Investigating the efficacy of Semi-Supervised learning to generate pseudo labels using US Elections Data

Muhammad Abu Bakar Aziz

Abstract

In this project, I classify YouTube videos into voter fraud and videos containing information about the US elections/voter fraud using a novel technique similar to pseudo-labeling. Pseudo labeling utilizes semi-supervised learning methods such as k-means clustering. First, on the pre-processed voter fraud dataset, I generated word embeddings using different sentence transformer models. Then, I applied k-means clustering using word embeddings to generate 200 clusters. To automatically label the clusters, I used pseudo-labeling. To test how effectively the clusters Ire labeled, I trained the data on neural-based state-of-the-art transformer models such as BERT and RoBERTa. I demonstrate that my best model (RoBERTa) achieved up to 83% Accuracy-Precision Area (AUPRC). My results validate the pseudo-labeling technique.

1 Introduction

Data is essential for any kind of Natural Language Processing, Deep Learning, and Machine Learning tasks. These tasks require labeled data. Labeling data is often a time-consuming and expensive task [1]. Even in the labeled data, the inter-annotator agreement betIen the annotators may not be very high. Hence, there has been a need to efficiently and quickly label data in a cost-effective way.

Recently, semi-supervised learning is becoming quite prevalent. Semi-supervised learning can be used to reduce the need for manually labeled data by allowing a model to train a small set of data to predict labels for an unlabeled dataset[2]. For our project, similar to the semi-supervised learning, I propose a novel technique using clustering to generate pseudo labels for unlabelled data using the voter fraud dataset[3]. Our significant contribution in this project is:

- Building a pipeline to preprocess voter fraud data
- Using state of the art sentence transformers to build word embeddings.
- Using k-means to generate clusters using word embeddings

- Proposing a novel technique for labelling clusters using euclidean distance.
- Evaluating our labelled data of clusters on state-of-the-art neural based models and achieving up to 83% Precision-Recall Curve Area (AUPRC) .

I used state-of-the arts sentence transformers[4] to generate word embeddings for our clusters. The clusters Ire labelled using novel technique similar to pseudo labels. I tested the labels using state-of-the arts neural based transformer models such as BERT. Our analysis shoId that models trained on pseudo labels accurately predict labels with Accuracy-to-Precision Area as high as 83 %. I believe voter fraud dataset can be labelled using pseudo labels efficiently and effectively if cost and time are a constraint. The project data and code are available on GitLab. Furthermore, individual contribution of this project is mentioned at the end of the report.

2 Methodology to Generate Clusters

2.1 Dataset

I used VoterFraud2020 [3] data which contains multi-modal dataset of 7.6M tIets and 25.6M retIets from 2.6M users related to voter fraud claims. Within those tIets, it consisted of all the URLs, images, and YouTube videos that people shared about the voter fraud. For our project purpose, I focused on YouTube videos that appeared within those tIets.
In the tIets there Ire around 12k voter fraud YouTube videos. These videos Ire present as a separate csv file on VoterFraud2020. These video data contain many different meta data about YouTube video such as title, date, and video description.
Our project was mainly focused on the titles that appeared on the YouTube videos. Since our project was focused on the voter fraud, I searched for those YouTube titles that included either promoting voter fraud or videos related to the US elections/voter fraud. To find these type of videos, I searched a set of common voter fraud keywords that are mentioned in the appendix section at the end of the report. After searching for these keywords, I have around 7358 YouTube video titles.

2.2 Clustering

Recently, there has been extensive use of state-of-the-art sentence transformers[4]. Transformer models are based on the self-attention mechanism. I used different models to generate word embeddings for our titles. These word-embeddings are vector representations of all the titles that capture their semantic meanings. After experimenting with different word embeddings, I selected the state-of-the-art paraphrase-distill Roberta-base sentence transformer from hugging face library[5] to generate word embeddings . Then, these word embeddings Ire used as inputs to k-means clustering. K-means clustering is an unsupervised technique that tries to group the datasets into k-clusters. K defines the number of centroids; the K-means tries to find and add data points nearest to each cluster. The algorithm is described in detail in this paper[6]. After some experimentation and seeing the generated clusters, I selected the k-value to be 200. Hence, our video title dataset was divided into 200 clusters with the assumption that the titles within the same clusters would have the same semantics.

3 Cluster Evaluation

3.1 Data labelling

To evaluate the effectiveness of the generated 200 clusters, I decided to pseudo-label our datasets. As I wanted to detect voter fraud videos from the given dataset, I used two labels: Videos Promoting Voter fraud (label 0) and Videos containing information about Elections/VoterFraud(1). Since labelling a large data set of about 7k titles for a single person over a short period was time-consuming and not feasible, I decided to use a novel technique which is similar to pseudo labelling to label clusters using euclidean distance. Euclidean distance is defined as the square root of the sum of squared differences betIen corresponding elements (titles) of the two vectors[7]. Pseudo labelling process is describe in detail in Figure 1

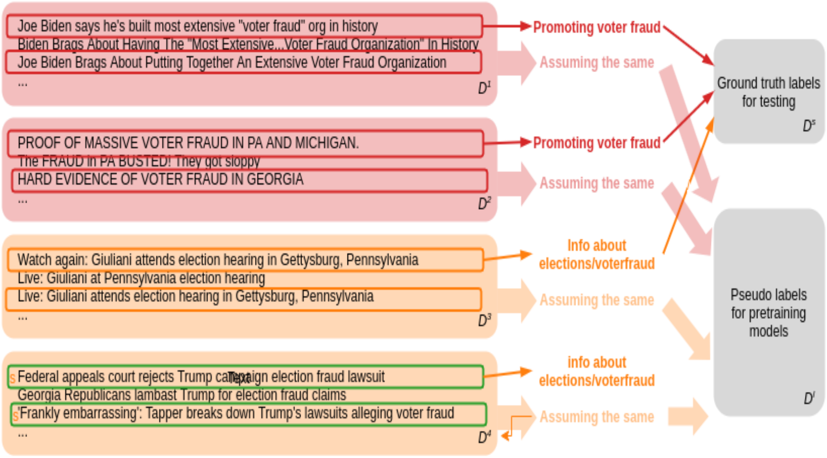
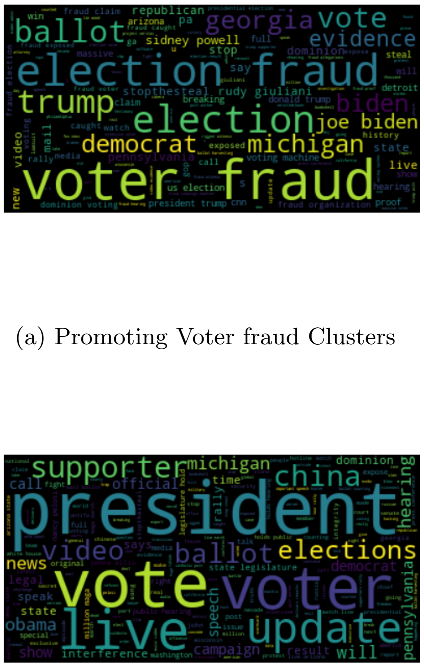

Joe Biden says he's built most extensive "voter fraud" org in history
Biden Brags About Having The "Most Extensive...Voter Fraud Organization" In History
Joe Biden Brags About Putting Together An Extensive Voter Fraud Organization
...
D^1
Promoting voter fraud
Assuming the same
PROOF OF MASSIVE VOTER FRAUD IN PA AND MICHIGAN.
The FRAUD in PA BUSTED! They got sloppy
HARD EVIDENCE OF VOTER FRAUD IN GEORGIA
...
D^2
Promoting voter fraud
Assuming the same
Watch again: Giuliani attends election hearing in Gettysburg, Pennsylvania
Live: Giuliani at Pennsylvania election hearing
Live: Giuliani attends election hearing in Gettysburg, Pennsylvania
...
D^3
Info about elections/voterfraud
Assuming the same
Federal appeals court rejects Trump campaign election fraud lawsuit
Text
Georgia Republicans lambast Trump for election fraud claims
'Frankly embarrassing': Tapper breaks down Trump's lawsuits alleging voter fraud
...
D^4
info about elections/voterfraud
Assuming the same
Ground truth labels for testing
D^s
Pseudo labels for pretraining models
D^l

Figure 1: Pseudo Labelling Process

The first step in our labeling process consisted of picking up two video titles that Ire farthest apart based on the Euclidean Distance. If both of the titles agreed on the same labels, then I labeled the whole clusters with the same label. Otherwise, I randomly sampled 3-5 titles from the same cluster and gave that label to the complete cluster on which the majority of the titles agreed upon.

In total, out of 200 clusters, 152 clusters Ire successfully labeled using euclidean distance and the remaining Ire labeled using random sampling. I kept both the datasets separately to investigate if random sampling introduced noisy data or not. The labeling process was carried out by a single person and it took 3 days to label both the datasets.



(a) Promoting Voter fraud Clusters

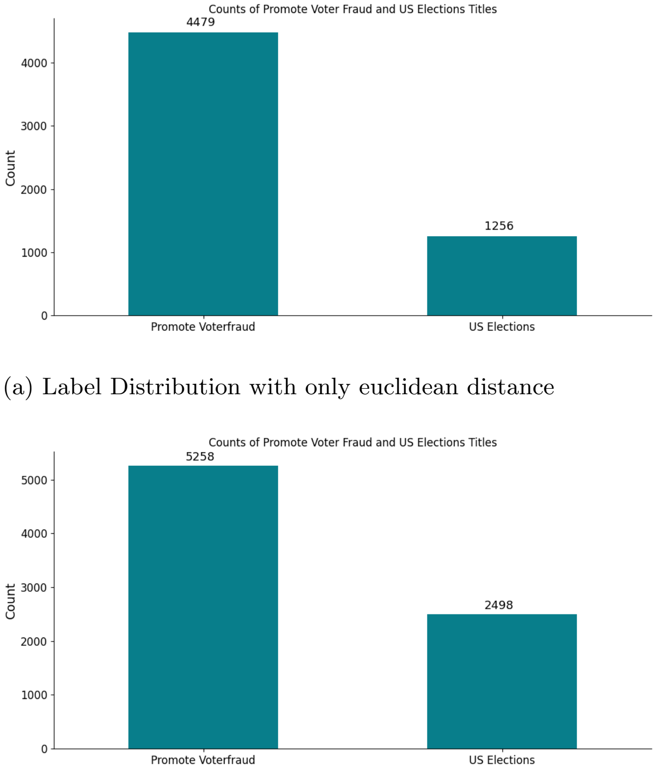
(b) Election/voter fraud cluster

Figure 2: Word Cloud of Labels

I also carried out the word analysis of the generated clusters after excluding the most common same words in both clusters. Figure 2(a) shows a word cloud of the clusters that promoted voter fraud and Figure 2(b) shows a word cloud of clusters that contain information about elections/voter-fraud. Promoting Voter fraud Clusters contain many of the expected words that usually appear on the videos that promote voter fraud such as ”fraud”, ”election fraud”, ”rudy giulani”, ” sidney powIll”. Furthermore, figure 1(b) shows a cluster of words that contain information about elections/voter fraud. These clusters show general words about the US elections such as ”president”, ”voter”, ”president”.

3.2 Training Data

Overall, I have two datasets. One which I labelled only using euclidean distance (data 1). Second which contained remaining clusters which Ire randomly sampled (data 2).



(a) Label Distribution with only euclidean distance

(b) Label Distribution with only euclidean distance and random sampling

Figure 3: Label Distribution

In data 1 as shown in figure 3 (a) , out of 5734 video titles for training, 4479 (78 %) of the video titles Ire labelled as promoting voter fraud and 1256 (21%) Ire labelled as information about voter fraud/elections. In data 2 as shown in figure 3(b) out of 7756 video titles for training, 5258 (67.8%) of the titles Ire labelled as promoting voter fraud and 32% Ire labelled as titles about voter fraud or US-elections. Our testing set consisted of all the manually labelled titles. For data 1 I have around 304 manually labelled titles for testing and for data 2 I have around 400 titles for testing. Moreover, our testing dataset was not part of the training dataset.

3.3 Experiments

For testing whether our clusters correctly predicted labels on our testing dataset, I experimented with 5 pre-trained state-of-the-art neural models: : BERT[8], RoBERTa [9], XLNet[10], Albert[11], DistilBert [12]. These are pre-trained models that are trained on large, unlabeled corpus with specific objectives [13]. During fine-tuning, these same models are initialized with pre-trained Iights and further trained on labeled dataset Since these models Ire already pre-trained on large datasets, so they train very Ill on small datasets to give better results. For our project, I used simple transformer pretrained models[5] available on Hugging-Face Library[5] which provides all the pre-trained models. Our neural models used the same neural architecture, hyper-parameters and tokenizers as the base models, and I trained them for 2 epochs. I fed our preprocessed data into those models. The training time took about two hours. I trained our model on a single RTX A6000.

3.4 Metrics

I fed all the titles into our neural-based models. For every title, the models predicted whether the title was promoting voter fraud (label 0) or whether the title contained information about voter fraud/US elections (label 1). I cannot simply calculate accuracy to evaluate our models since I have imbalance classes. For example, for our dataset 1, (78%) of the video titles Ire labeled as promoting voter fraud and 1256 (21%) of the videos Ire labeled as information about voter fraud/elections. To counter class imbalance, I calculated the area under the precision-recall curves, and I specifically made minority class a positive label.

3.5 Baseline

To establish a baseline, I trained our best classifier Roberta on testing dataset 2. Testing dataset 2 contained around 400 manually labeled examples. I further randomly split this dateset into 90% (360) training set and 10% (40) testing set. Then, I carried out three experiments with 2 epochs. The average AUPRC for three experiments was 43 %.

3.6 Results

The results of our training and evaluation are shown in 4. On Figure 4, on the x-axis are different models (Bert, Roberta, Albert, Distilbert, Xlnet) and on the y-axis is the area of precision-recall curve (AUPRC). Bert and Roberta classifiers are the best performing with AUPRC 82% respectively. and 77%. Compared to our baseline AUPRC (43%), all our models performed better. This demonstrates that our pseudo-labeling process can be effectively used to label our clusters. Our results also shows that our clusters can differentiate betIen YouTube videos promoting voter fraud and YouTube videos about voter fraud/elections.

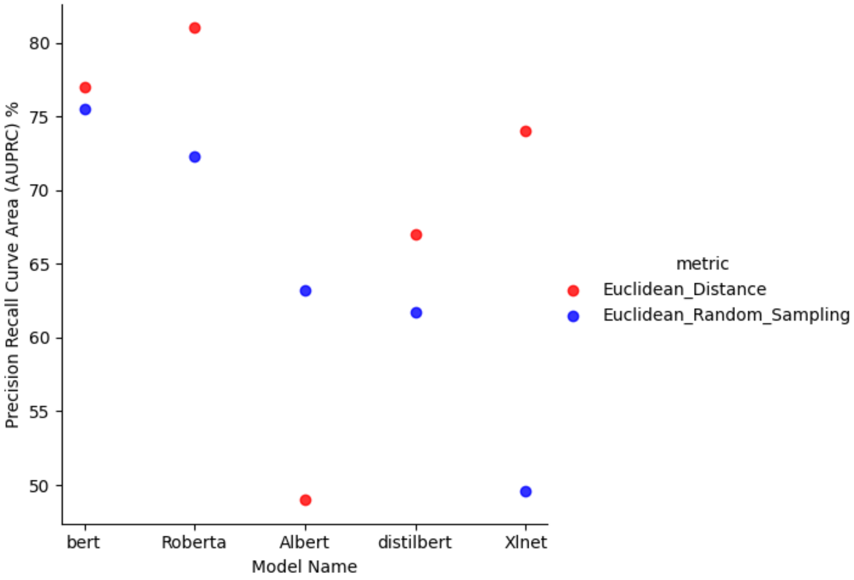


Figure 4: Area under the Precision-Recall Curve of five classifiers evaluated on two testing sets

Moreover, figure 4 also shows our results of the two datasets: one in which I only labeled data using Euclidean distance and the second using euclidean distance and random sampling. Our results show that for four classifiers (Bert, Roberta, distilbert, Xlnet) euclidean distance performed better than euclidean distance and random sampling. This may be because random sampling might have introduced more noisy data in the clusters. Hence, those clusters might not have video titles belonging to the same semantics. Overall, one classifier(Albert) had the least AUPRC better on both euclidean distance, and euclidean distance with random sampling. This may be because Albert is a light version of BERT and has a few parameters. Hence, in our case, Albert might not learnt very Ill on our limited voter fraud dataset.

4 Conclusion/Inference

From our analysis, I can conclude that pseudo labeling using clusters can have a high AUPRC on different neural-based classifiers which can differentiate betIen Youtube videos promoting voter fraud and Youtube videos containing information about the US elections. This effectively demonstrates that our youtube videos in the same clusters have nearly similar semantics videos.

Moreover, for future work, it could be useful to incorporate more features from the youtube videos such as youtube video description and transcripts and see if these new features help to make classifiers perform any better or not. I could also explore several questions that a social computer scientist would like to know. One such question is to see if there is a time effect on the generated clusters. For example, if the videos in the same clusters appear around the same time.

Furthermore, there are also several limitations of our project. In the data-labeling process, only one person labeled the whole data process. This could be improved in the future by carrying out the labeling process by more than one person. Furthermore, our dataset size is limited. The dataset may be increased by finding more voter fraud videos online.

5 Individual Contribution

Muhammad Abu Bakar Aziz : He worked on setting up the data pipeline, pre-preprocessed data, worked on generating different clusters, carried out data labeling process, ran different models and experiments, carried out the complete analysis, plotted different graphs, gave presentation, and wrote the complete report.

Shan Jiang: He is a Phd student and not part of the course. He helped with the pseudo-labelling idea, helped in setting up the data pipeline, and was available for project discussions.

References

[1] X. Zhou and M. Belkin, "Semi-supervised learning," in *Academic Press Library in Signal Processing.* Elsevier, 2014, vol. 1, pp. 1239–1269.

[2] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.

[3] A. Abilov, Y. Hua, H. Matatov, O. Amir, and M. Naaman, "Voterfraud2020: a multi-modal dataset of election fraud claims on twitter," *arXiv preprint arXiv:2101.08210*, 2021.

[4] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[6] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[7] W. H. Gomaa, A. A. Fahmy *et al.*, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013.

[8] 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.

[9] 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.

[10] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[11] 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.

[12] 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.

[13] S. Jiang, M. Metzger, A. Flanagin, and C. Wilson, "Modeling and measuring expressed (dis) belief in (mis) information," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 315–326.

A Appendix

Keyword used to look for voter fraud videos:

Vote	elect	mail
congress	expos	fraud
steal	stole	rig
rig	cheat	discard
donald	trump	joe
biden	ilhan	pelosi
obama	democrat	republican
gop	qproof	qanon