbalanced data using weighting and undersampling. *Amino acids*, 39:1385–91.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002a. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002b. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. 2021. Probing BERT in hyperbolic spaces. *CoRR*, abs/2104.03869.

Kewen Chen, Zuping Zhang, Jun Long, and Hao Zhang. 2016. Turning from tf-idf to tf-igm for term weighting in text classification. *Expert Systems with Applications*, 66:245–260.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

David Díaz-Vico, Aníbal R. Figueiras-Vidal, and José R. Dorronsoro. 2018. Deep mlps for imbalanced classification. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021a. Simcse: Simple contrastive learning of sentence embeddings.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Gao, Yi-Fan Li, Yu Lin, Charu Aggarwal, and Latifur Khan. 2021c. Setconv: A new approach for learning from imbalanced data.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. 2019. Data-efficient image recognition with contrastive predictive coding.

E.L. Iglesias, A. Seara Vieira, and L. Borrajo. 2013. An hmm-based over-sampling technique to improve text classification. *Expert Systems with Applications*, 40(18):7184–7192.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies*, 9(1).

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. 2021. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020a. Supervised contrastive learning.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020b. Supervised contrastive learning. *CoRR*, abs/2004.11362.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Bartosz Krawczyk. 2016. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. BertGCN: Transductive text classification by combining GNN and BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1456–1462, Online. Association for Computational Linguistics.

Manuel Lopez-Martin, Antonio Sanchez-Esguevillas, Juan Ignacio Arribas, and Belen Carro. 2022. Supervised contrastive learning over prototype-label embeddings for network intrusion detection. *Information Fusion*, 79:200–228.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *CoRR*, abs/1705.08039.

Maximilian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. *CoRR*, abs/1806.03417.

Nikolaos Pappas and James Henderson. 2019. GILE: A generalized input-label embedding for text classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.

Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. 2021. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 735–744.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Michael R. Smith, Tony R. Martinez, and Christophe G. Giraud-Carrier. 2013. An instance level analysis of data complexity. *Machine Learning*, 95:225–256.

Jia Song, Xianglin Huang, Sijun Qin, and Qing Song. 2016. A bi-directional sampling based on k-means method for imbalance text classification. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5.

Muhammad Atif Tahir, Josef Kittler, and Fei Yan. 2012. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10):3738–3750.

Jiachen Tian, Shizhan Chen, Xiaowang Zhang, Zhiyong Feng, Deyi Xiong, Shaojuan Wu, and Chunliu Dou. 2021. Re-embedding difficult samples via mutual information constrained semantically oversampling for imbalanced text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3148–3161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive multiview coding.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. 2022. Label anchored contrastive learning for language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1449, Seattle, United States. Association for Computational Linguistics.