



Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction

Ayaan Haque^{1*}, Viraj Reddi^{1*}, Tyler Giallanza² (*equal contributions)

¹Saratoga High School, ²Princeton University



ICANN 21

Overview

Problem

- Distinguishing between suicide and depression is a challenging and unaddressed task
- Web-scraped data allows for increased data, enabling the use of DNNs
- Data for this task inherently has noisy labels, requiring label correction

Solution: SDCNL

Contributions:

- Neural Network (NN) sentiment analysis for depression vs suicidal ideation classification
- Novel unsupervised, clustering-based label correction process
- Extensive experimentation, ablation on multiple datasets

• **Paper:** <https://arxiv.org/abs/2102.09427>

• **Code:** <https://github.com/ayaanzhaque/SDCNL>

Embedding Models & Classifiers

- We translate the raw text to numerical representations using embedding models. We classify the text by depression or suicide using classifiers.
- **Embedding Models:**
 - BERT — State of the Art, bidirectionally trained transformer
 - Sentence-BERT — Extension of BERT optimized for longer inputs
 - Google Universal Sentence Encoder (GUSE) — transformer optimized for greater-than-word length text
- **Deep Classifiers:** CNN, BiLSTM, GRU
- **Classical Classifiers:** LogReg, MNB, SVM

Datasets

Primary Dataset:

- Used for suicide and depression classification
- Reddit-based dataset web-scraped from subreddits r/SuicideWatch and r/Depression
- Posts from r/SuicideWatch labeled as suicidal; posts from r/Depression labeled as depressed

Reddit Suicide C-SSRS dataset:

- 500 Reddit posts from r/Depression that are clinically labeled by psychologists according to the C-SSR Scale (severity of depression)

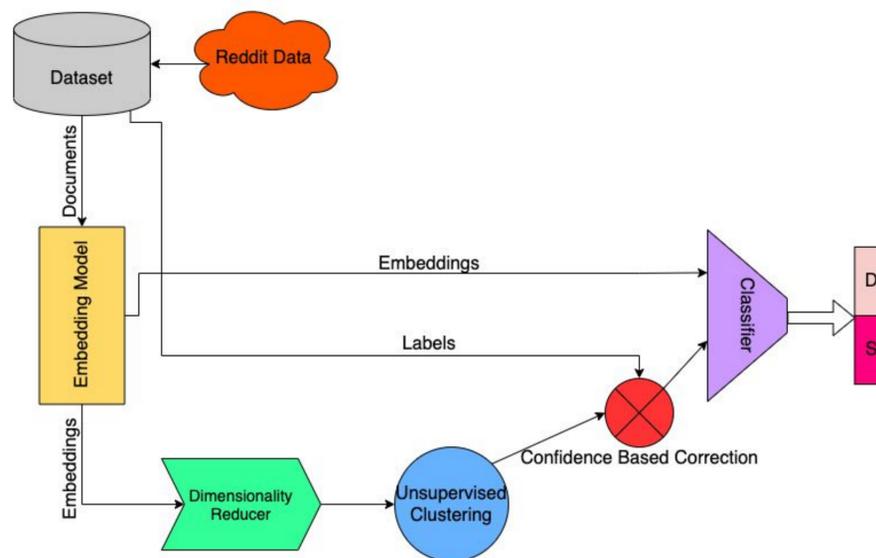
IMDB Dataset:

- Baseline large movie dataset for label correction validation

Conclusion and Ethics

- We present a novel method for deep neural network classification of depressive sentiment vs suicidal ideation with unsupervised noisy label correction
- An applied setting would be as second opinions and a supplementary tool for therapists
- The ethical concern with our research are false negative and positive predictions in eventual applications.

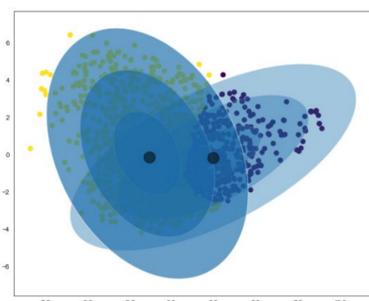
SDCNL



- Data is **web scraped from Reddit** for large amounts of samples
- Data is pre-processed and **fed to a state-of-the-art embedding model**
- To remove noise, embeddings are passed to dimensionality reducing algorithms and then a **clustering algorithm to create new labels**
- Clustering labels are used to correct the noisy labels based on a **thresholding scheme**
- Embeddings and corrected labels are **passed to a classifier** to distinguish between suicidal and depressive sentiment

Label Correction

- **“Curse of Dimensionality”:** Word embeddings have many features, but for clustering algorithms, dimensionality reduction methods are required
- **Unsupervised Clustering:** The reduced embeddings are fed to an unsupervised clustering algorithm and new labels are predicted
- **Correction Scheme:** Labels are corrected using a thresholding scheme: if the confidence of the clustering algorithm is above the threshold τ , then the label is swapped to the clustering label, otherwise it is preserved
- The underlying feature distributions of each class are very similar, making clustering a challenging task, as shown below

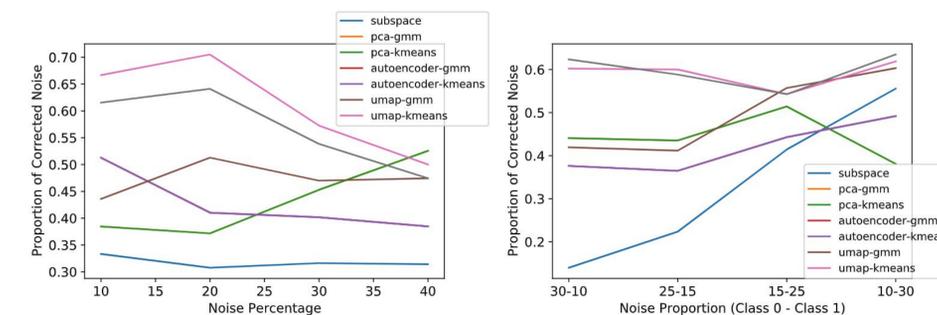


- This graphic uses BERT embeddings and PCA reduction to 2 dimensions
- The graphic shows the difficulty of the clustering task
 - There is little differentiation between the clusters
 - The clusters heavily overlap

Results

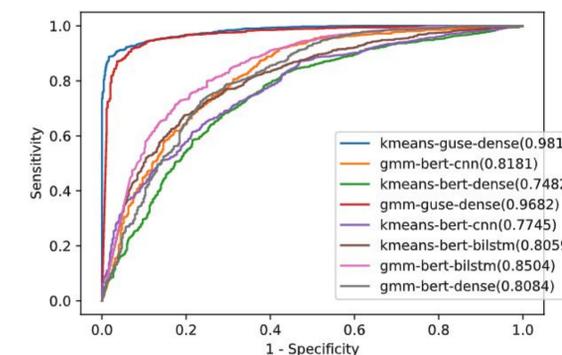
Clustering Performance

- Our noise correction method is able to consistently remove > 50% of injected noise while remaining below a 10% false-correction rate
- The performance does not degrade heavily at higher noise percentages, which is challenging to achieve



Classification Performance after Label Correction

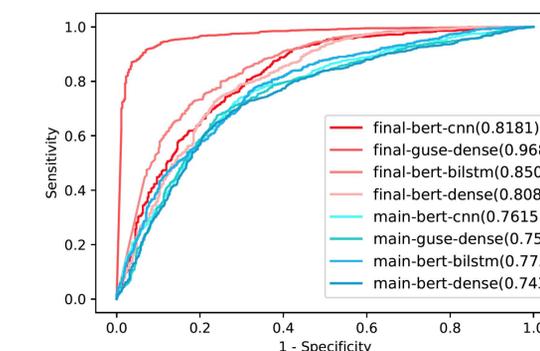
- We compare the classification accuracy of our model on uncorrected labels versus labels corrected using the label correction method.



Classification improved by 10-20% for every model proposed. The 4 best combinations of models are shown with the 2 final label correction methods (GMM vs. K-Means).

Classification Performance

- We found that a combination of GUSE with a Dense NN was the best for our proposed task of suicide vs. depression classification



ROC curves of performance of top 4 models with label correction (red) against the same models without label correction (blue).