Field Lab YunoAI: Startup Analytics

A Machine Learning Approach to Predict Startup Success based on Founders' Features

Benjamin Schmidt

Work project carried out under the supervision of:

Qiwei Han

15/01/2025

Abstract

This thesis examines the impact of founders on startup success, focusing on education, professional experience, and LinkedIn metrics. It analyzes university and workplace prestige, including consultancy or VC experience, and LinkedIn follower count as a connectivity measure. Using advanced text analysis with SBERT and TF-IDF embeddings, it evaluates LinkedIn posts and profiles—a novel approach in predicting startup success. Founder personality traits are predicted through X platform data linked to LinkedIn profiles via Crunchbase. A three-layered success definition offers a comprehensive framework for evaluating outcomes, providing new insights into the relationship between founder characteristics and startup success.

Keywords

Startup Success, Funding Metrics, Failure Prediction, Founder-Market Fit (FMF), Machine Learning, Natural Language Processing (NLP), Startup Dynamics

1. Problem Statement

Accurately predicting startup success requires comprehensive information about founders to better understand their characteristics. Previous studies have utilized X data to predict founders' personalities as a feature for startup success, but the limited availability of X links in Crunchbase restricts its applicability. Moreover, prior research often examines individual founder features in isolation, making it challenging to determine their relative importance. Startup success is also frequently defined by singular metrics, such as M&A, IPOs, or funding raised, overlooking other successful startups.

This thesis addresses these limitations by leveraging LinkedIn data—including comments, posts, reactions, and profile descriptions—which has never been used before in the context of startup success prediction and provides broader accessibility to founder information compared to social media data like X posts. It also evaluates a diverse set of founder features to assess their relative importance. Finally, a novel three-layered success definition is introduced, capturing a more inclusive and multidimensional view of startup success.

2. Institutional Background

This study is conducted in collaboration with YunoAI, a pioneering AI technology company specializing in decision-making and market analysis solutions. YunoAI empowers stakeholders with actionable, timely data, guided by its values of innovation, integrity, customer-centricity, and sustainability. Its mission is to redefine business opportunity evaluation through data-driven approaches, fostering responsible economic growth and social impact.

YunoAI’s expertise lies in its SaaS platform, which integrates structured and unstructured data to enhance the sourcing, screening, and evaluation of companies. With advanced NLP algorithms, its semantic search uncovers nuanced insights, enabling in-depth analyses of

financial, operational, and qualitative metrics. The platform automates analytical report generation using a generative AI engine, streamlining the interpretation of diverse data sources like social media, financial filings, and market reports into actionable insights.

This collaboration demonstrates YunoAI's commitment to data-driven innovation in the startup ecosystem. Leveraging its cutting-edge platform, the study develops a predictive model for startup success, analyzing founder. The partnership advances academic understanding and practical methodologies for decision-making in startup and investment contexts, reinforcing YunoAI's role as a leader in data-driven solutions for dynamic markets.

3. Literature Review

Startups, with their growth potential and industry disruption, face failure rates of up to 90% (Gage, 2014), highlighting the importance of predicting their success.

3.1 Early Models of Startup Succes

Early models of predicting startup success focused on as key indicators (MacMillan, Siegel, & Subba Narasimha, 1985; Gompers & Lerner, 2001).

Today, the concept of success has expanded. Investors want is an increase of startup's future value. Beyond revenues and financial milestones, factors like team dynamics, social and environmental impact, product innovation, and market influence now play critical roles in startup valuation. As a result, startup valuation measures have become a common target variable for assessing success, capturing what matters to investors.

3.2 Literature on Importance of Founders in Startup Success

Research highlights the pivotal role of founders in startup success, emphasizing the influence of their characteristics, experiences, education, and personalities on venture outcomes (McCarthy et al., 2023; Brahmana et al., 2024). This focus on founders is especially critical in

the absence of financial or historical performance data during the early stages, making founder analysis a key predictor of success (Freiberg & Matz, 2023).

Key factors influencing startup success are categorized within the KECCT framework: Knowledge, Experience, Competence, Characteristics, and Team dynamics (Virágh et al., 2024). Founder personalities are significant predictors (Stenson, 2023; Freiberg & Matz, 2023). This review synthesizes these insights, highlighting the critical role of founder-related factors in entrepreneurial success, offering valuable contributions to research and practice.

3.2.1 Founder Education and Professional Background

The educational background and professional experience of founders have been shown to significantly influence startup oucomes (Pinelli et al. 2022; Kimani 2024).

Educational Impact

Research by AngelList reveals that startups led by alumni from less traditionally prestigious institutions, such as the University of Washington and University of Waterloo, often achieve higher markup rates compared to those from elite universities like Stanford, MIT, and Harvard (Othman & Speiser, 2021). While elite university affiliations provide advantages, such as higher initial valuations in early-stage funding due to access to robust networks, these benefits do not guarantee long-term entrepreneurial survival. In fact, high initial valuations may result in lower markup rates in subsequent funding stages.

Entrepreneurial education fosters innovation, persistence, and effective networking with mentors, supporting founders in creating successful ventures (Kimani, 2024). While the impact of university prestige is limited, there is a clear correlation between founders holding a university degree and an increased likelihood of success (Shanar, 2023).

Professional Experience

Professional experience plays a pivotal role in determining founder success and startup performance, with several key aspects emerging as particularly influential:

Prior Work Experience

The length and relevance of a founder's work experience strongly influence startup success. Extensive experience helps secure funding through industry knowledge and networks, enhancing market navigation and connections (Alfons, 2022; Stenson, 2023).

3.2.2 Founder Interests and Hobbies

Founders' interests and habits, such as regular physical activity or sports, contribute to startup success by enhancing stress management, discipline, and well-being (Freiberg & Matz, 2024).

3.3.3 Founder Personality Research

The Big Five personality traits significantly influence startup success predictions. Research identifies six founder personality types—fighters, operators, accomplishers, leaders, engineers, and developers—demonstrating diverse paths to success without a single ideal personality (Banerji & Reimer, 2019a; Otterlei & Wik, 2023; Brahmana et al., 2024).

3.3.4 Machine Learning Models and datasets

Research on predicting founder success emphasizes individual traits, with Crunchbase—a platform offering business information on private and public companies—and LinkedIn data as key sources. LinkedIn data, including education, experience, and connectivity metrics, enhances models by reflecting network reach (Banerji & Reimer, 2019b; Te et al., 2022).

Predicting founder personality is challenging, but the PANDORA dataset facilitates analysis using X (formerly Twitter) posts labeled with Big Five personality scores, applicable for founders linking X profiles on Crunchbase (Gjurković et al., 2021). Traits like high openness and low neuroticism strongly correlate with success (Braesemann & Stephany, 2023). Models such as XGBoost, random forests, and neural networks using TF-IDF or BERT embeddings

are commonly employed for text vectorization (Stenson, 2023; McCarthy et al., 2023b).

Personality traits, alongside education, experience, interests, and network connectivity, remain critical predictors, highlighting the multifaceted nature of entrepreneurial success.

4. Data

4.1 Data Overview and Sources

The data pipeline of the company and founder data is visualized in the following figure. The text will refer to this figure.

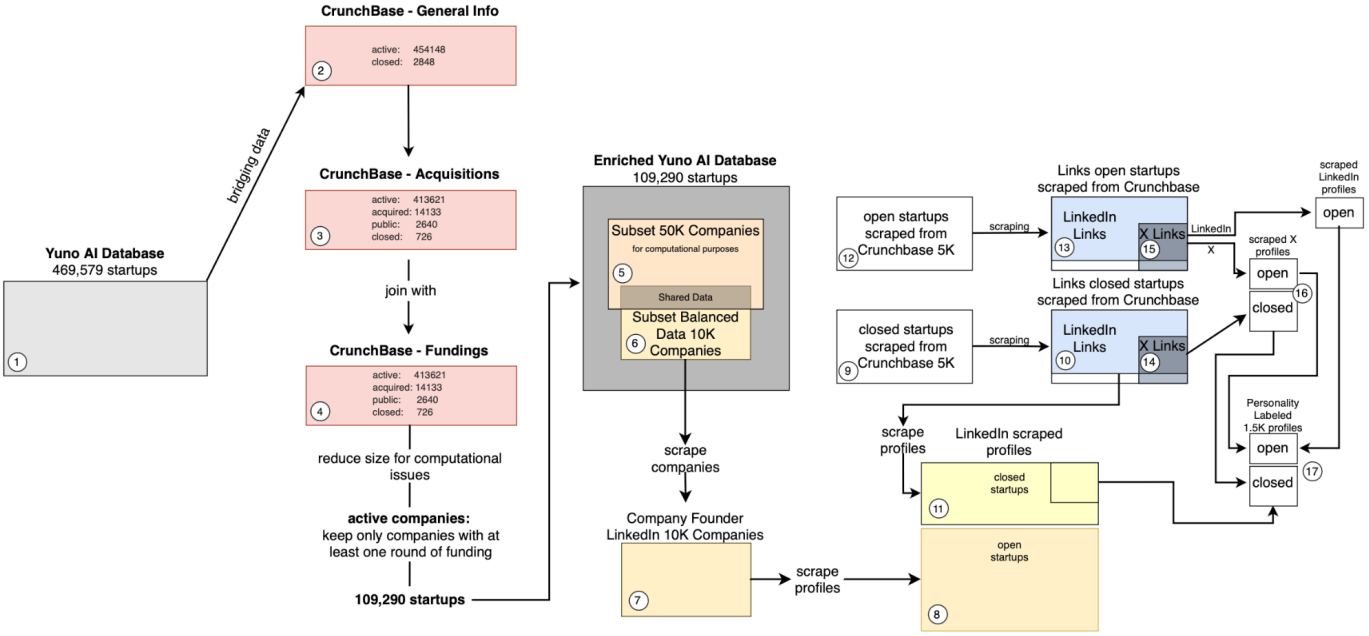


Figure 1: Data Flow, Nr. 1-17

The data collection for this study began with the YunoAI dataset, which was primarily based on data from Crunchbase. We expanded and refined this dataset by incorporating additional data from Crunchbase, LinkedIn, and X (formerly Twitter). These multiple sources allowed us to create a comprehensive framework for analyzing startup success, with a focus on features related to funding, founders, and investors.

4.2 YunoAI Dataset

The YunoAI dataset was the backbone of our analysis, comprising 469,579 records across 30 variables in JSON format (Figure 1: Data Flow, Nr. 1). From this dataset, we retained the most relevant columns for our analysis.

4.3 CrunchBase Datasets Integration

To enhance the initial dataset, we incorporated additional data from CrunchBase. CrunchBase offers insights into investments, funding rounds, acquisitions, and key industry players, making it an essential resource for professionals in venture capital, private equity, and entrepreneurship to track market trends and assess business performance.

The dataset was enriched by integrating four specific data sources from CrunchBase:

1. *General Company Data*: Companies' operating status (Figure 1: Data Flow, Nr. 2).
2. *Acquisitions data*: Companies' acquisitions status (Figure 1: Data Flow, Nr. 3).
3. *Funding Data*: funding round data –type, amount, and date– (Figure 1: Data Flow, Nr. 4).
4. *Investors Data*: Individual and institutional investors of funding rounds

4.4 Subset Dataset of enriched YunoAI Data

A subset of the enriched dataset (Figure 1: Data Flow, Nr. 1-5) was created, containing 10,000 companies (Figure 1: Data Flow, Nr. 6). This subset included features like funding round dates, amounts raised, and operating status (active, IPO, or M&A) and focused on western English-speaking companies in the USA, UK, Canada, and Australia, founded between 2010 and 2020 (Western Countries, 2024).

Focusing on English-speaking companies enabled consistent analysis of LinkedIn text data, such as descriptions and posts, for predicting founders' personalities and their impact on startup success. The 2010–2020 timeframe provided relevant data while balancing historical

and contemporary insights. Key variables, such as employee range, last round type, and founding year, were balanced to minimize biases.

4.5 Founders' LinkedIn URLs Dataset

Using a subset of 10,000 startups, LinkedIn company page URLs (Figure 1: Data Flow, Nr. 7) were used to extract founder information (Figure 5, Appendix). Automated Selenium scripts searched the "People" section for keywords like "founder" or "co-founder," creating a CSV file linking startups to LinkedIn profiles (startup_name and profile_link, Figure 6) and producing a dataset of founders (Figure 1: Data Flow, Nr. 8).

From 10,000 startups, 38,800 potential founders or team members were identified across 6,811 startups, as many lacked valid LinkedIn URLs due to closures, renaming, or acquisitions. Google Lens OCR and reverse image search recovered inaccessible profiles, and rule-based filtering removed irrelevant entries. Ethical guidelines, including LinkedIn policy compliance and data anonymization, were adhered to (Matić & Jukić, 2020; Islam & Kang, 2022).

4.6 Founders' LinkedIn Profiles Dataset

Professional and educational data were collected from LinkedIn profiles using automated web scraping via Selenium. The data, stored in JSON format for seamless analysis (Figure 7, Appendix), included basic information such as names, locations, and follower counts. It also encompassed professional experience, including job titles, companies, LinkedIn URLs, and employment durations. Additionally, education details were captured, such as degrees, institutions, URLs, and attendance years. Engagement metrics, including posts, comments, and likes, were also included.

Dynamic content loading, particularly for engagement metrics, required up to five scrolls per profile. Scraping adhered to ethical guidelines, capturing only public information and using multiple LinkedIn Premium accounts to comply with daily limits (~300 profiles/day). This

time-intensive yet ethical process spanned several months to ensure high-quality data collection.

4.7 University Rankings Dataset for calculating University Prestige Score

University prestige was quantified using three ranking levels: global rankings (QS, Times Higher Education, ARWU Shanghai, CWUR), regional rankings (e.g., QS Regional Rankings), and subject-specific rankings (e.g., QS Subject Rankings, Financial Times Rankings). These rankings were combined into a single ranking score, offering a nuanced evaluation of educational pedigree across geographic and disciplinary contexts.

4.8 Five Thousand Closed Startups scraped from Crunchbase

For a prediction model we do not only need open, but also closed companies which failed to have an unbiased prediction. For this we scraped the company profiles from Crunchbase (Figure 1: Data Flow, Nr. 9). In this case we did not have to exclude companies which were founded after 2020 because we can already see their performance outcome, even if they got founded after 2020. They got closed. "Closed" can indicate bankruptcy, merger, or acquisition, as detailed in the company status column. This dataset also provides information on funding raised, funding rounds, founder Crunchbase profiles, investors, industry, and company description.

4.9 Links open Startups scraped from Crunchbase

Based on the Crunchbase profile links (Figure 1: Data Flow, Nr. 10) we scraped the Crunchbase founders' profiles to get their LinkedIn links. While scraping the LinkedIn links we also scraped the twitter links of the founders. While most of the time the founders put their LinkedIn link on Crunchbase only few founders put their X (Twitter) link on Crunchbase.

4.10 LinkedIn scraped Profiles closed

Based on the LinkedIn links extracted from the founders we could also scrape the LinkedIn

profiles of the closed startups (Figure 1: Data Flow, Nr. 11). In the end there were approximately 3 thousand profiles, because some of the LinkedIn links were inactive already, or sometimes there was no LinkedIn link in the founder's Crunchbase profile.

4.11 Open Startups scraped from Crunchbase

Some founder Crunchbase profiles included both Twitter and LinkedIn links, presenting an opportunity to predict founder personality using X posts labeled with PANDORA's personality database. By extending the predicted personality labels to LinkedIn profiles, we aimed to enable personality prediction of LinkedIn profiles, which currently lacks linked personality data. To avoid bias, the model included open startup founders with both LinkedIn and X links. The same Crunchbase scraping process was applied to open and closed startups, resulting in a dataset of open startups with founder Crunchbase profiles (Figure 1: Data Flow, Nr. 12, 13).

4.12 Link open Startups scraped from Crunchbase

Similar to the closed startups, we scraped founder Crunchbase profiles from the open startups to obtain founder LinkedIn and X links. Of the 800 profiles with both links, we scraped both X and LinkedIn profiles of the founders (Figure 1: Data Flow, Nr. 16). We also scraped X profiles for the 1,200 closed startup (Figure 1: Data Flow, Nr. 16) founders with X links that were already scraped from Crunchbase profiles.

4.13 Personality Labelled 1.5 thousand profiles

Personality prediction for X profiles was conducted using various models, including Random Forest, XGBoost, combined with SBERT, and TF-IDF vectorization. The pretrained **BertForSequenceClassification** model with uncased tokenization ([BERT Transformer Library]) achieved the best performance, with an MSE of 0.0781, MAE of 0.2308, and Training Loss of 0.0683, indicating approximate personality labels. A major limitation was the dataset's small size (Figure 1: Data Flow, Nr. 18), as only 1,500 labeled LinkedIn profiles

were derived from 10,000 scraped companies, constrained by the limited number of founders with both LinkedIn and X links. This small training set restricted the effectiveness of the LinkedIn personality model and contributed to the approximations in labeling.

5. Definition Startup Success

This part of the thesis defines success in three layers, each focusing on the performance achieved after the startup's first five years. The focus on performance after five years is motivated by two reasons. First, as the literature review has shown, the founder's role is a significant and critical factor, especially in the early stages of a startup's life. Second, examining performance after five years helps mitigate potential bias related to the varying ages of startups.

5.1 Success through M&A, IPO, and Failure through Bankruptcy

The first layer predicts whether the company will be acquired, merged, achieve an IPO within five years, or close because of bancruptcy. If a company closes due to being merged into the acquiring company, it is considered a success because the original company’s closure is a result of the merger. For this reason, this part of the thesis defines a company as successful if it closes within the first five years due to a merger or acquisition. Additionally, for this thesis, achieving an IPO is defined as a success, even if the company closes afterward.

5.2 Success measured in Funding Amount raised

The second layer of success prediction assesses how much money the company raised within its first five years. This data is extracted from the initial financial dataset, which was merged with the dataset shared by YunoAI. It is crucial to consider the timing of the funding rounds, as companies founded from 2010 onwards are included. For this success metric, the amount of money raised by each company founded in the first year 2010 is calculated until 2015.

However, a limitation of this success variable is that it relies on Crunchbase datasets, where not all companies disclose the funding amounts for each round. In some cases, only the number of investment rounds is reported.

5.3 Success measured in Number of Funding Rounds raised

For this reason, this part of the thesis introduces a third level of success: the number of funding rounds the startups raised.

5.4 Benefit of the Three-Layered Success Definition

This framework addresses weaknesses in individual predictors. Solely tracking acquisitions, mergers, or IPOs overlooks unicorns with high revenue but no exit. Adding funding amount predictions captures such companies, while funding round counts offer insights for companies not disclosing amounts. However, these alone are insufficient since fundraising success may not prevent bankruptcy. Combining all three ensures a more comprehensive evaluation of well-funded and sustainable startups.

5.5 Implementation of this three layered Success Definitions

The choice of a 5-year period, rather than 2, 3, or 4 years, was due to data limitations in the closed dataset, which only provided the total funding raised over a startup's lifetime without time-specific details. A shorter timeframe would have excluded many older companies, significantly reducing the dataset. By using a 5-year window, most closed companies—specifically those that ceased operations within 6 years—could be included, under the assumption that startups typically do not raise funds in their final year before closure. This assumption is supported by research indicating an average startup runway of 18 months (Vick, 2019; Brex, 2024), and further validated by analysis showing that 86.6% of merged or acquired companies did not raise funds in the year preceding these events (Appendix 1).

For IPOs, only two companies achieved this status between years 5 and 6, and they were

excluded to maintain consistency. Success for the first layer is defined as whether a company went bankrupt, achieved an IPO, was merged, or acquired within 5 years, with events occurring in the sixth year treated as neutral. For the second layer, total lifetime funding was used, assuming no additional funding in the final year, which also allowed for determining the number of funding rounds.

Restricting the dataset to companies closed within 6 years introduced some bias, excluding 1,500 closed companies and reducing the closed dataset to 2,780 companies. Similarly, open companies founded in 2020 were excluded, reducing the open dataset by 250 entries.

6. Experience and Education

As highlighted in the literature review, previous research indicates that a high level of education is a predictor of founder success, particularly in the early stages. However, the prestige of a university does not always correlate with long-term success, as alumni from less prestigious institutions, such as the University of Washington and the University of Waterloo, have also demonstrated notable success (Othman and Speiser, 2021).

To assess the influence of educational and professional experiences on startup success, this thesis introduces next to a unified score of company and university prestige, a scoring system for universities and companies. Universities are assigned success and failure scores based on the percentage of alumni in the dataset who succeeded or failed as founders. To balance accuracy and coverage, scores are calculated at three thresholds: universities with at least 5, 10, and 20 alumni founders in our dataset. This prevents unreliable scores from universities with too few alumni while maintaining sufficient predictive coverage. Similarly, a scoring system is applied to companies, complementing the previously calculated company prestige score for founder-market fit.

For each founder, the average of success scores derived from their education and professional experience is computed. Additionally, the number of universities attended and job experiences

is captured as separate features to provide further predictive power. This metric accounts for the founder's education level (e.g., whether they attended university) and offers insights into their experience level. While the number of jobs includes both internships and full-time positions without distinction, this limitation can be addressed in future research.

The scoring system, while limited by dataset size, offers valuable predictive variables for the final model and interesting insights. For instance, universities with the highest success scores tend to rank higher on average compared to those with the highest failure scores, although prestigious institutions appear in both groups (Appendix 2: Figure Success; Appendix 3: Figure Failure).

7. Founders connectivity

The literature review highlighted that connectivity is a key factor in a founder's success. In the context of founder connectivity, previous research has used the number of LinkedIn followers as a predictor of startup success. This part of the thesis will also incorporate the number of LinkedIn followers as a predictor to evaluate its relative importance.

8. LinkedIn Comments, Posts, Reactions, and Profile Descriptions

8.1 Numerical features

From the founders' LinkedIn activity, numerical metrics were extracted from posts, comments, and reactions to quantify their online engagement. For posts, the metrics include the total number of posts created by the founder, the number of reposts, repost frequency per post, timing of the posts (measured in days since posting), and engagement indicators such as the number of comments and overall interactions. These metrics are utilized to assess the founder's connectivity and popularity within their professional network.

For comments, the analysis captures the total number of comments posted by the founder and the engagement metrics (likes, comments, and reposts) associated with the posts they

commented on. This provides insights into whether the founder engages with niche, personalized content or broader, widely viewed content posted by individuals with higher popularity.

For reactions, the analysis includes the number of posts the founders reacted to and the timing of these posts, offering additional data on their engagement patterns and temporal activity.

8.2 Text Features

Textual predictors were extracted from founders' LinkedIn profile descriptions, comments, posts, and the posts they reacted to using a systematic four-layered approach. First, unsupervised learning techniques were employed to derive textual features. Second, sentiment analysis was performed using pre-trained models to evaluate the emotional tone of the texts. Third, supervised learning methods were applied to classify the texts and predict the likelihood of founder success based on the first-layer success definition. Lastly, personality traits were predicted for the founders based on their textual content.

To predict personality traits, a model was trained on labeled LinkedIn data and applied to the remaining profiles. The labels for LinkedIn profiles were obtained by matching X profiles to LinkedIn profiles via Crunchbase. The X data personality predictions were generated using a model trained on publicly available X posts annotated with personality traits of the respective authors.

The same analysis pipeline was applied to posts, comments, profile descriptions, and reacted-to content. Models requiring text embeddings were executed twice—once using TF-IDF embeddings and once with BERT Sentence Transformer embeddings. TF-IDF captures the relevance of words in a text (Ali and Quaiser, 2018), while BERT Sentence Transformers provide advanced embeddings that also encapsulate the semantic meaning of the text (Yin and Zhang, 2024)**.**

8.2.1 Unsupervised text feature training

Unsupervised models were implemented using K-Means clustering to identify broad topics within the textual data. Cluster topics were determined by extracting the most frequently occurring words in each cluster. Hyperparameter optimization was conducted through grid

search, using the Silhouette Score as the performance metric. The Silhouette Score measures the cohesion of data points within a cluster relative to their separation from other clusters, where -1 indicates poor clustering, 1 indicates perfect clustering, and values near 0 suggest overlap between clusters.

For founder posts, the Silhouette Score ranged from 0.013 to 0.0336, with TF-IDF vectorization outperforming BERT embeddings. For posts reacted to by founders, scores ranged from 0.0084 to 0.0271, where BERT embeddings demonstrated slightly better performance (Appendix 5: Kmeans Clusters of text features). Profile descriptions achieved Silhouette Scores ranging from 0.0085 to 0.0405, with SBERT embeddings combined with optimized K-Means clustering yielding scores more than twice as high as other methods. Comments exhibited scores between 0.02667 and 0.1262, with TF-IDF vectorization consistently producing superior results (Appendix 5: Kmeans Clusters of text features).

The results demonstrate the feasibility of clustering textual data, with some Principal Component Analysis (PCA) cluster visualizations showing distinct clusters, while others exhibit significant overlap (Appendix 5: Kmeans Clusters of text features). The combined use of TF-IDF and BERT embeddings provides flexibility, as performance varies depending on the dataset. In all cases, hyperparameter optimization improved the Silhouette Scores, indicating better clustering performance. Examining the TF-IDF PCA plots suggests that the clustering quality appears to deteriorate after hyperparameter optimization. This effect is primarily due to the increased number of clusters, which makes it visually challenging to distinguish which data points belong to specific clusters. The increased Silhouette Score suggest an improvement of the clusters (Appendix 5: Kmeans Clusters of text features).

Higher Silhouette Scores observed for comments, compared to other text types, are likely attributable to the higher amount of comment data since the scraping algorithm appears to extract more comment data.

Latent Dirichlet Allocation (LDA) was applied using Gensim to perform probabilistic topic modeling. LDA aims to uncover hidden topics within a collection of texts by assuming that each text is generated by a mixture of topics, with each topic represented by a distribution of words (Yu and Xiang 2023). LDA does not require embeddings and is therefore independent of TF-IDF and BERT vectorization.

For BERT vectorization, BERTopic was utilized. BERTopic integrates transformer-based embeddings, such as BERT sentence transformer embeddings, with traditional clustering techniques to generate coherent and interpretable topics. Additionally, BERTopic maps the distance between topics, providing a visual representation of their relationships (Appendix 6: BERTopic Clusters of Text Features).The visualization indicates that the distance between topics in LinkedIn profile descriptions appears smaller, with the topics being more dispersed compared to comments, which exhibit three distinct and well-separated topic areas. This observation aligns with the word cloud analysis, where comments frequently feature words such as “thank,” “love,” “congrats,” and “congratulations,” reflecting a consistent theme of congratulatory or appreciative remarks related to posts. In contrast, the word cloud for profile descriptions includes terms like “technology,” “company,” “product,” “experience,” and “business,” suggesting a broader and more diverse range of topics, as reflected in the BERTopic results (Appendix 4: Word Clouds).

8.2.2 Sentimental text feature training

For sentiment analysis, the Valence Aware Dictionary and Sentiment Reasoner (VADER), a rule-based model, was employed. VADER analyzes text using a predefined lexicon and heuristic rules, returning probabilities for positive, neutral, and negative sentiment categories. Additionally, it provides a compound score, a normalized metric summarizing the overall sentiment of the text. To complement this analysis, an API from Weights & Biases (W&B SDK) was utilized to further enhance sentiment evaluation, offering a broader perspective on

the sentiment distribution within the dataset.

8.2.3 Supervised text feature training

To ensure the integrity of the supervised learning process and prevent data leakage, an additional split of the training data was performed during feature training. The data used to train the final model was entirely separate from the data used to train the supervised features. These features were trained on the same half of the training data from which the success and failure scores for universities and companies were extracted. This approach ensured that the final dataset never implicitly encountered the target variable, even through features trained on it. As a result, the final prediction using the supervised features relied solely on the second half of the training dataset.

For supervised models using TF-IDF and BERT sentence embeddings, logistic regression, random forest classifier, XGBoost, and a neural network were implemented. Initially, basic versions of these models were run, followed by fine-tuning their hyperparameters using randomized search to save time, as grid search would have been too time-consuming. Since the dataset was imbalanced—with most companies categorized as neutral (neither failing nor succeeding according to the success definitions in this thesis)—additional steps were taken to address this issue.

To handle the imbalance, the hyperparameters were optimized to improve Macro-Averaged Accuracy. This metric calculates the accuracy for each category independently and then averages them, ensuring equal weight for each category regardless of its size. Furthermore, specific adjustments were made for each model.

For logistic regression and the random forest classifier, the class_weight parameter was set to "balanced," which adjusts weights by dividing the total number of samples by the product of the number of classes and the class frequency. This adjustment ensured that minority classes contributed more during training.

For XGBoost, the scale_pos_weight parameter was set to the ratio of negative to positive classes, biasing the model towards minority classes.

For the neural network, class_weight during training were set up to balance the importance of each class.

Additionally, SMOTE (Synthetic Minority Oversampling Technique) was applied to oversample the minority classes, creating a more balanced training set. This combination of techniques allowed the models to handle the dataset's inherent imbalance effectively while enhancing their ability to generalize to unseen data.

The results indicate that, for posts, comments, posts the founder reacted to, and profile descriptions, the Logistic Regression model based on SBERT embeddings achieves the best performance in terms of **Macro-Average Accuracy** (Appendix 7: Supervised Feature Training Performance). This superior performance can be attributed to a relatively high **Macro-Average Recall**, demonstrating that the model captures a relatively bigger proportion of actual true values in the classification task.

Furthermore, the results suggest that the profile description (highlighted in blue in Appendix 7) exhibits the highest overall predictive power, as evidenced by its consistently superior performance across **Macro-Average Accuracy**, **Macro-Average Precision**, and **Macro-Average Recall**. It is important to note that accuracy, as a performance measure, is less informative in this case due to the imbalanced nature of the dataset.

An evaluation of the overall performance of the classification models reveals that predicting the first layer of success definition using a single classification algorithm applied to individual text elements—such as a post, comment, post a founder reacted to, or a LinkedIn profile description—does not yield promising results, as the performance remains close to random (Appendix 7: Supervised Feature Training Performance). This outcome is expected, as it is unlikely that a founder's success can be accurately predicted based on a single text element

alone.

However, combining these text elements with a diverse range of extracted features provides a foundation for a more robust prediction model. Therefore, a comprehensive variety of feature extraction techniques has been applied to these texts. All the results from the supervised models, regardless of their individual performance, will be incorporated as features into the final prediction models. The final model will then determine the most relevant features for accurately predicting startup success.

Despite the limited predictive power of individual text elements, the results of the supervised feature training are still valuable. They demonstrate that even a single text component can provide slight indications of a startup's success, reinforcing the potential utility of these classifications as input features for the final prediction model.

8.2.4 Personality Prediction of the Founders the Texts Belong to

For the personality prediction, linear regression was employed as a baseline model, along with random forest, XGBoost, and neural networks, for both TF-IDF and BERT sentence transformer embeddings. While hyperparameter optimization was attempted using both grid search and randomized search during the model development process, neither approach yielded significant improvements in model performance. Across all models, the performance metrics remained almost identical. The mean squared error (MSE) consistently ranged between 0.002 and 0.003, while the R-squared value ranged between -0.5 and -0.0025.

A negative R-squared value indicates that the model performs worse than a simple baseline that predicts the mean of the target variable. This suggests that the models fail to explain the variance in the personality scores, meaning the target labels exhibit very low variance. Consequently, even the average prediction results in a high MSE, as evidenced by the negative R-squared value for models with an MSE of 0.002, where the average prediction outperforms the models.

This lack of variance poses a significant challenge for feature building, as the objective is to

explain the deviation from the mean using predictive features, which the models are unable to achieve. This limitation is likely attributed to substantial biases in the data, as previously discussed in the data pipeline part of the thesis. While the idea of labeling data based on predictions from X remains conceptually appealing, it appears to be impractical for effective application in this context.

The negative R-squared value indicates that the variance in personality cannot be explained by the personality predictions from individual text elements, rendering the prediction of personality based on these features infeasible. The extracted features do not provide meaningful information about the variance in founder personality. Consequently, this thesis does not pursue a final personality prediction based on these personality features.

However, these features will still be included in the final startup success prediction model, as they capture some variation in the text that may contribute to the overall prediction. It is important to note that the personality features do not reflect the actual personality of the founder but instead capture unrelated variations in the text that may offer **some** insight into the final prediction outcome.

For the supervised models, the final predictions were not directly extracted as features; instead, the probabilities of a given text belonging to each target variable were utilized. This approach ensures maximal information extraction from the features before integrating them into the final models. To obtain probabilities, the predict_proba function was applied for the Logistic Regression, Random Forest Classifier, and XGBoost Classifier models. For the neural network, the probabilities were derived from the softmax layer outputs. All predicted probabilities were subsequently incorporated as features in the final prediction model.

9. Final Success Prediction Model

In the final model, all features were merged into a single DataFrame with 785 features, including cluster and classification elements. The first layer of startup success, a categorical

variable requiring class imbalance handling, was predicted separately, while the second and third layers, both continuous variables, were predicted simultaneously to leverage shared patterns and dependencies. Models capable of handling missing values internally were used to avoid introducing bias from imputation. XGBoost and a neural network were compared: class imbalance in XGBoost was managed using the scale_pos_weight parameter, while in the neural network, class weights were calculated using compute_class_weight and applied during training. XGBoost achieved superior performance with a Macro-Averaged Precision of 0.57 and a Macro-Averaged Recall of 0.42 (Appendix 8: Final Models XGBoost First Layer Success).

For the prediction of the second and third layers of startup success, a comparison of **Random Forest**, **XGBoost**, and **LGBMRegressor** showed **LGBMRegressor** as the best-performing model. To assess feature importance, several data subsets were tested: experience and education alone, LinkedIn profile descriptions, posts/reactions/comments, and all data excluding followers (Appendix 8 and Appendix 10).

For the first-layer success prediction (mergers, acquisitions, IPOs), posts, reactions, and comments outperformed education and LinkedIn profile descriptions, highlighting the potential of text-based features for success prediction. However, caution is warranted due to **forward-looking bias**: recent LinkedIn activity may be influenced by a founder's success or failure. While **followers** slightly improved the model's Macro-Averaged Precision from **0.56** to **0.57** (recall remained at **0.42**), they also exhibit this bias. In contrast, experience, education, and static features like the LinkedIn “About” section are less prone to such issues.

Feature importance analysis revealed that for the first-layer success definition, **unsupervised text features** were more influential than supervised ones, while for the second and third layers, **supervised features** played a larger role. Consistent with prior research, **university prestige** was a top predictor for fundraising (second/third layers) but not for first-layer

success. Instead, the percentage of successful founders from an institution proved more insightful, underscoring the value of institutional success rates over rankings. Notably, **VC experience** emerged as the most relevant experience variable for predicting first-layer success but showed limited impact on fundraising outcomes, offering a novel insight.

The results demonstrate a notable predictive capacity, highlighting the importance of founders in startup success. For the first-layer success definition, a model with limited forward-looking bias—using only experience, education, and LinkedIn profile descriptions—achieved a significant performance improvement over the random baseline, with **Macro-Average Precision** increasing from **0.34** to **0.49** and **Macro-Average Recall** from **0.33** to **0.35**. This confirms that founders influence startup success, though the predictability remains limited. When including features with forward-looking bias, the performance further improved, with **Macro-Average Precision** reaching **0.57** and **Macro-Average Recall 0.42** (Appendix 8: Final Models XGBoost First Layer Success).

For the second and third success layers, education and experience explained some variance in the number of funding rounds but showed limited predictive power for posts, comments, and reactions. Combining LinkedIn profile descriptions with experience and education substantially improved performance. The final model reduced the **Mean Absolute Error (MAE)** for the third layer from **69.568 million** to **54.282 million** and improved **R-squared** from **-1.3758** to **0.0747**. For the second layer, the MAE improved from **1.0836** to **0.8309**, and R-squared increased from **-0.8959** to **0.1005** (Appendix 10: Final Models LGBMRegression Second, Third Layer Success). Overall, the results show stronger predictive performance for the first-layer success definition compared to the second and third layers.

10. Limitations

This study has several limitations that may impact the results and generalizability of the findings:

10.1 Small and Biased Personality-Labeled Dataset:

The LinkedIn profiles with personality labels represent a small and biased dataset, making it infeasible to achieve reliable personality predictions. The extracted text features failed to explain the variance in personality traits, limiting the predictive capacity of personality-based features.

10.2 Assumptions in Success Definition:

For the closed dataset, it was assumed that companies do not raise funds during their final year before closure. While practical, this assumption introduces bias, as it does not always reflect reality. Additionally, excluding companies that fail later than six years post-founding introduces further bias into the dataset.

10.3 Data Availability Bias:

The analysis was restricted to founders with profiles on both Crunchbase and LinkedIn, as well as corresponding company profiles on either platform. This exclusion inherently biases the results, as not all founders meet these criteria.

10.4 Reliance on Crunchbase Data:

The study assumes the accuracy of Crunchbase data, though funding information is frequently incomplete or missing. To mitigate this limitation, a third success metric—based on the number of funding rounds—was introduced.

10.5 Assumption of Founder Status:

LinkedIn scraping was conducted using filters to identify individuals labeled as "founders." However, this filter was not always accurate, leading to the inclusion of individuals who may not actually hold founder roles.

10.6 Forward-Looking Bias:

Models predicting future founder success, with the exception of those based on the experience

and education subset, are subject to forward-looking bias. This bias arises because the LinkedIn text features were extracted from recent posts, reactions, and comments, which may have been influenced by the founder's current success or failure.

These limitations should be considered when interpreting the results and highlight areas for improvement in future research.

11. Conclusions

This part of the thesis showed with the more diverse perspective of startup success the different features that are relevant for different facets of startup success. This allowed to examin, that university prestige plays a bigger role for fundraising, while for predicting if a founder will achieve an IPO or Mergers & Acquisition it is more relevant to look at the performance of previous university graduates of the university. The key insight of this thesis is that it emphasizes the potential importance of posts, reactions, and comments. This opens dors to new research possibilities. LinkedIn comments, posts, and reactions have never been used as startup success predictor in scientific papers before. Now we know that this might be an area with full potential. In general in previous research posts of X were used to predict the personality of founders with which they predict the personality as a feature of the success prediction of a startup. But why should be add a human made personality framework inbetween, if we can predict founder success of based on posts immediately. This might cause less loos of indications in the text that the founders are going to be successful. This thesis showed that based on posts comments and reactions one can predict if a founder is successful or not. The main question for further research is, if the posts, research, and comments section in LinkedIn are also good predictors of founder future success.

References

Alfons, Sirius Araya. 2022. „Impact of Previous Entrepreneurial Experience on Start up Evaluation & Success“. *Catolica Lisbon*. Available at: https://repositorio.ucp.pt/bitstream/10400.14/40420/1/203133005.pdf

„BERT — transformers 3.0.2 documentation“. o. J. Zugegriffen 11. Dezember 2024. https://huggingface.co/transformers/v3.0.2/model_doc/bert.html.

Braesemann, Stephany, Fabian, Fabian. 2023. „Personality of Founders Could Predict Start-up Success, Finds New Study“. 2023. https://www.oii.ox.ac.uk/news-events/personality-of-founders-could-predict-start-up-success-finds-new-study.

Brahmana, Rayenda Khresna, Doddy Setiawan, Evan Lau, and Maria Kontesa. 2024. "Do the Characteristics of Startup Founders Matter for Funding Performance?" *Journal of Indonesian Economy and Business*, 39(3): 328–346. Available at: https://doi.org/10.22146/jieb.v39i3.11841.

Devika Banerji, and Torsten Reimer. 2019a. „Startup founders and their LinkedIn connections: Are well-connected entrepreneurs more successful?“ *Computers in Human Behavior* 90 (Januar):46–52. https://doi.org/10.1016/j.chb.2018.08.033.

Devika Banerji, Torsten Reimer. 2019b. „Startup founders and their LinkedIn connections: Are well-connected entrepreneurs more successful?“ *Computers in Human Behavior* 90 (Januar):46–52. https://doi.org/10.1016/j.chb.2018.08.033.

Freiberg, Brandon, and Sandra C. Matz. 2023. “Founder Personality and Entrepreneurial Outcomes: A Large-Scale Field Study of Technology Startups.” *Proceedings of the National Academy of Sciences* 120 (19): e2215829120. https://doi.org/10.1073/pnas.2215829120.

Freiberg, Matz, Brandon, Sandra. 2024. „Startup Success: How Founder Personalities Shape Venture Outcomes | Columbia Business School“. 1. April 2024. https://business.columbia.edu/research-brief/research-brief/startups-founder-personalities-vc.

Gage, Deborah. 2014. "Why Startups Fail, According to Their Founders." *Fortune*, September 25. Available at: https://fortune.com/2014/09/25/why-startups-fail-according-to-their-founders/.

Gjurković, Matej, et al. 2021. "PANDORA Talks: Personality and Demographics on Reddit." *arXiv preprint arXiv:2004.04460*. Available at: https://doi.org/10.48550/arXiv.2004.04460.

Gompers, Paul, and Josh Lerner. 2001. "The Venture Capital Revolution." *Journal of Economic Perspectives*, 15(2): 145–168. Available at: https://www.aeaweb.org/articles?id=10.1257/jep.15.2.145.

Kimani, Bonface. 2024. „Impact of Entrepreneurial Education on Startup Success Rates in Kenya“. *International Journal of Entrepreneurship* 7 (2): 45–55. https://doi.org/10.47672/ije.2104.

MacMillan, Ian C., Robin Siegel, and P.N. Subba Narasimha. 1985. "Criteria Used by Venture Capitalists to Evaluate New Venture Proposals." *Journal of Business Venturing*, 1(1): 119–128. Available at: https://www.sciencedirect.com/science/article/abs/pii/0883902685900114.

McCarthy, Paul X., Xian Gong, Fabian Braesemann, Fabian Stephany, Marian-Andrei Rizoiu, and Margaret L. Kern. 2023a. „The Impact of Founder Personalities on Startup Success". *Scientific Reports* 13 (1): 17200. https://doi.org/10.1038/s41598-023-41980-y.

Othman, Abe, and Matthew Speiser. 2021. „How Does a Founder's Alma Mater Impact Their Startup's Markup Rate?" AngelList. 11. August 2021. https://www.angellist.com/blog/founder-schools.

Otterlei, Wik, Håkon, Hogstad. 2023. „Founder Success in Norwegian Startups: A Machine Learning Approach". Norwegian School of Economics. Available at: https://openaccess.nhh.no/nhh-xmlui/bitstream/handle/11250/3090327/masterthesis.pdf?sequence=1

Pinelli, Michele, Francesco Cappa, Stefano Franco, Enzo Peruffo, and Raffaele Oriani. 2022. „Too Much of Two Good Things: Effects of Founders' Educational Level and Heterogeneity on Start-Up Funds Raised". *IEEE Transactions on Engineering Management* 69 (4): 1502–16. https://doi.org/10.1109/TEM.2020.2991607.

Shanar, Hanna. 2023. "Exploring the Intricate Interplay of Education, Income and Entrepreneurial Success." *Entrepreneur. April 21, 2023.* https://www.entrepreneur.com/starting-a-business/exploring-the-intricate-interplay-of-education-income-and/449324.

Stenson, Daniel. 2023. „USING SOCIAL MEDIA AND PERSONALITY PREDICTIONS TO ANTICIPATE STARTUP SUCCESS", *LUP Student Pap*ers. Available at: https://lup.lub.lu.se/student-papers/search/publication/9141885

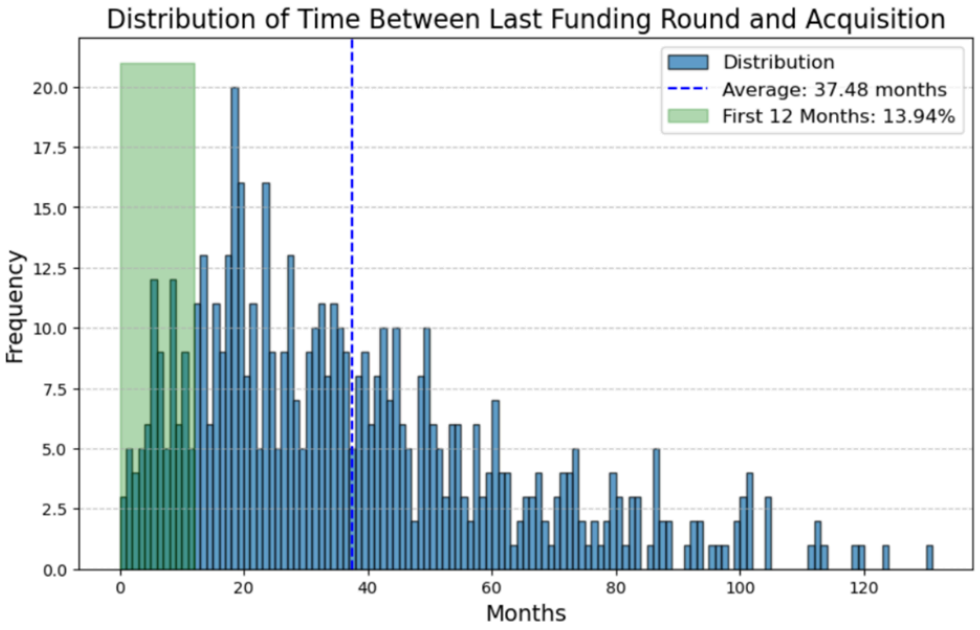
Virágh, Enikő Anna, Gigi Tímár, and Krisztina Pecze. 2024. „Startup Success from the Founder's Perspective", Februar. https://doi.org/10.1556/204.2023.00029.2002

Xian Gong, Fabian Braesemann, Fabian Stephany, Marian-Andrei Rizoiu, andMargaret L. Kern. 2023b. „The Impact of Founder Personalities on Startup Success". *Scientific Reports* 13 (1): 17200. https://doi.org/10.1038/s41598-023-41980-y.

Western Countries. 2024. „Western Countries 2024". https://worldpopulationreview.com/country-rankings/western-countries.

Appendix

Appendix 1: **Success definition including closed startups that got merged and acquired of 6 years after they go funded.**



Appendix 2: **Figure Success**

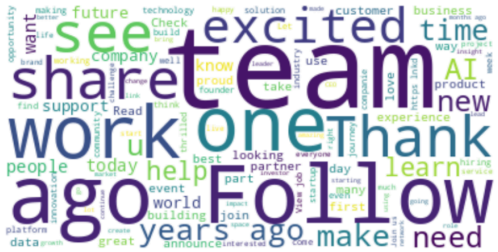
	University	Success_Score	Failure_Score	Count
0	University of Toronto - Rotman School of Manag...	25.000000	6.250000	16
1	Caltech	22.222222	14.814815	27
2	Brandeis University	18.181818	0.000000	11
3	Wesleyan University	18.181818	0.000000	11
4	Delft University of Technology	18.181818	0.000000	11
5	William & Mary	18.181818	18.181818	11
6	Concordia University	16.666667	0.000000	12
7	University of Warwick	16.666667	8.333333	12
8	Y Combinator	16.438356	32.876712	73
9	Stanford University Graduate School of Business	15.555556	2.222222	90
10	University of London	15.384615	0.000000	13
11	University of Michigan - Stephen M. Ross Schoo...	15.000000	0.000000	20
12	Western Governors University	14.285714	0.000000	14
13	The University of British Columbia	14.285714	7.142857	42
14	The Johns Hopkins University	14.285714	10.714286	28
15	University of San Francisco	14.285714	7.142857	14
16	Indian Institute of Technology, Madras	13.636364	4.545455	22
17	Brown University	13.513514	8.108108	37
18	University of Wisconsin-Madison	13.513514	8.108108	37
19	Carnegie Mellon University	12.500000	19.642857	56
20	Tsinghua University	12.500000	18.750000	16
21	University of Washington	12.121212	10.606061	66
22	Birla Institute of Technology and Science, Pilani	11.764706	17.647059	17
23	Loyola Marymount University	11.764706	5.882353	17
24	Tufts University	11.538462	11.538462	26
25	University of North Carolina at Chapel Hill	11.538462	19.230769	26
26	Stanford University	11.336032	9.311741	247
27	Virginia Tech	11.111111	22.222222	18
28	University of Alberta	11.111111	0.000000	18
29	New York University	10.909091	9.090909	55
30	Harvard Business School Online	10.714286	7.142857	28
31	Texas McCombs School of Business	10.526316	10.526316	19
32	University of Massachusetts Amherst	10.344828	3.448276	29
33	The University of Texas at Austin	10.169492	3.389831	59
34	Queen's University	10.000000	0.000000	20
35	UCLA Anderson School of Management	9.677419	3.225806	31
36	Technion - Israel Institute of Technology	9.523810	19.047619	42
37	Rensselaer Polytechnic Institute	9.523810	0.000000	21
38	Pontificia Universidad Católica de Chile	9.090909	9.090909	11
39	Bangalore University	9.090909	9.090909	11
40	University of Victoria	9.090909	0.000000	11
41	University of Maryland Baltimore County	9.090909	0.000000	11
42	Georgetown University	9.090909	18.181818	33
43	University of California, Berkeley, Haas Schoo...	9.090909	3.030303	33
44	Harvard University	8.759124	10.948905	137
45	Ben-Gurion University of the Negev	8.695652	0.000000	23
46	NYU Stern School of Business	8.571429	0.000000	35
47	University of Colorado Boulder - Leeds School ...	8.333333	0.000000	12
48	Reichman University	8.333333	8.333333	24
49	Penn State University	8.333333	10.416667	48

Appendix 3: **Figure Failure**

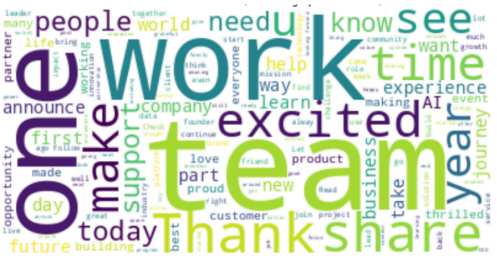
	University	Success_Score	Failure_Score	Count
0	University College Dublin	0.000000	42.857143	14
1	Reforge	0.000000	36.363636	11
2	Y Combinator	16.438356	32.876712	73
3	University of Nebraska-Lincoln	0.000000	30.769231	13
4	Università Bocconi	0.000000	30.769231	13
5	Indian Institute of Technology, Delhi	0.000000	29.411765	17
6	USC Marshall School of Business	5.882353	23.529412	17
7	University of Central Florida	7.692308	23.076923	13
8	Oregon State University	7.692308	23.076923	13
9	MIT Sloan School of Management	6.666667	22.222222	45
10	Virginia Tech	11.111111	22.222222	18
11	The University of Georgia	0.000000	20.000000	15
12	Southern Methodist University	0.000000	20.000000	15
13	Carnegie Mellon University	12.500000	19.642857	56
14	University of North Carolina at Chapel Hill	11.538462	19.230769	26
15	Technion - Israel Institute of Technology	9.523810	19.047619	42
16	University of Miami	6.250000	18.750000	16
17	Tsinghua University	12.500000	18.750000	16
18	William & Mary	18.181818	18.181818	11
19	Georgetown University	9.090909	18.181818	33
20	Savannah College of Art and Design	0.000000	18.181818	11
21	National University of Singapore	0.000000	18.181818	11
22	Birla Institute of Technology and Science, Pilani	11.764706	17.647059	17
23	Massachusetts Institute of Technology	7.042254	17.605634	142
24	Tulane University	0.000000	16.666667	18
25	Florida International University	0.000000	16.666667	12
26	Saïd Business School, University of Oxford	8.000000	16.000000	25
27	Michigan State University	0.000000	15.789474	19
28	University of Chicago	0.000000	15.384615	26
29	The Hebrew University of Jerusalem	3.846154	15.384615	26
30	Boston College	5.000000	15.000000	20
31	Babson College	3.703704	14.814815	27
32	Caltech	22.222222	14.814815	27
33	Georgia Institute of Technology	6.451613	14.516129	62
34	University of Michigan	4.285714	14.285714	70
35	Northwestern University - Kellogg School of Ma...	4.081633	14.285714	49
36	Columbia Business School	2.222222	13.333333	45
37	The University of Chicago Booth School of Busi...	6.666667	13.333333	30
38	UCLA	7.228916	13.253012	83
39	Duke University	4.347826	13.043478	46
40	Washington University in St. Louis	6.250000	12.500000	16
41	University of Virginia	0.000000	12.500000	48
42	University of California, Berkeley	6.666667	12.000000	150
43	Harvard Business School	6.569343	11.678832	137
44	Tufts University	11.538462	11.538462	26
45	Northeastern University	2.272727	11.363636	44
46	DePaul University	0.000000	11.111111	18
47	University of Missouri-Columbia	0.000000	11.111111	18
48	Harvard Business School Executive Education	5.555556	11.111111	18
49	Syracuse University	0.000000	11.111111	18

Appendix 4: **Word Clouds**

Posts



Reactions



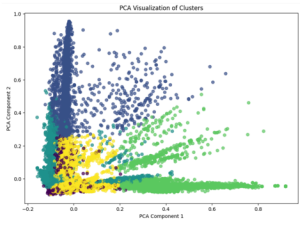
Comments

Profile description

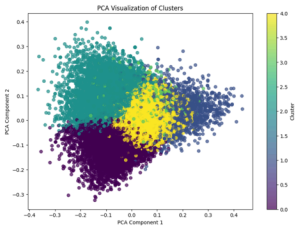
Appendix 5: **Kmeans Clusters of text features**

Kmeans Clustering TF-IDF

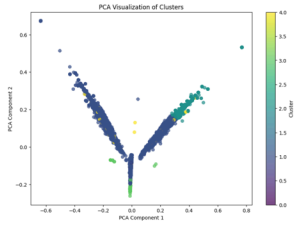
Posts: Silhouette Score 0.0228



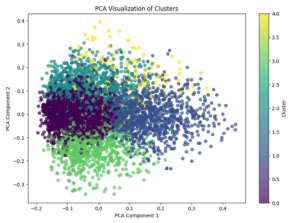
Reactions: Silhouette Score 0.0084



Comments: Silhouette Score 0.0648

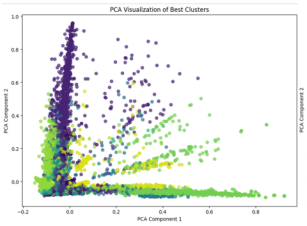


Profile Description: Silhouette Score 0.0085

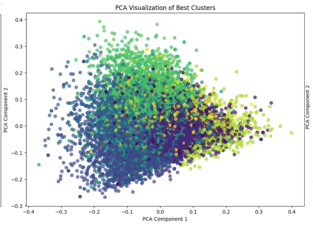


Kmeans Clustering TF-IDF optimized

Posts: Silhouette Score 0.0324



Reactions: Silhouette Score 0.0132

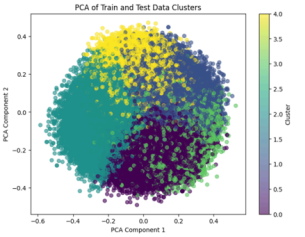


Comments: Silhouette Score 0.1262

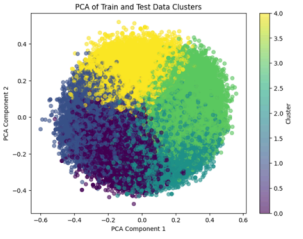
Profile Description: Silhouette Score 0.0143

Kmeans Clustering SBERT

Posts: Silhouette Score 0.0130



Reactions: Silhouette Score 0.0115

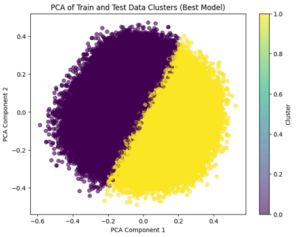


Comments: Silhouette Score 0.0267

Profile Description: Silhouette Score 0.0236

Kmeans Clustering SBERT optimized

Posts: Silhouette Score 0.0233

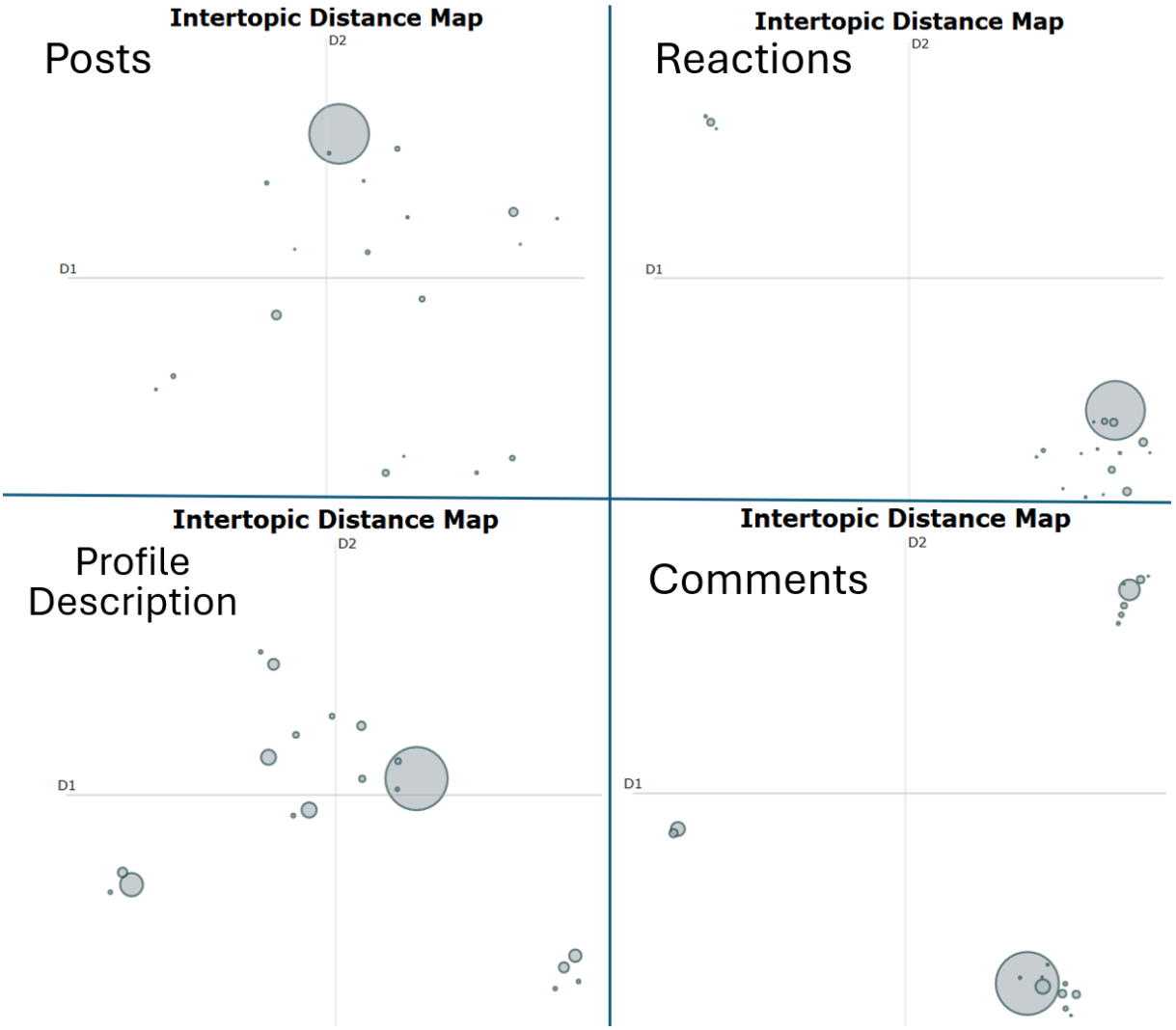


Reactions: Silhouette Score 0.0271

Comments: Silhouette Score 0.0408

Profile Description: Silhouette Score 0.0405

Appendix 6: **BERTopic Clusters of Text Features**



Appendix 7: **Supervised Feature Training Performance**

Modelname_posts	Accuracy	*Macro-Averaged Accuracy*	Precision (Macro avg)	Recall (Macro avg)	F1-Score (Macro avg)	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
Logistic_Regression_TF-IDF_posts	0.5847	*0.3656*	0.35	0.37	0.32	0.79	0.58	0.67
Random_Forest_TF-IDF_posts	0.8342	*0.3513*	0.37	0.35	0.35	0.79	0.83	0.81
Gradient_Boosting_TF-IDF_posts	0.7473	*0.3695*	0.36	0.37	0.36	0.79	0.75	0.77
Neural_Network_TF-IDF_posts	0.8743	*0.3366*	0.41	0.34	0.32	0.79	0.87	0.82
Logistic_Regression_Optimized_TF-IDF_posts	0.5781	*0.3642*	0.35	0.36	0.32	0.79	0.58	0.66
Random_Forest_Optimized_TF-IDF_posts	0.8297	*0.3531*	0.37	0.35	0.35	0.79	0.83	0.81
XGBoost_Optimized_TF-IDF_posts	0.7711	*0.3712*	0.36	0.37	0.36	0.79	0.77	0.78
Logistic_Regression_BERT_posts	0.4845	*0.3899*	0.36	0.39	0.3	0.81	0.48	0.59
Random_Forest_BERT_posts	0.8657	*0.3447*	0.43	0.34	0.34	0.8	0.87	0.82
Gradient_Boosting_BERT_posts	0.8077	*0.3619*	0.37	0.36	0.36	0.79	0.81	0.8
Neural_Network_BERT_posts	0.8770	*0.3336*	0.38	0.33	0.31	0.79	0.88	0.82
Logistic_Regression_Optimized_BERT_posts	0.4824	*0.3913*	0.36	0.39	0.3	0.81	0.48	0.58
Random_Forest_Optimized_BERT_posts	0.8644	*0.3448*	0.42	0.34	0.34	0.79	0.86	0.82
XGBoost_Optimized_BERT_posts	0.6957	*0.3826*	0.36	0.38	0.36	0.8	0.7	0.74
Modelname_comments	Accuracy	*Macro_Averaged_Accuracy*	Precision_macro_avg	Recall_macro_avg	F1_Score_macro_avg	Weighted_Precision	Weighted_Recall	Weighted_F1_Score
Logistic_Regression_TF-IDF_comments	0.4896	*0.3561*	0.34	0.36	0.29	0.78	0.49	0.59
Random_Forest_TF-IDF_comments	0.7503	*0.3404*	0.34	0.34	0.34	0.78	0.75	0.76
Gradient_Boosting_TF-IDF_comments	0.7001	*0.3497*	0.34	0.35	0.34	0.78	0.7	0.74
Neural_Network_TF-IDF_comments	0.8723	*0.3342*	0.36	0.33	0.31	0.78	0.87	0.82
Logistic_Regression_Optimized_TF-IDF_comments	0.4873	*0.3558*	0.34	0.36	0.29	0.78	0.49	0.59
Random_Forest_Optimized_TF-IDF_comments	0.7506	*0.3392*	0.34	0.34	0.34	0.78	0.75	0.76
XGBoost_Optimized_TF-IDF_comments	0.7991	*0.3440*	0.35	0.34	0.34	0.78	0.8	0.79
Logistic_Regression_BERT_comments	0.4116	*0.3620*	0.34	0.36	0.27	0.79	0.41	0.52
Random_Forest_BERT_comments	0.8569	*0.3361*	0.35	0.34	0.33	0.78	0.86	0.81
Gradient_Boosting_BERT_comments	0.7996	*0.3378*	0.34	0.34	0.34	0.78	0.8	0.79
Neural_Network_BERT_comments	0.8739	*0.3341*	0.38	0.33	0.31	0.78	0.87	0.82
Logistic_Regression_Optimized_BERT_comments	0.4089	*0.3586*	0.34	0.36	0.27	0.79	0.41	0.51
Random_Forest_Optimized_BERT_comments	0.8556	*0.3358*	0.34	0.34	0.33	0.78	0.86	0.81
XGBoost_Optimized_BERT_comments	0.7058	*0.3492*	0.34	0.35	0.34	0.78	0.71	0.74
Neural_Network_Optimized_BERT_comments	0.7124	*0.3378*	0.34	0.34	0.33	0.78	0.71	0.74
Modelname_about	Accuracy	*Macro_Averaged_Accuracy*	Precision_macro_avg	Recall_macro_avg	F1_Score_macro_avg	Weighted_Precision	Weighted_Recall	Weighted_F1_Score
Logistic_Regression_TF-IDF_about	0.6407	*0.3962*	0.36	0.4	0.35	0.81	0.64	0.71
Random_Forest_TF-IDF_about	0.8822	*0.3377*	0.4	0.34	0.33	0.81	0.88	0.83
Gradient_Boosting_TF-IDF_about	0.8407	*0.3471*	0.37	0.35	0.35	0.8	0.84	0.82
Neural_Network_TF-IDF_about	0.878	*0.3376*	0.39	0.34	0.33	0.8	0.88	0.83
Logistic_Regression_Optimized_TF-IDF_about	0.6826	*0.3764*	0.36	0.38	0.35	0.81	0.68	0.74
Random_Forest_Optimized_TF-IDF_about	0.8799	*0.3363*	0.36	0.34	0.32	0.8	0.88	0.83
XGBoost_Optimized_TF-IDF_about	0.8401	*0.3509*	0.36	0.35	0.35	0.8	0.84	0.82
Logistic_Regression_BERT_about	0.5762	*0.4005*	0.36	0.4	0.33	0.82	0.58	0.67
Random_Forest_BERT_about	0.8796	*0.3416*	0.39	0.34	0.33	0.81	0.88	0.84
Gradient_Boosting_BERT_about	0.8729	*0.3431*	0.39	0.34	0.34	0.8	0.87	0.83
Neural_Network_BERT_about	0.8874	*0.3332*	0.3	0.33	0.31	0.79	0.89	0.83
Logistic_Regression_Optimized_BERT_about	0.5835	*0.3867*	0.35	0.39	0.33	0.82	0.58	0.67
Random_Forest_Optimized_BERT_about	0.8798	*0.3381*	0.38	0.34	0.33	0.8	0.88	0.83
XGBoost_Optimized_BERT_about	0.8483	*0.3640*	0.39	0.36	0.37	0.81	0.85	0.83
Custom_Model_BERT_about	0.8281	*0.3483*	0.35	0.35	0.35	0.8	0.83	0.81
Modelname_reactions	Accuracy	*Macro-Averaged Accuracy*	Precision (Macro avg)	Recall (Macro avg)	F1-Score (Macro avg)	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
Logistic_Regression_TF-IDF_reactions	0.6025	*0.3530*	0.34	0.35	0.33	0.76	0.6	0.67
Random_Forest_TF-IDF_reactions	0.8311	*0.3422*	0.36	0.34	0.34	0.76	0.83	0.79
Gradient_Boosting_TF-IDF_reactions	0.7478	*0.3477*	0.35	0.35	0.35	0.76	0.75	0.75
Neural_Network_TF-IDF_reactions	0.8620	*0.3363*	0.4	0.34	0.32	0.77	0.86	0.8
Logistic_Regression_Optimized_TF-IDF_reactions	0.5974	*0.3522*	0.34	0.35	0.32	0.76	0.6	0.67
Random_Forest_Optimized_TF-IDF_reactions	0.8255	*0.3484*	0.35	0.34	0.34	0.76	0.83	0.79
XGBoost_Optimized_TF-IDF_reactions	0.7515	*0.3393*	0.34	0.34	0.34	0.76	0.75	0.76
Logistic_Regression_BERT_reactions	0.4763	*0.3802*	0.35	0.38	0.3	0.78	0.48	0.57
Random_Forest_BERT_reactions	0.8476	*0.3452*	0.39	0.35	0.34	0.77	0.85	0.8
Gradient_Boosting_BERT_reactions	0.8069	*0.3530*	0.36	0.35	0.35	0.77	0.81	0.79
Neural_Network_BERT_reactions	0.8644	*0.3336*	0.38	0.33	0.31	0.77	0.86	0.8
Logistic_Regression_Optimized_BERT_reactions	0.4744	*0.3850*	0.36	0.38	0.3	0.79	0.47	0.57
Random_Forest_Optimized_BERT_reactions	0.8457	*0.3453*	0.38	0.35	0.34	0.77	0.85	0.8
XGBoost_Optimized_BERT_reactions	0.7046	*0.3683*	0.36	0.37	0.36	0.77	0.7	0.74

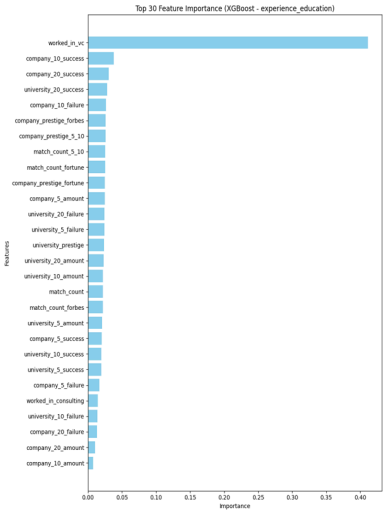
Appendix 8: **Final Models XG Boost First Layer Success**

Metric	Random_Model	experience_education	education_experience_about	posts_reaction_comments	all_data_excluding_followers	all_data_including_followers
Accuracy	0,8447	0,9037	0,9326	0,9739	0,9135	0,9159
Precision (Class -1)	0,05	0.09	0.35	0.08	0.37	0.43
Recall (Class -1)	0,04	0.01	0.03	0.02	0.17	0.19
F1-Score (Class -1)	0,05	0.02	0.05	0.03	0.23	0.27
Precision (Class 0)	0,92	0.92	0.94	0.98	0.93	0.93
Recall (Class 0)	0,92	0.99	1.0	1.0	0.99	0.99
F1-Score (Class 0)	0,92	0.95	0.97	0.99	0.96	0.96
Precision (Class 1)	0,04	0.12	0.18	0.46	0.38	0.34
Recall (Class 1)	0,04	0.02	0.02	0.08	0.09	0.09
F1-Score (Class 1)	0,04	0.04	0.04	0.14	0.15	0.14
Macro Avg Precision	0,34	0.37	0.49	0.51	0.56	0.57
Macro Avg Recall	0,33	0.34	0.35	0.37	0.42	0.42
Macro Avg F1-Score	0,33	0.34	0.35	0.39	0.45	0.46
Weighted Avg Precision	0,84	0.85	0.89	0.96	0.88	0.89
Weighted Avg Recall	0,84	0.9	0.93	0.97	0.91	0.92
Weighted Avg F1-Score	0,84	0.87	0.91	0.97	0.89	0.9

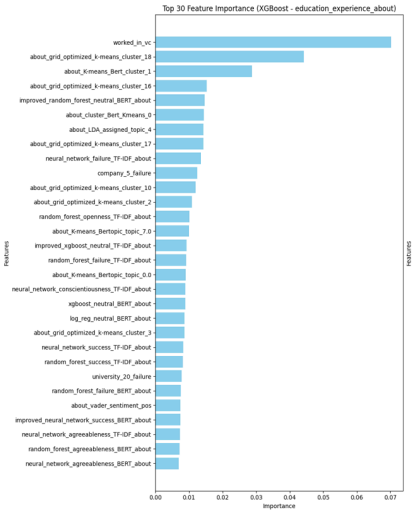
Appendix 9: **Feature Importance XG Boost First Layer Success**

XG Boost First Layer Success Definition

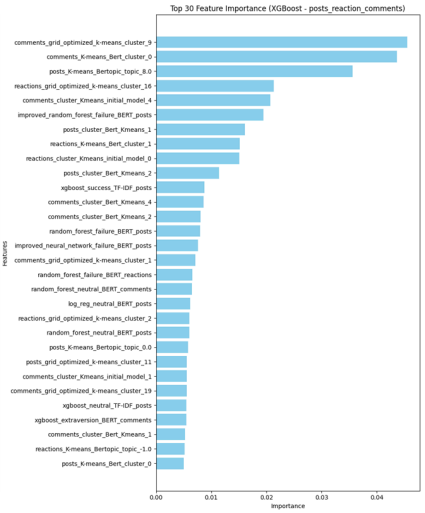
experience_education



education_experience_about

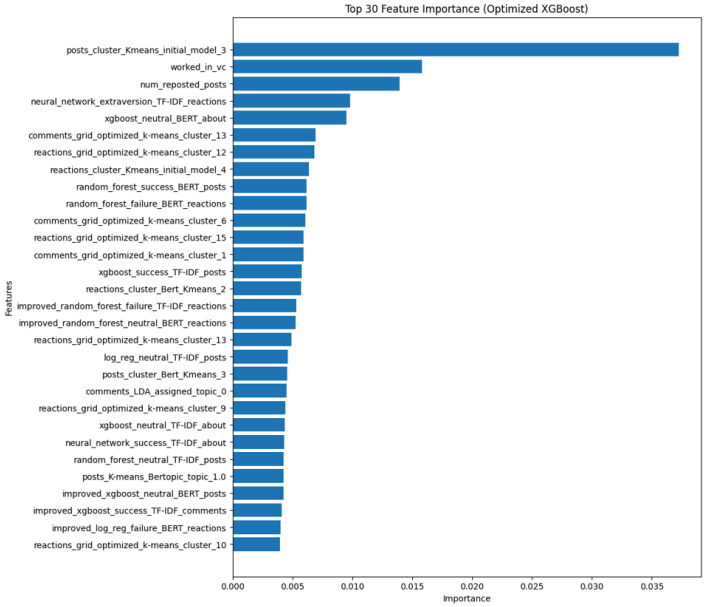


posts_reaction_comments

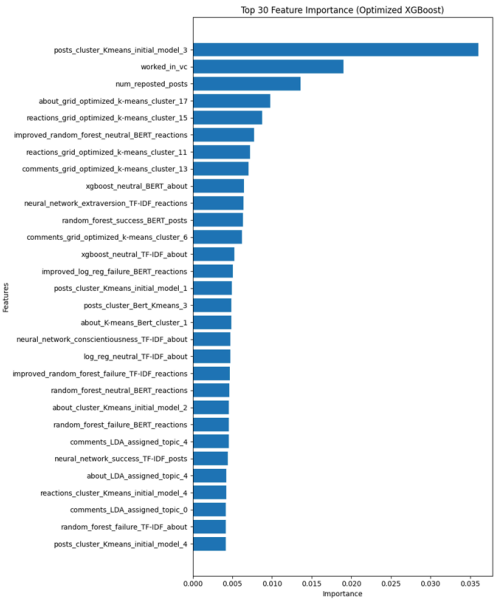


XG Boost First Layer Success Definition

all_data_excluding_foll



all_data_including_follower



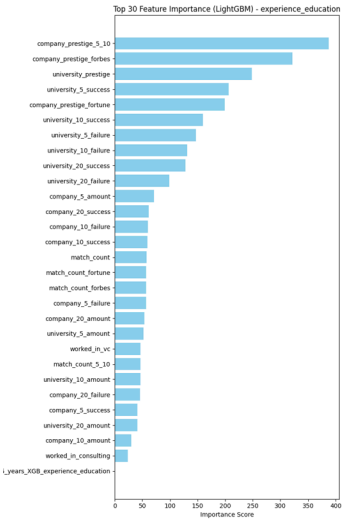
Appendix 10: **Final Models LGBMRegression Second, Third Layer Success**

Metric	Random_Model	experience_education	education_experience_abo	posts_reaction_comments	all_data_excluding_follower	all_data_including_followers
MAE_layer2	69.568.985.92	57.281.879,84	57.556.339,39	45.746.448,25	54.501.139,19	54.282.250,40
MSE_layer2	5,93E+31	29.748.830.817.636.000	2.635.429.563.455.220	21.130.436.639.163.300	27.856.305.520.221.600	27.441.172.031.206.900
R_layer2	-1,3758	-0,0031	0,0141	-0,0242	0,0607	0,0747
MAE_layer3	1,0836	0,8637	0,8096	0,7753	0,8291	0,8309
MSE_layer3	0,2576	1,3136	1,2024	1,0254	1,2136	1,2223
R_layer3	-0,8959	0,0333	0,0460	-0,0122	0,1069	0,1005

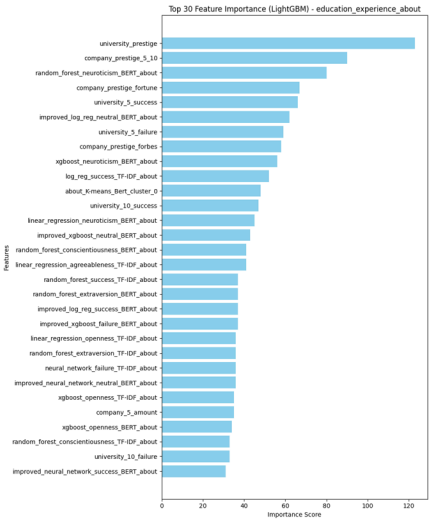
Appendix 11: **Feature Importance LGBMRegression Second, Third Layer Success**

LGMBRegressor Second & Third Layer Success Definition

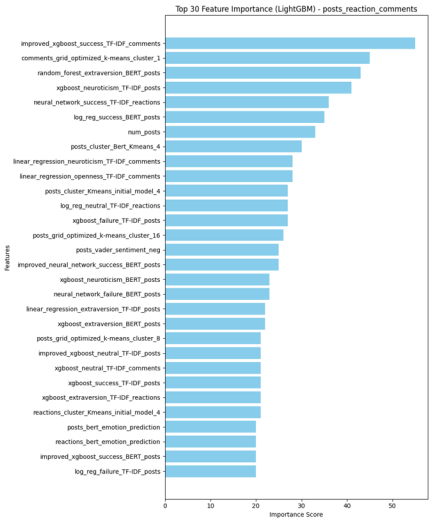
experience_education



education_experience_about

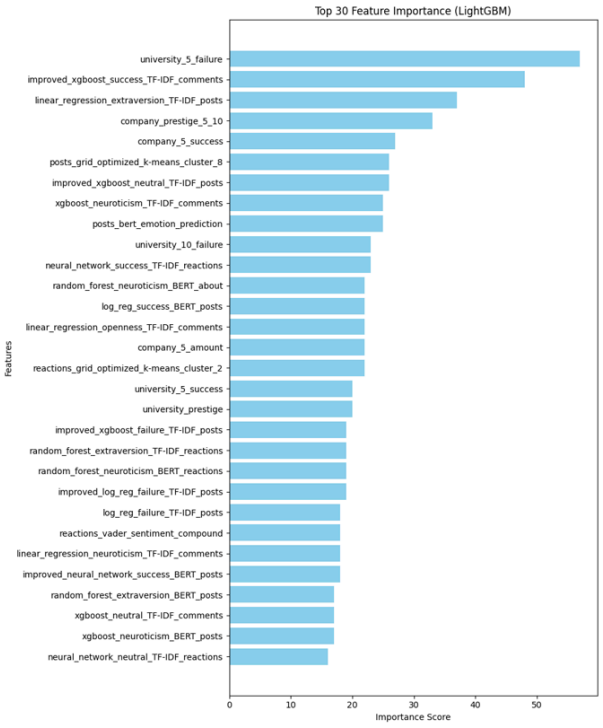


posts_reaction_comments



LGMBRegressor Second & Third Layer Success Definition

all_data_excluding_followers



all_data_including_follower