Predicting Employability based on LinkedIn Profile Summary

Manvitha Kola
kolam@tcd.ie

Vardan Kaushik
kaushikv@tcd.ie

Sachin Arvind Gade
gades@tcd.ie

Yifan Pei
peiy@tcd.ie

Wei Hu
huwe@tcd.ie

Abstract

Personality traits of an individual tend to have a great influence on their career growth and the job positions they hold. The language of a person contains cues to their personality traits which is one of the most fundamental dimensions that distinguish two individuals. The current study aims to determine if there is a relationship between employability and the linguistic features extracted from the user-written profile summaries. LinkedIn is a source so far overlooked for personality prediction. For this study, data is collected from a LinkedIn corpus consisting of close to 60k user profiles. The correlation between extracted linguistic features from the profile summary and their corresponding job roles is analyzed to determine the relationship. The methodology, findings and further research are discussed.

Keywords - Linguistic features, Employability, Career growth, LinkedIn

1 Introduction

In this paper, we discuss the employability prospects of a user on a professional networking site, LinkedIn. We are using the profile summaries of LinkedIn users to extract linguistic features to identify personality traits. Linguistic feature extraction is performed on the collected corpus and the employability likelihood of a user is calculated based on these extracted features.

LinkedIn is a social networking platform that focuses on professional networking and career growth. In LinkedIn, people are required to write a summary about themselves which is the first thing that other people and recruiters see when they see their profile. A summary is written entirely by the user and depends heavily on what the user thinks about themselves. We believe that the personality type of an individual is reflected in their summary. In this paper, we want to identify the linguistic patterns and then evaluate if it has any correlation with employability status of the user.

Predicting the employability of a user based on a profile summary is a major step in our research process. Language-based features are extracted from the text to analyze the personality traits of an individual and to verify if there exists any correlation between language and employability.

2 Literature Review

In this part, we searched papers and journals which are related to our research topic: “To what extent is LinkedIn profile correlated with employability?”

LinkedIn Profile Data and Career Development With the rapid increment of data from social media, those data collections have become a huge resource for academic research.

Dai, Nespereira, Vilas, and Díaz Redondo (2015) used scraping and clustering techniques to study LinkedIn’s profile. Their work focuses on categorizing the education level and employability of users in the profile and providing data support for related research in the future.

Seibert, Crant, and Kraimer (1999) are classical in this cross-field, although their research is more closely related to psychology, their research still provided an effective methodology to study the relationship between personality and career. In later time, Judge, Kammeyer-Mueller, and Eliot (2007) made a full introduction, not only mentioning commonly used “Big Five test“, but also discovered potential relation between “social behavior”, “personality”, “career success”, etc. (Fig 2).

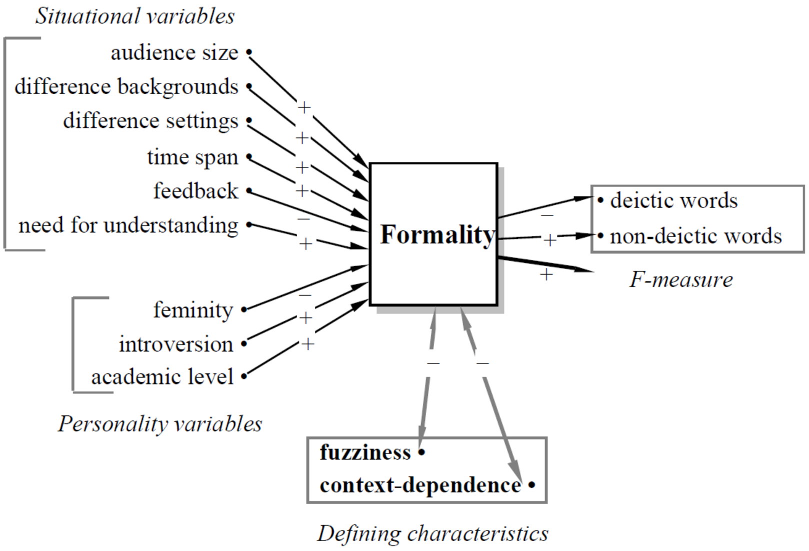


Figure 1: Possible factors which will influence the formality of language Source:Heylighen (1970)

van de Ven, Bogaert, Serlie, Brandt, and Denissen (2017) conducted a social survey based on “Big Five” and found that inferences based on personal data on LinkedIn were correlated with self-assessment scores for these characteristics. This means that information about important personality traits (extraversion and self-expression) will be leaked out through personal profile data on work-related social networks.

Pan, Peng, Hu, and Luo (2017) collected social media LinkedIn data, combined with psychology-based language analysis tools LIWC and SVR algorithms to predict personal future work development, and finally got a preliminary conclusion that personality can affect career development to a certain extent.

3 Research Question

In this research paper, we are trying to analyze if there exists any correlation between the linguistic features of a user and their job prospects. The main research question we are trying to answer in this research is : “To what extent does language in profile summary influence employability?”

Predicting the user’s employability is the key research objective of this study. However, several other objectives need to be fulfilled to achieve this. The following sections describe the summary of the objectives in detail:

a) Extract and explore linguist features from LinkedIn user profile summaries

b) Perform linguistic analysis to determine the most relevant features that influence employability

4 Data Processing

4.1 Data Collection and Pre-processing

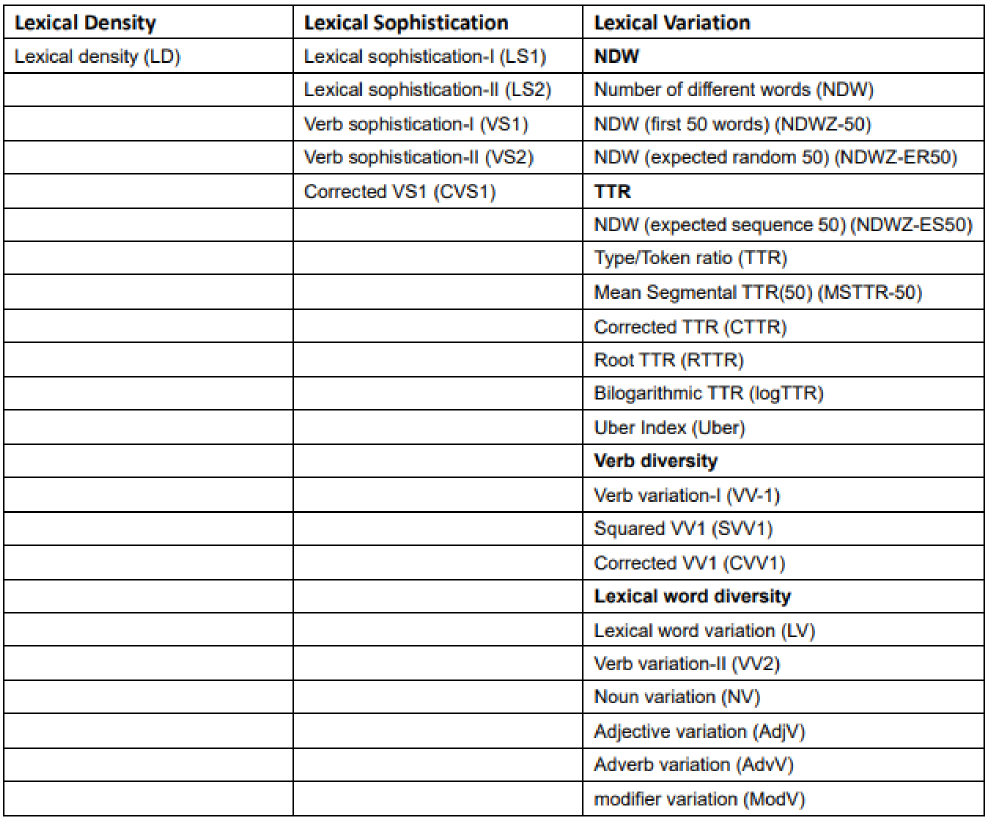
We used a pre-existing dataset which was scraped by using the scraping tool “GRPHIN” which is written in JAVA for this research. The datasets are publicly available on Amazon Simple Storage Service (Amazon S3). It contains about 66k user profiles along with their profile summary, education, industry, job title, previous experience, and employability in JSON format. All the datasets obtained are merged into a single file, which is then restructured into a CSV format. Basic pre-processing steps are performed for data cleaning. To create the final dataset, only the profile summaries with a lot of detail are considered, to be able to analyze the linguistic features of text and their correlation with employability.

4.2 Feature extraction

To identify the linguistic patterns, we identified 89 linguistic features from the given text corpus. The features can be mainly classified into 3 major categories:

- Lexical Complexity
- Parts of speech Tagging
- Word count-based features

LCA (Lexical Complexity Analyzer) is a tool for analyzing the lexical complexity of English text provided by Lu (2012). This tool divides the three types of lexical density, lexical variation and lexical sophistication into 25 specific measurement methods, as shown in Table 1.

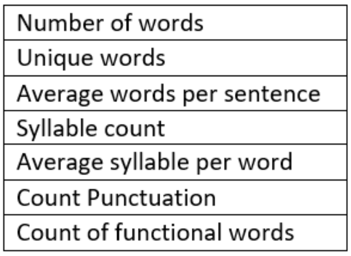


Lexical Density	**Lexical Sophistication**	**Lexical Variation**
Lexical density (LD)	Lexical sophistication-I (LS1)	**NDW**
	Lexical sophistication-II (LS2)	Number of different words (NDW)
	Verb sophistication-I (VS1)	NDW (first 50 words) (NDWZ-50)
	Verb sophistication-II (VS2)	NDW (expected random 50) (NDWZ-ER50)
	Corrected VS1 (CVS1)	**TTR**
		NDW (expected sequence 50) (NDWZ-ES50)
		Type/Token ratio (TTR)
		Mean Segmental TTR(50) (MSTTR-50)
		Corrected TTR (CTTR)
		Root TTR (RTTR)
		Bilogarithmic TTR (logTTR)
		Uber Index (Uber)
		Verb diversity
		Verb variation-I (VV-1)
		Squared VV1 (SVV1)
		Corrected VV1 (CVV1)
		Lexical word diversity
		Lexical word variation (LV)
		Verb variation-II (VV2)
		Noun variation (NV)
		Adjective variation (AdjV)
		Adverb variation (AdvV)
		modifier variation (ModV)

Figure 2: The 25 Measures using Lexical Complexity Analyzer

In our research, a total of 34 features are extracted using LCA which include types and tokens based on parts of speech (wordtypes, swordtypes, lextypes, slextypes, wordtokens, swordtokens, lextokens, slextokens) , one for lexical density (ld), two features (ld1, ld2) for lexical sophistication, 3 features (vs1,vs2,cvs1) for verb sophistication, 4 for lexical diversity(ndw,ndwz,ndwerz,ndwesz), 6 for type token ratio(ttr, msttr, cttr, rttr, logttr, uber), 3 for verb diversity (vv1, svv1, cvv1), and 6 for lexical diversity (lv, vv2, nv, adjv, advv, modv)

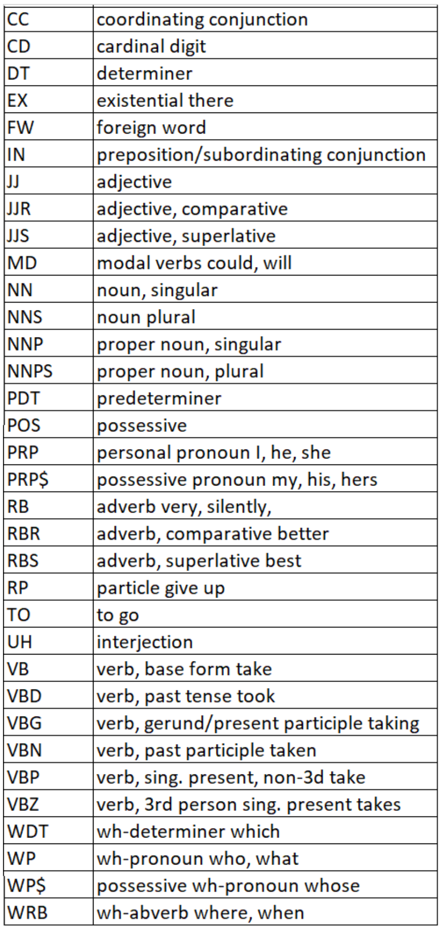
Word count-based features that are extracted include TotalWords, ComplexWords, UniqueWords and average words per sentence as shown in Table 2



Number of words
Unique words
Average words per sentence
Syllable count
Average syllable per word
Count Punctuation
Count of functional words

Figure 3: Word count based features

35 features are extracted using POS (part-of-speech) tagging which include Conjunction, digits, determiners, prepositions, adjectives, modal verbs, nouns, pronouns, adverbs, verbs, wh-words and interjections.



CC	coordinating conjunction
CD	cardinal digit
DT	determiner
EX	existential there
FW	foreign word
IN	preposition/subordinating conjunction
JJ	adjective
JJR	adjective, comparative
JJS	adjective, superlative
MD	modal verbs could, will
NN	noun, singular
NNS	noun plural
NNP	proper noun, singular
NNPS	proper noun, plural
PDT	predeterminer
POS	possessive
PRP	personal pronoun I, he, she
PRP$	possessive pronoun my, his, hers
RB	adverb very, silently,
RBR	adverb, comparative better
RBS	adverb, superlative best
RP	particle give up
TO	to go
UH	interjection
VB	verb, base form take
VBD	verb, past tense took
VBG	verb, gerund/present participle taking
VBN	verb, past participle taken
VBP	verb, sing. present, non-3d take
VBZ	verb, 3rd person sing. present takes
WDT	wh-determiner which
WP	wh-pronoun who, what
WP$	possessive wh-pronoun whose
WRB	wh-abverb where, when

Figure 4: POS based features

5 Research Methods

5.1 Statistical Analysis

Now that the features have been extracted from the summary of the user, we begin with determining if there exist any correlation between the features and the employability of the user. The resultant feature vector has 82 features, a hybrid collection of Parts-Of-Speech, Lexical Complexity Analyzer and Word Count Based features.

We first begin with calculating the correlation coefficient between each feature and the target variable (employability status), then we determine the p-values for the same. The correlation coefficient measures the strength of the relationship between two variables and the p-value indicates the statistical significance of the calculated correlation coefficients. All the features with p-values smaller than 0.05 are considered to be highly correlated and are shown in the Results and Conclusion section. We further plot some bar-plots of the features with respect to their employability status. As the features are just total count in the domain of natural numbers therefore they can be plotted on a bar graph. This is a simpler and illustrative way to determine which features correlate highly with our target variable.

5.2 Classification Problem

Having few features highly correlated with our target variable is not enough to verify our hypothesis but if we can train a machine learning classification model on this dataset and get satisfactory results then we can conclude our hypothesis with confidence. Fortunately enough, our research question can be converted to a binary classification problem with target variable having values "employed" and "unemployed. However, we face two problems with our original dataset: first, it is unevenly distributed - with 57,000 "employed" users and 8,700 "unemployed" users. If we are to train a machine learning model on this dataset it will most often than not just predict the "employed" value of the target variable. Second, there are few users who have not considered writing a summary in their profile and so their summary field is empty. We want an even dataset and we want to ignore those users who have an empty summary field.

5.3 Data Processing for Machine Learning

As all the features values are just the total count of those features in the domain of natural numbers, therefore we took the maximum sum of each row to determine users with a content rich summary. We did that for both "employed" users and "unemployed" users to avoid any discrepancy in our dataset. By doing this, we ignored all the users with an empty summary field and then we took the first 8,000 observations with maximum feature values for both the "employed" and "unemployed" users. The resultant dataset has 16,000 observations and is evenly distributed.

5.4 Machine Learning Models

Now we use this dataset of 16,000 observations to train three machine learning classification models : Logistic Regression (LR), Support Vector Machine Classifier (SVC) and K Nearest Neighbors Classifier (KNN). We also use a dummy classifier which always predicts the most frequent class to define our baseline model.

The dataset is divided into 80-20 train-test datasets and the final accuracy score is calculated for each model which is given in the Results and Conclusions section.

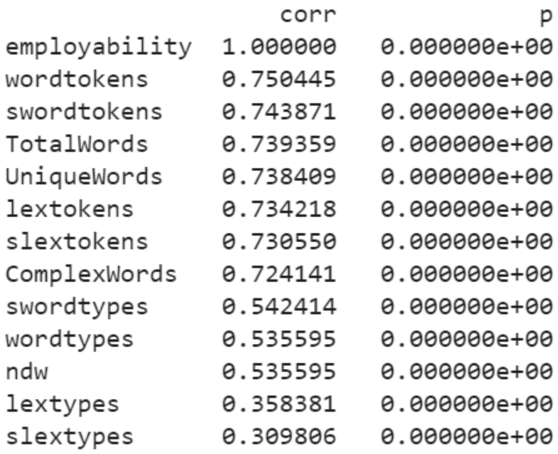
5.5 Cross-validation

We use K-fold cross validation with a number of folds equal to five to prevent overfitting the data as well as to get the optimal value of the hyper-parameters of the three models used. For LR, the optimal value of hyper-parameter C = 0.1, for SVC, optimal value of hyper-parameter C = 0.01 and for KNN the optional number of neighbors is equal to 9.

6 Results and Discussion

6.1 Statistical Analysis

The correlation coefficient measures the strength of the relationship between 2 variables. Point biserial correlation coefficients are calculated to measure the relationship between the feature variables which are continuous and a target variable, 'employability', which is binary. the p-value indicates the statistical significance of the calculated correlation coefficients. Considering only the features where p-value (0.05), i.e when the correlation is statistically significant.



```
                   corr              p
employability  1.000000   0.000000e+00
wordtokens     0.750445   0.000000e+00
swordtokens    0.743871   0.000000e+00
TotalWords     0.739359   0.000000e+00
UniqueWords    0.738409   0.000000e+00
lextokens      0.734218   0.000000e+00
slextokens     0.730550   0.000000e+00
ComplexWords   0.724141   0.000000e+00
swordtypes     0.542414   0.000000e+00
wordtypes      0.535595   0.000000e+00
ndw            0.535595   0.000000e+00
lextypes       0.358381   0.000000e+00
slextypes      0.309806   0.000000e+00
```

Figure 5: Point biserial correlation values between feature scores and employability. Significant correlations are shown when p 0.05. Only features that correlate significantly are shown in the figure.

We also plotted bar plots on our features divided by the employability status of the user. We calculated the average number of features for each category and plotted them. The plots in the below Figure 6 are for the eight features with a maximum difference between both categories.

6.2 Machine Learning

For the dataset that we choose, we can see that there is some correlation between the features extracted from the summary of the user and their employability status. However, we feel that this is not enough to verify our hypothesis. Therefore, we trained three classification models with target variable as employability status of the user and features extracted from the summary as input.

Figure 7 below shows the classification accuracy results for all the models trained along with the accuracy metric calculated on the training dataset -

We can see that KNN Classifier gives the highest accuracy, Logistic Regression and Support Vector Classifier shows the same results, however, SVC doesn't converge even with 10000 iterations so we won't consider the results of SVC to conclude our hypothesis. We can also see that the dummy classifier, always predicting the most frequent class, gives 49% accuracy on testing data and 50% accuracy on training data.

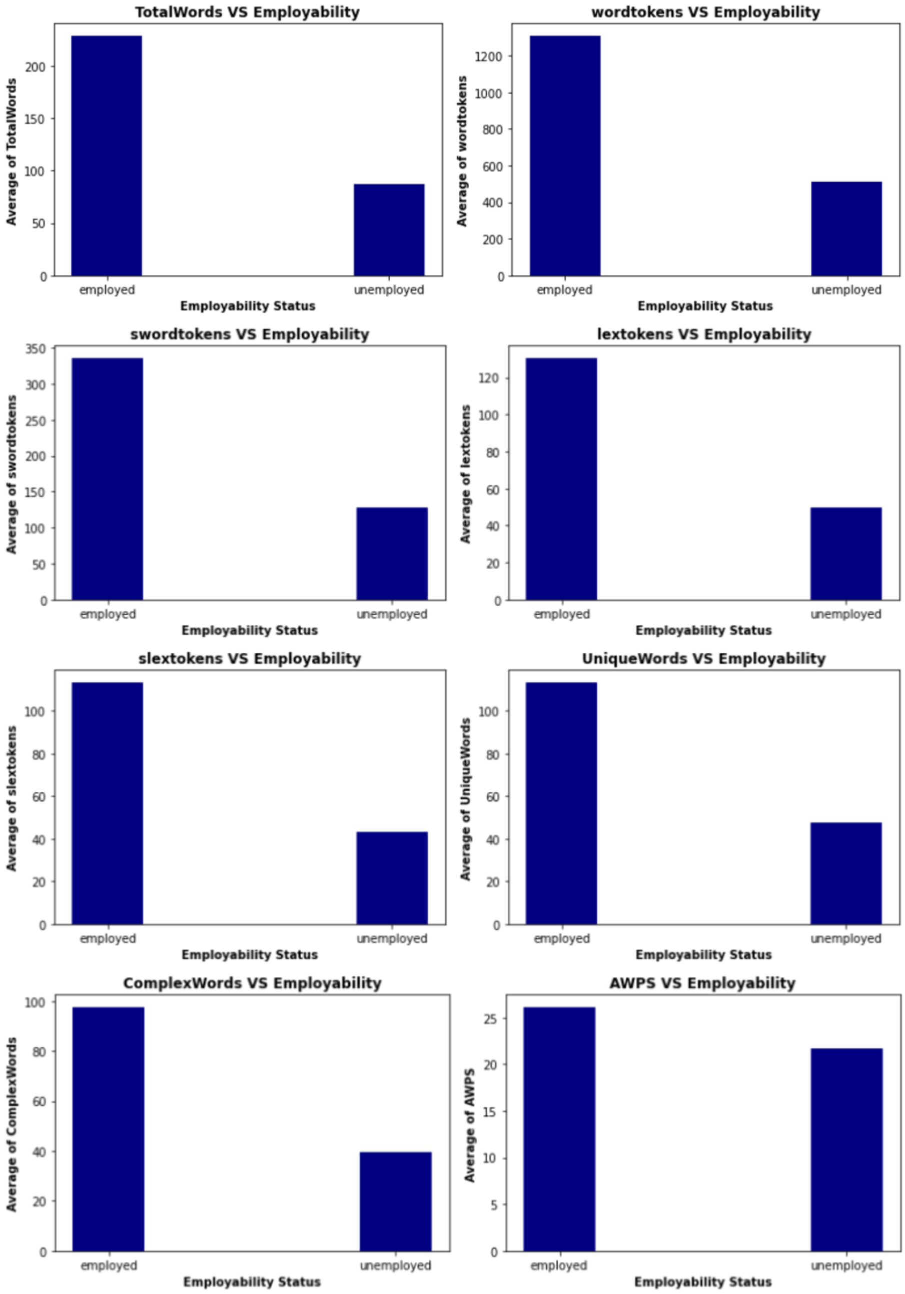
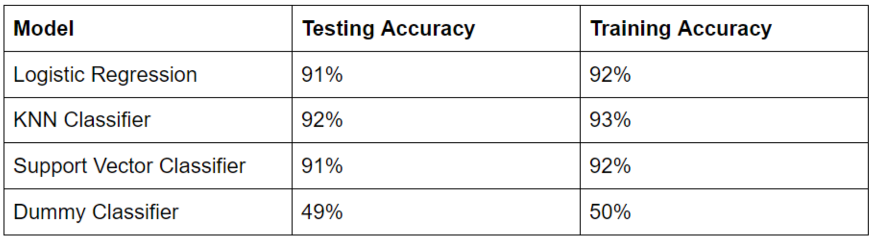


Figure 6: Feature differences in Categories



Model	Testing Accuracy	Training Accuracy
Logistic Regression	91%	92%
KNN Classifier	92%	93%
Support Vector Classifier	91%	92%
Dummy Classifier	49%	50%

Figure 7: Classification Model Results

50% accuracy of the Dummy classifier on training data proves that the dataset is evenly distributed. Moreover, we also know that the high accuracy of our models is not because we overfitted our data as the training data accuracy of all the models is lower than 95%. Furthermore, the high accuracy of all three classifiers shows that the input data can successfully be separated based on the target variable, i.e. there is some correlation between the employability status of the user and their LinkedIn summary.

7 Conclusion and Future Work

In this paper, we have shown that the employability of a LinkedIn user can be predicted from the profile description they share on LinkedIn. We collected publicly accessible data from LinkedIn user's profiles and extracted Language-based features from the collected profile descriptions. Using the generated feature set, we trained 3 machine learning algorithms - Logistic Regression, KNN and SVC - to predict the employability of a LinkedIn user. All the algorithms achieved an accuracy close to 90%. Such high accuracy on testing data, provided that the accuracy on training data didn't cross the 95% threshold, verifies that the model didn't overfit and there indeed is some underlying dependency between the summary of the user and their employability status.

With the promising results from current research, future work can be focused on trying to improve the model. When using classification models, feature selection has a high impact on the predictions. A direction for future work is to investigate more on what features that could be added to improve the results. Another area that this research can be further extended is to analyze the management position of the user based on the linguistic features in their profile summary. That is, to determine if there is any underlying dependency between the summary of the user and the management position they hold in their company, CEO being the top level management and intern being the lowest level of management.

References

K. Dai, C. Nespereira, A. Vilas, and R. Díaz Redondo, "Scraping and clustering techniques for the characterization of linkedin profiles," 01 2015.

S. Seibert, J. Crant, and M. Kraimer, "Proactive personality and career success," *The Journal of applied psychology*, vol. 84, pp. 416–27, 07 1999.

T. Judge, J. Kammeyer-Mueller, and T. Eliot, "Personality and career success," 01 2007.

F. Heylighen, "Formality of language: definition, measurement and behavioral determinants," 02 1970.

N. van de Ven, A. Bogaert, A. Serlie, M. Brandt, and J. Denissen, "Personality perception based on linkedin profiles," *Journal of Managerial Psychology*, vol. 32, 09 2017.

Y. Pan, X. Peng, T. Hu, and J. Luo, "Understanding what affects career progression using linkedin and twitter data," 12 2017, pp. 2047–2055.

L. Viii, K. Intelligenz, and T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," 01 1999.

X. Lu, "The relationship of lexical richness to the quality of esl learners' oral narratives," *The Modern Language Journal*, vol. 96, pp. 190–208, 06 2012.

G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and bernoulli naïve bayes for text classification," 04 2019, pp. 593–596.

O. Aborisade and M. Anwar, "Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers," 07 2018, pp. 269–276.

Viii et al. (1999) Lu (2012) Singh et al. (2019) Aborisade and Anwar (2018)

Individual Contributions:

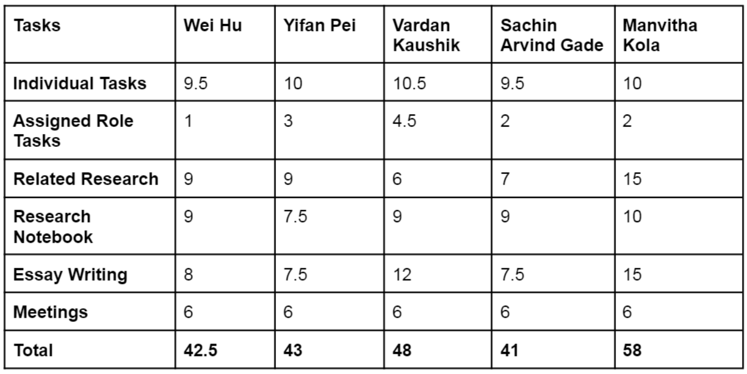
Wei Hu (Verifier) My role in the group is to determine whether other group members have fulfilled their responsibilities in each meeting and record it. In the current progress, I participated in reading related research articles, participated in writing a literature review, and read the literature on the selection of related models, and sorted out the characteristics of each model.

Yifan Pei(Recorder) My role in the group is to record details of every team meeting.During the midterm essay writing, I mainly wrote the 'literature review' and part of calculation of features importance.In the final essay writing, I was mainly responsible for the modification of "literature review".

Manvitha Kola(Chair) As the Group's Chair, I took responsibility for arranging meetings to discuss and review the progress of the team work. I contributed towards finalizing the research topic and created the setup for code and document sharing. I worked on analysing the collected data set to extract linguistic features for our analysis. I worked on the code to extract all the linguistic features from the text. I performed statistical analysis on the extracted features including correlation analysis and the statistical significance test. I contributed to the final essay structure and formatting I took the responsibility of finalizing the LaTex template and formatting the document to the required format.

Vardan Kaushik (Ambassador) My role as an Ambassador was to collaborate with other groups and share ideas along with resources. As for my contribution towards this essay, I first helped by giving ideas for the project. I further contributed to the data preprocessing process by classifying the data in six different management levels. Moreover, I collected all the feature files and merged them into one to create our original feature dataset. I also worked on developing the entire machine learning classification process from processing the dataset to cross validation and finally training the models to get an accuracy score. For writing the final essay, I contributed by writing the Results and Discussion section along with the Conclusion section. I also contributed by formatting the essay and improving its structure.

Sachin Arvind Gade (Accountant) As an accountant of the group, my job was to keep track of time devoted towards the project and to check contributions given by everyone. For this, I used the way to ask in the meetings and then followed up on the tasks assigned. For the midterm essay, after discussion, we decided to find an appropriate dataset for processing. I helped the project by finding and extracting the appropriate dataset. I also helped in the preprocessing of the dataset. I helped in finding correlation between the features in feature extraction.



Tasks	Wei Hu	Yifan Pei	Vardan Kaushik	Sachin Arvind Gade	Manvitha Kola
Individual Tasks	9.5	10	10.5	9.5	10
Assigned Role Tasks	1	3	4.5	2	2
Related Research	9	9	6	7	15
Research Notebook	9	7.5	9	9	10
Essay Writing	8	7.5	12	7.5	15
Meetings	6	6	6	6	6
Total	42.5	43	48	41	58

Figure 8: Time Spent(In hours)

Wei Hu
Wei Hu

YiFan Pei
Yifan Pei

Manvitha Kola

VK
Vardan Kaushik

Sachin Arvind Gade

Signed on: 20th April, 2021