Twitter User Embedding and Clustering - Final Report

Ian Argyle
Bryce Gillespie
Janaan Lake

1 INTRODUCTION

People use social media to share ideas, opinions, preferences, and more. Twitter is one of the most popular social media platforms and has over 330 million active users. Twitter users can share their thoughts and ideas by tweeting, which is a post of up to 280 characters. Previous research has shown the ability to represent Twitter users in low-dimensional embeddings using a document representation model [2]. These user embeddings tend to be semantically meaningful such that users with similar interests tend to be close in the embedding space. These embeddings could be useful to segment users by interest, find nearest neighbors, etc. Many techniques are used to create user embeddings. We examine the technique explored in [2], which is creating user embeddings by averaging the the document embedding of each tweet belonging to each different user. We expand upon this idea to see if other clustering techniques can better represent the data. Specifically, we clustered each user's tweets and found the centroids of each cluster. These centroids were then used as the user's embedding and compared to the results above. Additionally we cluster the data using the best silhouette score and compare these "natural" clusters to those created by the categories of the labelled tweets.

2 DATASET

We used the dataset that was created and used in [2]. This dataset contains 500k total tweets and 500 users categorized by interest into five categories: economy, cryptocurrency, technology, politics and fashion. The dataset was categorized by user, rather than by individual tweets. The dataset can be found at [1]. In hindsight, it would be better to apply our technique to individually labeled tweets, however we were unable to find such a dataset.

3 TECHNIQUES USED AND RESULTS OBTAINED

Since our dataset contained labels, we used them to gain insight into the data and to compare different embedding and clustering techniques to the ground-truth labels. First, we created an embedding for each user's tweets by using a pre-trained SBERT embedding model. Sentence-BERT (SBERT) is a modification of the BERT (Bidirectional Encoder Representations from Transformers) network using siamese and triplet networks that is able to derive semantically meaningful sentence embeddings [3]. This means that similar sentences are close in vector space.

Next, we averaged the embeddings of each user and then plotted them using the t-SNE dimensionality-reduction technique. The results are shown in Figure 1. Each dot represents a user embedding (average of the tweet embeddings), color coded by the user's labeled category. The embeddings appear to capture semantically similar tweets. The categories of fashion and politics are the most separable, while the categories of economy, crypto and technology have more overlap in their embeddings.

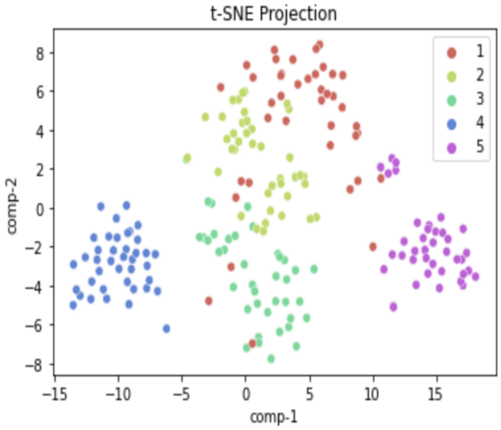

Figure 1: User Embeddings by Category

Next, we used the labels to measure the accuracy of the embeddings by comparing how many of the datapoints (average user embeddings) in each category are closest to the centroid of that category. To calculate the centroid of each category, all of the tweet embeddings for that category were averaged. For example, all of the tweet embeddings with the label of politics were averaged to calculate the centroid for the politics category. Then all of the embeddings associated with users who had a politics label were measured to see if the politics centroid was closer than any other category centroid to that user. The accuracy measures are reported in Table 1 below:

Category	**Accuracy**
1 - Economy	82.5%
2 - Crypto	90.0%
3 - Technology	95.0%
4 - Fashion	100.0%
5 - Politics	100.0%
Average	**93.5**

Table 1: Accuracy of User Embeddings using SBERT

To gain more insight into the separability of the labeled data, we calculated the similarity matrix across categories. This was done using the cosine similarity score between the centroids of different categories. The results are shown in Table 2.

The accuracy and similarity results confirm the empirical results of Figure 1. The economy, crypto and technology categories are

	Economy	Crypto	Technology	Fashion	Politics
Economy	1.00	0.86	0.79	0.58	0.70
Crypto		1.00	0.83	0.62	0.58
Technology			1.00	0.62	0.64
Fashion				1.00	0.47
Politics					1.00

Table 2: Similarity Matrix for Category Centroids

more similar and less separable. The politics and fashion categories form more distinct clusters and are less similar than any other categories. This intuitively makes sense because the economy, crypto and technology fields can intersect much more often than perhaps politics and fashion.

Next, we tried using a slightly different representation for each user. First, we clustered all of the embedded tweets for each user using an Agglomerative Clustering algorithm. We used the maximum silhouette score to determine the number of clusters for the algorithm. A graph of the sillhoutte scores is shown below in Figure 2. The best number of clusters for most users' tweets was between 3 and 4, with a steep decline after that. Again, these results make sense based on the visual representation in Figure 1.

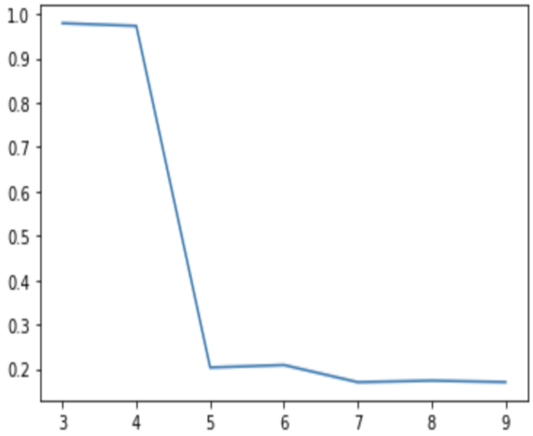

Figure 2: Sillhoutte Scores for Clustering of Users

Then for each user we found the centroids for each cluster. We then used the Agglomerative Clustering technique using the best silhoutte score to cluster all the centroids for each account. The results are shown in Figure 3. Surprisingly, the best number of clusters is eight. This result exposes the noise or variation in the data since the best number of clusters is eight rather than five. It also highlights the challenges of clustering tweets due to the variety and complexity of tweets even within each category. The similarity matrix for each cluster was calculated and the results are shown in Table 3. As can be seen from both the graph and similarity matrix, the data is somewhat separable but many of the clusters are fairly similar. Under this method, though, the data appears less separable than the using average embeddings as the user representation.

For the next method, rather than using all of the centroids for the user representation, we calculated the biggest centroid for each user and used that as the user representation. The biggest centroid was essentially the centroid that had the most datapoints associated with it. As we did with all of the centroids, we used an Agglomerative Clustering algorithm to cluster the biggest centroids. The clustering results are shown in Figure 4. Comparing this to the results in Figure 3, the largest centroids are much more easily separable and the dataset is best separated into three categories rather than five.

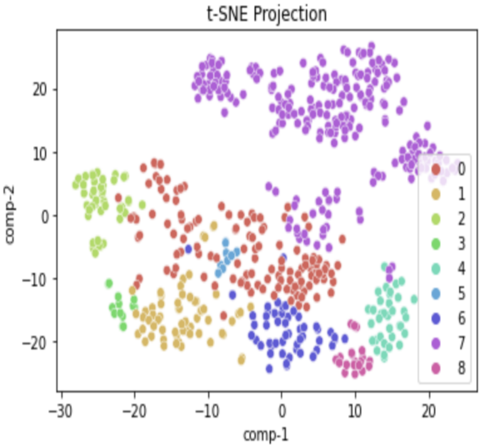

Figure 3: Agglomerative Clustering Using All Centroids as User Representation

	1	2	3	4	5	6	7	8	9
1	1.00	0.80	0.74	0.63	0.54	0.75	0.91	0.89	0.69
2		1.00	0.50	0.81	0.44	0.66	0.77	0.83	0.54
3			1.00	0.57	0.30	0.44	0.58	0.58	0.45
4				1.00	0.26	0.53	0.49	0.61	0.31
5					1.00	0.29	0.66	0.66	0.89
6						1.00	0.58	0.66	0.36
7							1.00	0.92	0.84
8								1.00	0.79
9									1.00

Table 3: Similarity Scores for Users' Centroids

The similarity matrix for these clusters is shown in Table 4. The similarity score is relatively low, which highlights the good separability among the three clusters.

	1	2	3
1	1.00	0.65	0.69
2		1.00	0.482
3			1.00

Table 4: Similarity Scores for Biggest Users' Centroids

Lastly, we wanted to explore the results of using the biggest centroid as the user representation with the ground-truth labels. A graph of this representation and their respective labels is shown in Figure 5, and a list of the accuracy scores are found in Table 5. The accuracy scores are not as high as they are for the SBERT embeddings reported in Table 1. A similarity matrix for the categories based on the cosine similarity score between the centroids of

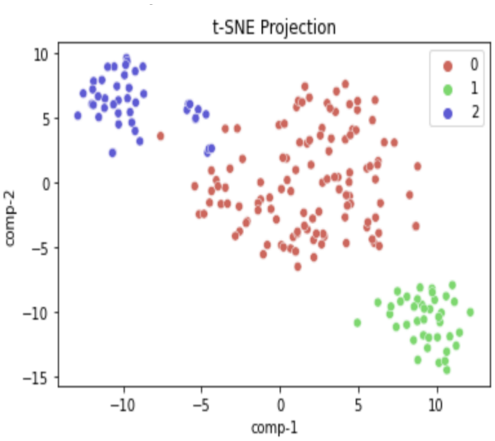

Figure 4: Agglomerative Clustering Using Biggest Centroids as User Representation

the different categories is reported in Table 6. Again, the similarity scores are higher for this method than for those using the SBERT embeddings shown in Table 2.

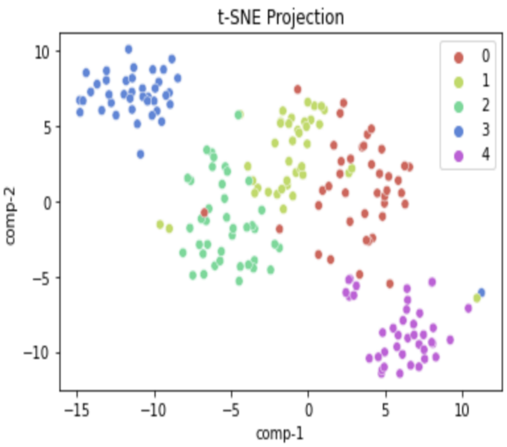

Figure 5: Biggest Centroids by Category

Category	Accuracy
1 - Economy	72.5%
2 - Crypto	82.5%
3 - Technology	85.0%
4 - Fashion	97.5%
5 - Politics	95.0%
Average	86.5

Table 5: Accuracy of User Embeddings using SBERT

	Economy	Crypto	Technology	Fashion	Politics
Economy	1.00	0.93	0.78	0.85	0.70
Crypto		1.00	0.91	0.91	0.67
Technology			1.00	0.86	0.65
Fashion				1.00	0.66
Politics					1.00

Table 6: Similarity Matrix for Category Centroids

4 ANALYSIS AND SUMMARY

This project was an exercise to explore and analyze users' tweets. The dataset was labelled, with each user assigned to one of five categories. Having a labelled dataset allowed us to compare different clustering techniques to a ground truth. Using the SBERT embeddings as our user representation produced the most accurate results when compared to the labels. As can be seen in Figure 1, some categories are much more separable than other categories. It makes sense that categories that are somewhat semantically similar and have more topic overlap are harder to cluster and separate. In this dataset, the categories of economics, crypto and technology are more similar semantically based on the users' tweet embeddings than the categories of politics and fashion. Other methods could be explored to determine how to better separate categories that are somewhat semantically similar.

Our analysis of the individual users' tweets produced some surprising results. Even though each user was assigned a category, their tweets didn't naturally fall into one cluster. In fact, the individual tweets displayed a lot of natural variety, and based upon silhouette scores best clustered into eight separate groups as seen in Figure 3. We attempted to reduce the noise and/or variation in the tweets and used only the biggest cluster centroid for each user. Again, rather than naturally clustering into five categories, the ideal number was three as shown in Figure 4. The accuracy of using this technique actually decreased as compared to using the average embeddings (See Tables 1 and 5). While somewhat perplexing, these results highlight the very complex nature of natural language processing and the challenges of clustering language datasets.

REFERENCES

[1] 2022. Dataset. http://buyukveri.firat.edu.tr/wp-content/uploads/2020/09/test_users.zip. Accessed: 2022-02-28.

[2] Makinist S. Ay B. Hallac, I. and G. Aydin. 2019. user2Vec: Social Media User Representation Based on Distributed Document Embeddings. *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)* (2019). https://doi.org/10.1109/idap.2019.8875952

[3] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084