Big Data Analytics: Project 2

Adhiraj Sood
Rochester Institute of Technology
as9125@rit.edu

Shweta Nateshan Iyer
Rochester Institute of Technology
si9808@rit.edu

Sidhartha Amperyani
Rochester Institute of Technology
ga5310@rit.edu

1 Introduction

In this project, we have explored label distributions. The label information is fetched from JSON file and a graph is created from the label distributions where there is an edge between 2 distributions when from a 10-item sample of 1 distribution, the second distribution is one of the 95 percent most likely distributions. For the purpose of computing the Cohen Kappa interannotator agreement, JobQ3a sheet has been selected. For each label in the sheet, a Cohen Kappa score has been calculated along with a Confusion Matrix to visualize the accuracy of out classification. Section 4 includes the discussion between 2 team members regarding a few tweets they disagreed on. For the purpose of clustering, we have implemented Agglomerative Hierarchical Clustering. In this clustering method, we begin with every node as a cluster and then find he two closest points and combine them into a cluster and so on [5].

2 Methods

The methods used for analysis of the graph are:

- number_of_nodes(): This function returns the number of nodes in the graph.
- number_of_edges(): This function returns the number of edges in the graph
- number_connected_components(): It returns the number of connected components in the graph which is given as a parameter in the function.
- connected_components(): This method returns multiple sets of nodes, one for each connected component. The graph is given as a parameter to this function
- density(): This networkx classes function returns the density of the graph which is given as a parameter to this function.
- is_directed(): Graph is given as a parameter to this networkx function. It returns true if the graph is directed.
- algorithms.cluster.clustering():
  This networkx function computes the clustering coefficient of each node.
  Since this is an undirected unweighted graph, the formula for calculating clustering coefficient of a node is given in 1 [1, 7].

$$c_u = \frac{2T(u)}{deg(u)(deg(u)-1)}$$

Figure 1. Clustering Coefficient of a node u.

  Here, T(u) denotes the number of triangles through node u and deg(u) is the degree of node u [1].
  The parameters of this function are:
  - G: the graph
  - nodes: This is an optional parameter which is for the container of nodes that the clustering coefficient should be computed for.
  - weight: This is an optional parameter which is a numerical value representing the weight of the edge.
- degree_histogram: It returns a list of frequency of each degree value. This is useful when we have to draw a histogram of degrees in the graph

The methods used for Cohen Kappa are:

- confusion_matrix():
  It helps in evaluating accuracy of a classification by computing confusion matrix where each cell represents the number of observations known to be in group i and predicted to be in group j i.e., the matrix would give a count of true positives, true negatives, false positives and false negatives [3].
  Parameters that we have given to the method are:
  - y_true: This is an array representing the ground truth i.e., correct target values.
  - y_pred: This is an array representing estimated targets which are returned by a classifier.
- ConfusionMatrixDisplay(): This function from sklearn.metrics helps in visualizing the confusion matrix.
  Parameters used from this method are:
  - confusion_matrix: The confusion matrix
  - display_labels: An array of display labels for the plot. If labels are not provided, then display labels are set from 0 to number of classes - 1.
- cohen_kappa_score(): It measures the Cohen Kappa inter-annotator agreement which is a score that expresses the level of agreement between two annotators.
  The formula for calculating the score is in Figure 2 [2], where Po denotes the observed relative agreement among raters, whereas Pe denotes the hypothetical probability of chance agreement [6].

$$\kappa = (p_o - p_e)/(1 - p_e)$$

Figure 2. Formula for Cohen Kappa Score.

Agglomerative Hierarchical Clustering - AgglomerativeClustering():
The library used here is sklearn.cluster.
The hyperparameters we have considered are affinity and linkage. Our function call is AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='ward')
This function provides the following parameters and hyperparameters [4]:

- n_clutser: This represents the number of clusters we have to find.
- affinity: This parameters specifies the metric that is used to compute the linkage. The value can be 'euclidean', 'l1', 'l2', 'manhattan', 'cosine'.
- memory: If str or object joblib.Memory interface is specified then it is used to cache the output of the computation of the tree.
- connectivity: This is used to specify the connectivity matrix
- distance_threshold: It specifies a linkage distance threshold. Clusters will not be combined if their linkage distance exceeds a certain threshold.
- complete_full_tree: At n_clusters, the tree development is halted. If the number of clusters is large in comparison to the number of samples, this is a good way to cut down on computing time.
- linkage: This hyperparameter is used to specify which linkage criteria to use. The algorithm then merges pairs to minimaize this criteria. The types of linkage are:
  - ward: This minimizes the variance of clusters that are being merged [4].
  - average: Utilizes the average distance of each observation [4].
  - complete: The maximum distance between all observations of the two sets is used. Also known as 'maximum' linkage.
  - single: The smallest distance between all observations of the two sets is used.

3 Results

Analysis of the graph created using pldl program:

Total number of nodes: 757
Total number of edges: 12462
Number of Connected Components: 2
Connected Component 1 with size: 755
Connected Component 2 with size: 2
Density of the graph: 0.04355119414564593
Average Degree: 16.462351387054163
Is the graph directed?: False
Average Clustering Coefficient: 0.4683566638015865

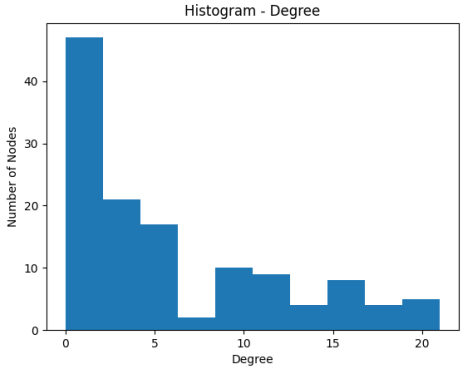

Figure 3. Histogram of the degree.

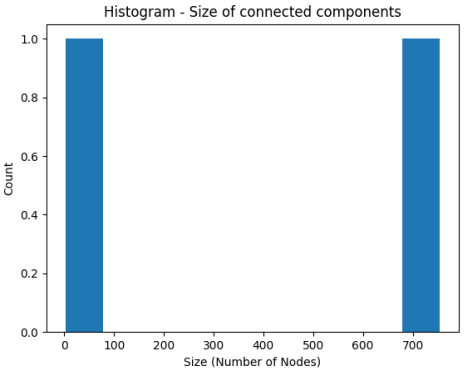

Figure 4. Histogram of size of connected components.

Figure 3 refers to a histogram of the degree plotted by using the networkx function degree_histogram() and Figure 4 refers to the histogram of the size of connected components. Since, there are 2 connected components, we see 2 bars in the histogram, one at 2 and one at 755. These are for the graph build by the code provided by professor.

Analysis of the graph created using the additional labels program:

Total number of nodes: 690
Total number of edges: 5500
Number of Connected Components: 4
Connected Component 1 with size: 676
Connected Component 2 with size: 4
Connected Component 3 with size: 8
Connected Component 4 with size: 2
Density of the graph: 0.0231379230558886
Average Degree: 7.971014492753623
Is the graph directed?: False
Average Clustering Coefficient: 0.0

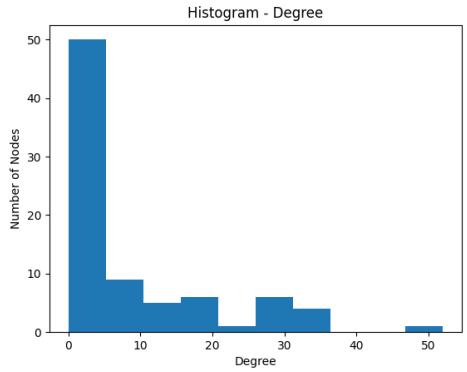

Figure 5. Histogram of the degree by additional labels.

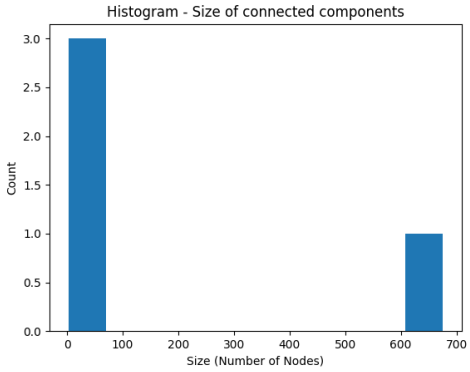

Figure 6. Histogram of size of connected components by adding additional labels.

Figure 5 refers to a histogram of the degree plotted by using the networkx function degree_histogram() and Figure 6 refers to the histogram of the size of connected components. Since, there are 2 connected components, we see 2 bars in the histogram, one at 2 and one at 755. These are for the graph build by adding the additional labels.

Cohen's kappa interannotator has been computed for JobQ3a spreadsheet between Shweta and Adhiraj. The Cohen Kappa Score has been computed for each of the 12 labels.

Label Number: 1, Cohen Kappa Score: 0.7434691745036572
Label Number: 2, Cohen Kappa Score: 0.7229499946624374
Label Number: 3, Cohen Kappa Score: 0.18190728217191032
Label Number: 4, Cohen Kappa Score: 0.7990176013098649
Label Number: 5, Cohen Kappa Score: 0.8403598135905495
Label Number: 6, Cohen Kappa Score: 0.3940765117235706
Label Number: 7, Cohen Kappa Score: 0.6190141885900089
Label Number: 8, Cohen Kappa Score: 0.1563573883161512
Label Number: 9, Cohen Kappa Score: 0.5472370247420479
Label Number: 10, Cohen Kappa Score: 0.7890544961829538
Label Number: 11, Cohen Kappa Score: 0.16559667673716016
Label Number: 12, Cohen Kappa Score: 0.7459249676584735

Word	Frequency
work	96
someon	64
today	17
like	12
call	12
time	10
want	9
good	8
manag	7
gonna	7

Table 1. Most Frequent words in Cluster 1.

Word	Frequency
someon	99
work	82
http	25
like	16
today	11
link	9
come	9
love	7
make	7
best	6

Table 2. Most Frequent words in Cluster 2.

Word	Frequency
work	102
someon	27
today	9
readi	7
come	7
hour	7
time	7
like	7
right	6
tomorrow	6

Table 3. Most Frequent words in Cluster 3.

We have 3 clusters and the most frequent words in each cluster are given in Table 1, 2 and 3 respectively.

4 Discussion

Some tweets we disagreed on were

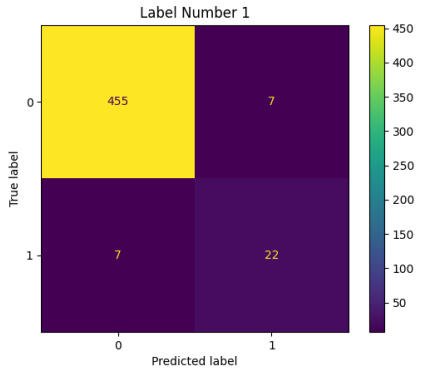

Figure 7. Confusion Matrix for Label 1.

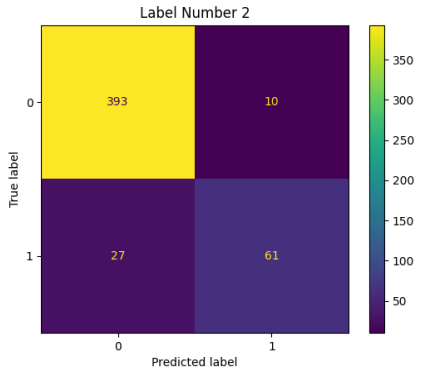

Figure 8. Confusion Matrix for Label 2.

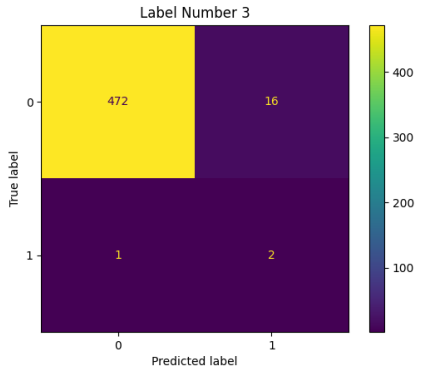

Figure 9. Confusion Matrix for Label 3.

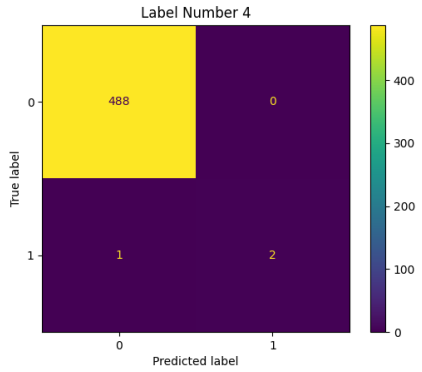

Figure 10. Confusion Matrix for Label 4.

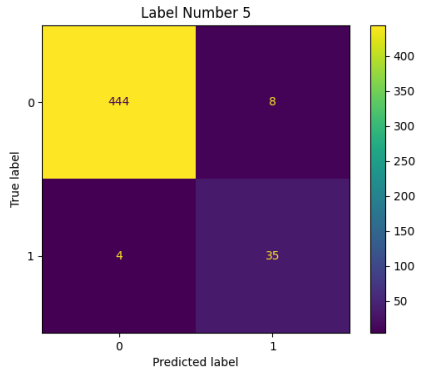

Figure 11. Confusion Matrix for Label 5.

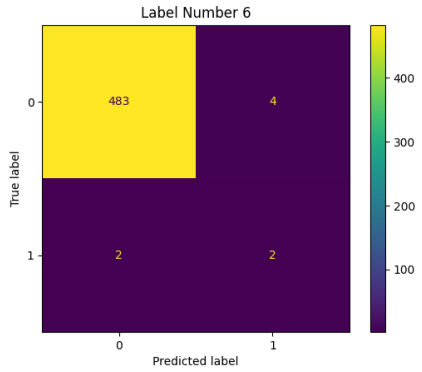

Figure 12. Confusion Matrix for Label 6.

- Item id: 476565359940890624
  Message: "@SOMEONE "@SOMEONE finally off work" slave"
  Reason for disagreement: Since the statement said finally off work, it seemed like it can be classified as coming home from work while the other though that it is a complaint about work due to the word slave.
  Result: It was relabeled as both 'coming home from work' and 'complaining about work'.
- Item id: 546171165765799936
  Message: At work Tonight ! AGAIN !
  Reason for disagreement: One person labeled it as 'complaining about work' and one labeled it as 'going to work'. This is because it can be thought of as the

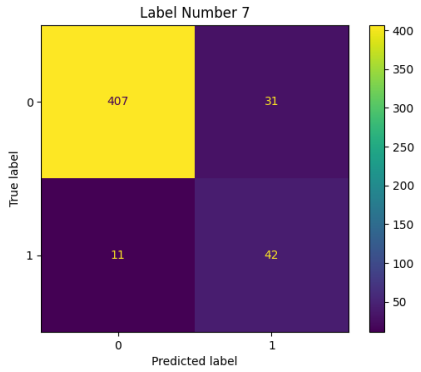

Figure 13. Confusion Matrix for Label 7.

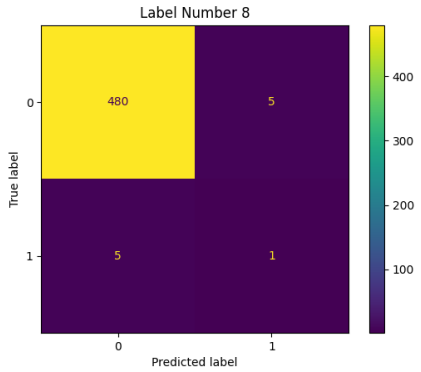

Figure 14. Confusion Matrix for Label 8.

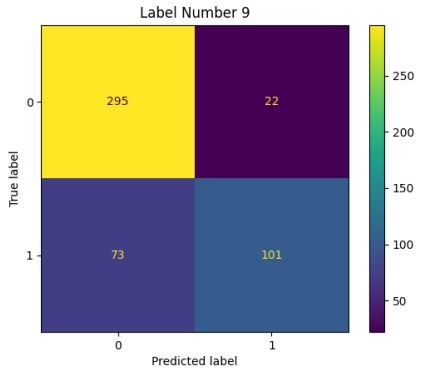

Figure 15. Confusion Matrix for Label 9.

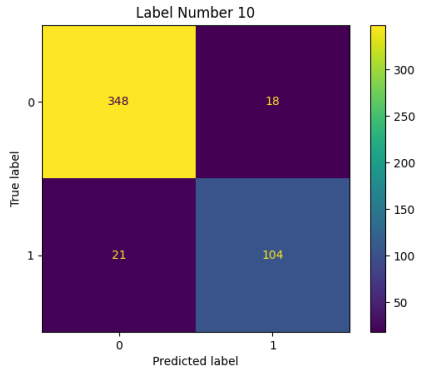

Figure 16. Confusion Matrix for Label 10.

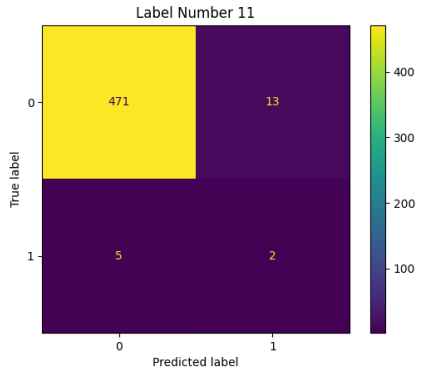

Figure 17. Confusion Matrix for Label 11.

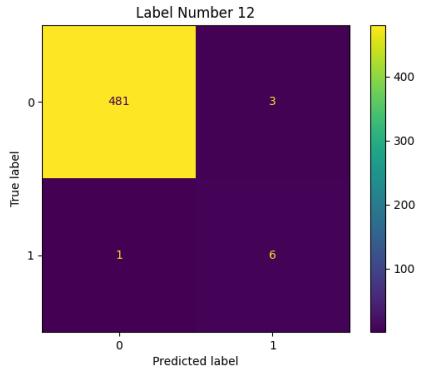

Figure 18. Confusion Matrix for Label 12.

person telling someone he is at work tonight, meaning that he is going to work while the other human thought that the again with exclamation denotes that the person is not happy about being at work.
Result: It was relabeled 'complaining about work'.

- Item id: 462270521917591552
  Message: Bored . At work . Ready to go home .
  Reason for disagreement: One person labeled it as 'coming home from work' while the other labeled it as 'complaining about work' because of the sentence 'Ready to go home'.
  Result: It was relabeled as 'complaining about work'.
- Item id: 396340668165287937
  Message: "@SOMEONE Glad i have to work a 9 hour

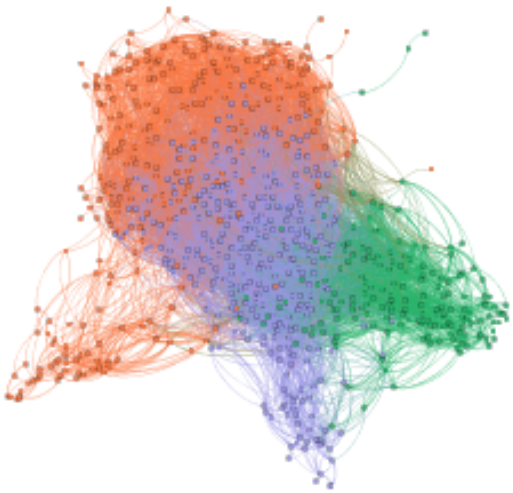

Figure 19. graph produced by the pldl program.

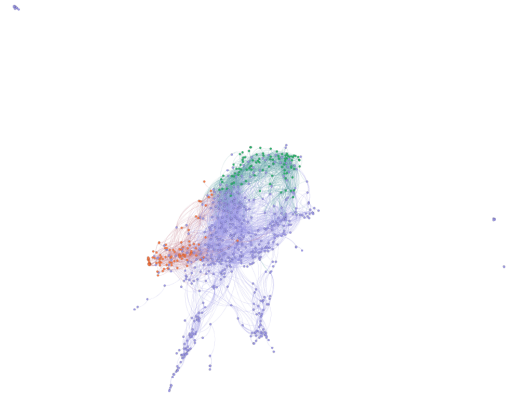

Figure 20. graph produced by the addition labels program.

shift tomorrow"
Reason for disagreement: We agreed that it was job related but one of us thought it meant getting cut in hours too, because of the mention of 9 hour shift. But upon discussion, we agreed that the tweet did not say that the person had a longer shift earlier so we decided to relabel the tweet.
Result: It was relabeled as 'none of the above but jobrelated'

- Item id: 434062023186726912
  Message: I want another job so bad
  Reason for disagreement: We disagreed on whether it should be labeled as 'complaining about work'. But in the end we decided to add this label because of the words' so bad' in the tweet.
  Result: A label was added as 'complaining about work'.
- Item id: 591339623655276544
  Message: Still gonna wear shorts to work a
  Reason for disagreement: We disagreed on whether it should be labeled as job related or going to work or both.
  Result: Labeled as 'none of the above but jobrelated'
- Item id: 456426940925480960
  Message: @SOMEONE I did the nurses gave me medicine at my job so I'm Good
  Reason for disagreement: We disagreed on whether it should be labeled as 'offering support' or 'not job related' or both. Giving the medicine constituted as offering support and since this was not related to a job, we relabeled it as both.
  Result: Relabeled as 'not job related' and 'offering support'
- Item id: 478494538181263362
  Message: I've got work today but all I'm thinking about is 6 o'clock tonight! #BeatGhana #USMNT #bluecollarsoccer
  Reason for disagreement: We agreed that the tweet was job related but one of us thought it related to coming home from work because of the end of the sentence and the mention of 6 o'clock. But since, there is no mention of going home from work, we relabeled the tweet.
  Result: Relabeled as 'none of the above but jobrelated'.
- Item id: 561604165080985600
  Message: Today my last day working til 3am
  Reason for disagreement: We disagreed on whether it should be labeled as 'getting cut in hours'. Since the tweet did not give any indication to claim that the person had more hours ealier, we decided to removed the label.
  Result: Relabeled as 'none of the above but jobrelated'.
- Item id: 518534416730423296
  Message: If you want this money gotta work for it !
  Reason for disagreement: One of us through it is job related but not related to getting raised. But upon discussion, we came to a conclusion that since it relates to working for more money, it can be labeled as getting raised.
  Result: Relabeled as 'getting promoted / raised'.
- Item id: 400703801327628288
  Message: So not in the mood to go to work tonight
  Reason for disagreement: We agreed that the tweet was related to going to work but one of also thought that it should be labeled as complaining about work because of 'so not in the mood'.
  Result: Labeled as 'going to work'.

- Item id: 553377690318028802
  Message: Work again tomorrow.
  Reason for disagreement: One of us labeled it as 'going to work' while the other labeled it as 'complaining about work'. Since, the tweet says working tomorrow, it can be labeled as going to work. Additionally, the word 'again' gave the idea of it being a complaint about working the next day.
  Result: Relabeled as both 'complaining about work' and 'going to work'.

References

[1] [n. d.]. Clustering Coefficient for Nodes. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.clustering.html

[2] 2011. sklearn.metrics.cohen_kappa_score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[3] 2011. sklearn.metrics.confusion_matrix. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

[4] 2011. ssklearn.cluster.AgglomerativeClustering. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[5] 2018. An Introduction to Clustering Algorithms in Python. https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097

[6] 2020. Inter-Annotator Agreement (IAA). https://towardsdatascience.com/inter-annotator-agreement-2f46c6d37bf3

[7] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and János Kertész. 2007. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.75.027105