**Mining Meaning from Meta Inc: Text Analysis on Employees’ Laid-off LinkedIn Post**

The project analyzes the profiles and laid-off posts of Meta employees on LinkedIn to predict factors that associate with people’s getting a new job, and

**Data Source**

Datasets include the information scraped from LinkedIn results by searching “I got laid off Meta” and “Impacted by Meta layoff.” Scraped content includes the laid-off posts and profile details of employees.

**Data Description**

Description of the data used in the project, including any data cleaning or preprocessing that was done before analysis.

The dataset consists of 732 observations (posts), and 20 variables of profile details and laid-off posts. The data cleaning process does not take long since the scraping functions have been written to get uniform types of values across variables. However, there are different date units that need to be converted to the same one.

1. Left join post details and profile details files into one data frame. Since one person can make more than a post, we want to left join to keep all content and drop all profile details that do not have url.
2. Data frame cleaning

* Convert post time into “days” unit: the time that posts are made comes in the form of “3d,” “5m.” Therefore, I convert all to the number of days
* Convert years of working experience into “months” unit: the working time comes in the form of “1 yr 8 mos,” “3yrs.” Therefore, I convert all to the number of months.
* Eliminate “\n” or any unnecessary keys in the content of the post
* Convert the Boolean value of new\_job into 0 and 1 instead of False and True

1. Insert positive and negative words count variables

* Import the positive and negative words files
* Use nltk package to tokenize the content, remove stop words, and lemmatize the tokens before creating a word matrix
* Match the words in the matrix with the positive and negative files to count the number of positive and negative words in a post
* Insert two new positive and negative count columns into the dataset
* Create two new variables measuring the positive and negative rate of the post by dividing the new positive and negative count columns with the total word counts of the post

**Variables**



**Analysis**

Classification:

* Classification models are run to see the classification efficiency of different models on the dataset. After using subset selection, the predictors including in the equation are meta\_worktime, word\_num, hastag\_num, para\_num, images, post\_days\_ago, pos\_words\_count, neg\_words\_count.
* Different classification models are run to predict the new job status: Logistic regression, Tree-based methods, Boosting, Linear Discriminant Analysis, Quadratic Discriminant Analysis. The dataset is split into training and test sets to evaluate the model, using AUC score and Accuracy score.

Text Analysis: Using Tokenize posts into a bag of words, represented by a Document-Term Matrix using nltk and quanteda packages

* + Sentiment Analysis: Determine the emotional tone expressed in a post using a set of predefined rules to identify positive or negative tones.
  + Topic Classification: Group posts into topics by collecting frequent words, evaluated using beta (the probability that a given term appears in a particular topic)

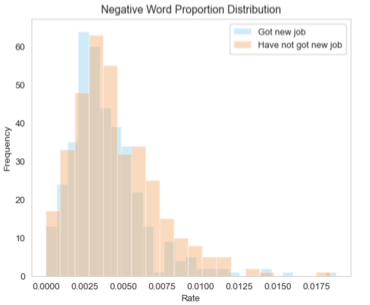
**Figures**

Chart, bar chart, treemap chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated 

**Conclusion**

* Meta layoffs affect thousands of employees regardless of demographics and experience years.
* Employees whose LinkedIn layoff posts have a high positive rate, demonstrate an active attitude, and a strong willingness to work stand higher chances of getting a new job after being laid off. However, more significant sentimental rates should be included with an extension on the dictionary of positive and negative words being matched
* Classification Trees and Boosting yield the highest accuracy score. Boosting is efficient when dealing with complex data by learning slowly based on the residuals, and therefore avoids overfitting. Besides, subset selections have been performed to drop unnecessary predictors.
* LDA and QDA results are similar, which can be explained by the differences in the covariance matrices between the classes are small. In this case, the linear decision boundary of LDA may be sufficient to accurately classify the data, and the increased flexibility of QDA may not lead to significant improvements in performance.