Data Analysis Report: Microsoft Employee Review

Samantha Paige Minchew

Professor Satvika Marrapu

*Data Analysis and Knowledge Discovery, INFO 4670, Section number 001*

Assignment number #5

14 November 2023

This is a report for the class Data Analysis and Knowledge Discovery. The thesis of this report is to use text-mining tools from Rapid Miner to analyze real data of Microsft Employee Reviews. Building on the last report, we can begin to understand the data analyzed to identify Microsoft's work culture. This report includes Data, Methods, Discussion, and conclusion sections.

**OBJECTIVES:**

To understand the process of gathering and organizing data through Rapid Miner, and to be able to create visualizations of Microsoft employee reviews. Using Rapid Miner, we will be able to data-mine the information to find and utilize a clustering procedure on a corpus of texts in the data set. After transcribing data, we should be able to examine clusters in regard to performance value. Using these data mining techniques will build a conclusion to the last assignment on confidence measurements. Compared to that assignment, we will analyze the “pros” listed by Microsoft employees about work-life at this company.

**DATA:**

The dataset used in this report is listed on Canvas, by Kaggle. “The method of identifying similar groups of data in large datasets is called clustering or cluster analysis. It is one of the most popular clustering techniques in data science used by data scientists. Entities in each group are comparatively more similar to entities of that group than those of the other groups (Kaushik 1).” Clustering is an important and easy way to analyze larger datasets. The data analyzed is taken from 2011-2016, by Microsoft employees during that time. This structured data set is being collected automatically on Rapid Miner. Using location, specific dates, job positions, summaries, pros and cons, management advice, overall ratings, work and life balance, culture value, career opportunities, and company benefits, it is safe to say that this data set is detailed and reliable. With Rapid Miner it will be easy to compare and categorize this data set into a cluster analysis.

**METHODS:**

Transferring the dataset from Excel to Rapid Miner will allow us to create visuals and be able to search through the data to utilize it properly. Using Rapid Miner functions, we will be able to analyze text pre-processed data to a visual report with graphs and clusters. With this tool, we will be able to filter out information to find the pros of working at this company. Rapid Miner gives us the opportunity to manipulate large data sets into something that is understandable and easy to analyze through the use of search functions and data filters. At the end of this report, we will have a data set that only lists the information we want to analyze.

**DATA PRE-PROCESSING:**

We should remember as we do this that means are particularly susceptible to undue influence by extreme outliers, so watching for inconsistent data when using the k-Means clustering data mining methodology is very important.

*Figure 1.a*

**A screenshot of a computer

Description automatically generated**

To begin, we must first drag the dataset from the Repository panel to the Process canvas. This step will load the dataset from its original form in Excel to the Rapid Miner layout. I used the original data sheet from my previous assignment, so I did not have to go through the data editor panel to arrange the attribute type again.

*Figure 1.b*

**A screenshot of a computer

Description automatically generated**

As shown in Figure 1.b, select “attribute” in the Operators panel and drag over to the main canvas. Double click the attribute then go to Parameters and select “include attributes” for *type*, “one attribute” for *attribute filter type*, and “pros” for *select attribute*. Changing this parameter type will filter the results and show only the positive reviews from Microsoft Employees.

*Figure 2.a*

**A screenshot of a computer

Description automatically generated**

Figure 2.a demonstrates adding the “process documents from data” attribute. Text processing helps generate word vectors from string attributes. Adding this process helps create the string attributes needed to process the data results and visualize the centroid table.

*Fifure 2.b*

**A screenshot of a computer

Description automatically generated**

Double-clicking on the “Process Documents from Data” attribute will bring you to a new window where we can begin to manipulate the text data to a string attribute. From here, click on the *Operators* panel, search for “Tokenize,” then add it to the second window canvas. Tokenize separates the collection of information into individual words. This process will allow us to solely seek information for “pros” since that was our original goal. Now in the data results, it will categorize information by selected “tokens” or words. Next, search for “transform cases” in the *Operators* panel. Then check the *Parameters* panel and make sure the “lowercase” attribute is selected. This text processing procedure allows the results to be viewed in all lowercase. Changing this will create more organized and readable results. Now, under the *Operators* panel add “Filter Separated Words (English). This will allow results to only show English words and take out any other data written in a foreign language. Adding this step is important so that we can be sure the data is understandable to the reader and is organized correctly. Lastly, add “Filter Tokens by Length.” This attribute trims the results to a number of characters. For this assignment, I changed the minimum number of characters shown to 4, and the maximum number of characters allowed to 99. Now, we can go back to the main canvas page by double clicking the “Process” button.

*Figure 2.c*

**A screenshot of a computer

Description automatically generated**

The next part of this assignment is to add a clustering attribute. Clustering is a technique for extracting information from unlabeled data. Clustering can be very useful in categorizing data. Adding the clustering attribute will allow our data to be processed into a centroid table. To add clustering as part of the canvas page, first go to the *Operators* panel and search for “K-Means.” In the Parameters panel, we will mark it as a cluster attribute, make “k” equal to 4 and “max runs” equal to 10. Check the “determine good start values” box, then change the measure values. For this assignment, we will change *measure type* to “Numerical Measure,” *numerical measure* to “Cosine Smilarity,” and *max optimization steps* to “100.” Altering these values will allow the data to measure the “average similarity between each cluster and its most similar cluster, ranging from 0 to infinity with a lower value indicating a better separation (North 1).” Clustering is one of the most common exploratory data analysis techniques used to get an intuitive understanding of structured data.

*Figure 3.a*

**A screenshot of a computer

Description automatically generated**

Now, we can add the performance attribute. Cluster Distant Performance operators are used “for performance evaluation of centroid-based clustering methods. This operator delivers a list of performance criteria values based on cluster centroids (GmbH 1).” This performance operator takes this centroid cluster model and clustered set as input and evaluates the performance of the model based on the cluster centroids. As shown in Figure 1.f, the *Parameters* panel displays the function under the “main criterion” setting. For performace operators, you can choose to calculate clusters distance by either average of centroid distance or by using the Davis-Bouldin index. Daves-Bouldin method finds clusters far from each other, compared to categrorizing them close to each other like the “average within centroid distance” method does. In RapidMiner I went ahead and examined both methods to understand the optimal performance of clusters. Seeing as it was negative, the results are good. Thus, I entered “average within centroid distance” changing this setting will allow the performance operator to calculate the average of centroid distances between clusters.

*Figure 3.b*

**A screenshot of a computer

Description automatically generated**

In RapidMiner, you can go ahead and run your process before adding the next step. Extracting Cluster Prototypes operators extract selected prototypes in the data analysis and generate an ExampleSet consisting of the Cluster Prototypes from the model. As seen in the image above, it is applied after the clustering process so the prototypes can be stored for further usage. “The information about the cluster prototypes can be seen in the cluster models generated by most clustering operators but the Extract Cluster Prototypes operator stores this information in the form of an ExampleSet thus it can be used easily (GmbH 1.)” After adding the “Exctract Cluster Prototypes,” we can add the “Write Excel” operator on the canvas page. This operator can be used to write an ExampleSet and transfer it into. A Microsoft Excel document. This operator is important for storing data and transferring data to Excel for extended use.

**RESULTS:**

*Figure 4.a*

A screenshot of a computer

Description automatically generated

This image shows the results of the centroid table from the “Average within Centroid Distance.” You can see the listed attributes from the Microsoft company contribute the clusteres based of off similarity. Thus, each attribute listed is calculated and measured based off the amount of keywords from employee reviews. “A cluster in the k-means algorithm is determined by the position of the center in the n-dimensional space of the n attributes of the Example Set (North 1).”

*Figure 4.b*

A screen shot of a graph

Description automatically generated

In this plotted graph you can see the color-coded diagram of the clusters. On the X-axis, blue represents cluster zero, teal equals cluster one, light green for cluster two, and red demonstrates cluster three. The Y-axis shows the decimal average calculation of the proximity of each of these clustered attributes. These attributes are normalized data calculated by the z-score normalization of clustering to find values that are similar to each other y finding the mean of 0 and the standard deviation of 1. The calculation: x' = (x - μ) / σ subtracts the mean other feature from each value and then divides by the standard deviation. This calculation is what determines the clustering attributes average to each other in regards to the axis plane. Overall, we can see that most of these attributes are clustered closely together except for just a few clusters. This means that most of the employee reviews have positive similarities to each other, minus a few attributes that are seen by ranging further away from the axis. These attributes can either include uncommon words or negative reviews.

**CONCLUSION:**

Overall, by examining the data set of Microsoft Employee Review, we can gather that many employees were happy with the work-life balance. Many of the reviews with the closest clustered words averages had words like “achievement” “accountability” “comfort” “balance” or “benefits.” We can look at these calculated attributes averages and determine that Microsoft has a lot of comfort for employees and focuses on a balanced work-life by offering benefits, the ability to achieve moving up in the company, and also holds employees accountable. We gathered this information by understanding Rapid Miner to analyze and data mine the given data set. From transcribing nominal values to text values, I was able to sort out data by using text processing tools. Learning how to process documents from data is a valuable tool that allows us to visualize and understand data. Analyzing the Microsoft Employee Review data set allows us to understand the cluster values by constructing a data analysis model. Using textual associations to construct data, Rapid Miner analyzes words that appear together to discover hidden or important relationships, so that we can manipulate the data to form visualized and interactive graphs.

**REFERENCES**

GmbH, R. (n.d.). *Cluster Distance Performance (rapidminer studio core)*. Cluster Distance Performance - RapidMiner Documentation. https://docs.rapidminer.com/latest/studio/operators/validation/performance/segmentation/cluster\_distance\_performance.html

Kaushik, S. (2023, July 11). *Clustering: Introduction, different methods, and applications (updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/#:~:text=The%20method%20of%20identifying%20similar,those%20of%20the%20other%20groups.

North, M. (2018). *Data Mining for the Masses*. CreateSpace.