PREDICTING STARTUP SUCCESS USING PUBLICLY AVAILABLE DATA

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Emily Gavrilenko

December 2022

COMMITTEE MEMBERSHIP

TITLE: Predicting Startup Success Using Publicly Available Data

AUTHOR: Emily Gavrilenko

DATE SUBMITTED: December 2022

COMMITTEE CHAIR: Foaad Khosmood, Ph.D.
Professor of Computer Science

COMMITTEE MEMBER: Mahdi Rastad, Ph.D.
Associate Professor in Finance

COMMITTEE MEMBER: Alexander Dekhtyar, Ph.D.
Professor of Computer Science

ABSTRACT

Predicting Startup Success Using Publicly Available Data

Emily Gavrilenko

This paper explores whether online data about a company, particularly general company data, previous funding events, published news articles, internet presence, and social media activity can be used to identify fast-growing and high-performing companies. Data collected from Crunchbase, the Google Search API, and Twitter was used to predict whether a company will raise a round of funding within a fixed time horizon.

A total of ten machine learning models were evaluated and the CatBoost ensemble method achieved the best performance at predicting future funding rounds. The adaptable prediction model can be used to predict funding 1-5 years into the future, with a variable cutoff threshold to favor either precision or recall. The culmination of this work is a real-time prediction pipeline that outputs the probability of a company raising funding, along with an extensive feature analysis showing what features are the most crucial to a startup's success.

ACKNOWLEDGMENTS

A huge thanks to:

- My amazing parents, for supporting me and throughout my educational journey and always believing in my success.
- My brothers, Dennis and Maxim, for always giving me a reason to smile.
- My advisor, Dr. Khosmood, for teaching me everything I know about Natural Language Processing and for keeping me on track for graduation.
- My committee members, Dr. Rastad and Dr. Dekhtyar, for all the great feedback and for taking part in my thesis defense.
- Kenny, for helping with the frontend for the Startup Tracker Website.
- Sadra, for all the great ideas and guidance during the past year.
- My co-founders, Johnny and Josh, for being the best Ryde partners and always being there when I needed emotional support.

Shoutout to Dr. Rastad for the project idea!

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The past few decades have seen an unprecedented number of startups, with technical centers such as the Silicon Valley attracting bright individuals seeking to bring the newest technology to market. Highly innovative, disruptive, and profitable companies are chased by investors, and accurately predicting these unicorns is extremely difficult due to the high level of uncertainty and risk associated with developing startups [29]. Success can be measured in many different ways, including revenue, merger and acquisition, and securing funding, with the last one being the dominant method amongst researchers [29]. 5-10% of startups typically fall under the successful classification, categorized by continuous growth and a growing user base. The unicorn milestone, a company valuation of 1 billion dollars, is the pinnacle of success, coined in 2013 by venture capitalist Aileen Lee to describe the rarity of such ventures [25]. Venture capitalists and investors spend countless hours interviewing founders, looking through released company data, and relying on their gut feelings to ultimately make a decision on whether or not to invest. These methods are often biased and subjective, difficult to teach and difficult to explain to founders why they didn't make the cut.

Recent developments in machine learning have ignited a new field of research into using available public data to predict successful companies. Statistical models are able to find hidden trends in large amounts of data, unconstrained by human social factors and cognitive limitations [29]. Startup databases such as CrunchBase contain information on hundreds of thousands of startups that can be mined for organizational data, founder information, and funding rounds received [25]. However, while current machine learning approaches dominate in analyzing "hard" data, they fail to capture “soft” skills such as creativity and innovativeness crucial in leading a business to success [29]. Previous research has explored using solely CrunchBase data [25], CrunchBase data and human computation [29], and finally CrunchBase data with web scraping to maximize the accuracy of machine learning models [41]. Recent studies have sought to utilize the “wisdom of crowds” to capture “soft” skills in organizations, with research showing that collective intelligence reduces the noise and biases of individual predictions [29]. A large group of untrained individuals performs as well if not better than a small group of specialized experts, and these human insights show huge promise in uncovering the hidden gems amongst startups.

This paper investigates whether the “wisdom of crowds” can be captured through online, publicly available data and combined with hard, factual data to improve model performance. This includes social media sources, specifically Twitter, and online news articles collected through Crunchbase and the Google Search Engine. Tweets and news headlines referencing a company are collected and linguistic features are mined to uncover the author’s sentiment and the topic being discussed. These linguistic features, along with hard statistical data about the company collected from Crunchbase, are used to predict whether or not a company is classified as “successful.”

Previous research in startup success prediction has focused on different metrics of success, ranging from an IPO, M&A, founding round, or continuous operation for 5+ years. This thesis uses a fundraising milestone to classify a company as successful. An investor’s main goal is to see a return on investment, and that is achieved when a company raises an additional funding round. Since very few companies go public or become acquired, this fundraising metric will best capture a startup’s success.

In all, this thesis makes five important contributions to the field of startup success prediction. The first is the largest collected feature set on startup companies to date, consisting of general company data, previous funding rounds, published news articles, Google Search results, and Twitter data. The second contribution is a fine-tuned, topic classifier that can identify major company events. The third contribution is a startup tracker website for viewing company Twitter trends, information that is currently extremely difficult to manually gather and compute. The fourth contribution is an extensive feature analysis, discussing what features had a high correlation to company success. The final contribution is a real-time prediction pipeline for evaluating a company's investment potential using a company name, start date, and allotted time frame. This tool can be used by investors and the general public to learn more about a startup's likelihood of success and can be used to help make informed decisions on future investments.

Chapter 2 of this thesis gives a detailed overview of the topics, methods, and data sources mentioned throughout the paper. Chapter 3 details a comprehensive literature review of related works in startup success prediction, stock market prediction, and event prediction using Twitter data. Chapter 4 discusses how the dataset was created and used to train the classification model. Chapter 5 explains the experimentation setup and Chapter 6 highlights the results of the prediction models. Chapter 7 concludes with closing remarks and directions for future work.

Chapter 2

BACKGROUND

2.1 Startup Industry

There are approximately 472 million entrepreneurs worldwide, working on developing their ideas into highly innovative, disruptive, and profitable companies [33]. On average, there are about 305 million total startups created each year, all competing for funding and consumers [33]. A typical startup begins its journey in the pre-seed phase, where the product is still an idea or prototype and operations are self-funded by the founders and family friends. After the idea and product starts gaining traction it enters the seed funding phase. During this stage, rich, risk-taking individuals invest their own money to finance these rising startups, helping fund market research and product development. Once a business has established a user base and started to bring in revenue, they may apply for Series A funding, typically from a Venture Capital firm, often receiving an average of $15.6 million in funding to continue growing their business. Further growth may lead to Series B and Series C funding, with hedge funds, investment banks, and private equity firms additionally financing the company's growth. During these stages, a company may acquire new businesses and expand to new markets to grow its product and user base. An initial public offering (IPO) is typically the final stage for a startup, a period during which market shares are opened up to the public on a general stock exchange [18]. The complete startup timeline from ideation to IPO can be viewed in Figure 2.1.

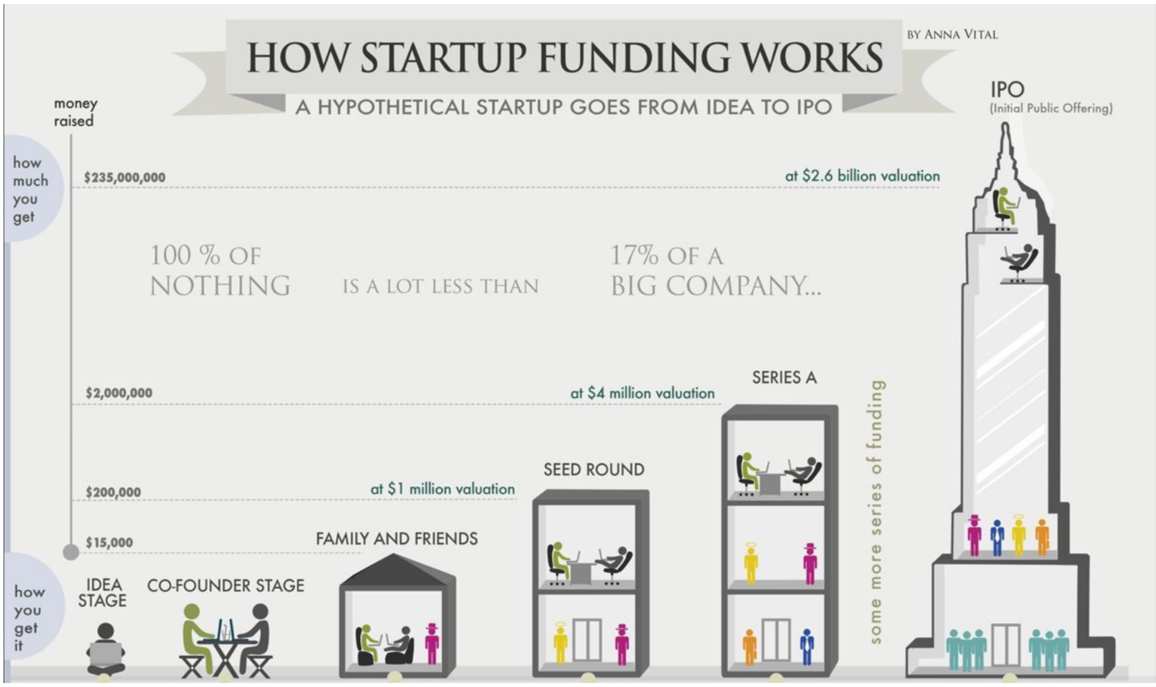


Figure 2.1: Funding Stages [10]

2.2 Predicting Startup Success

Predicting startup success has been the focus of many researchers and investors for decades. Initially, the methods of leading venture capitalists were studied and their advice on background research, founder interviews, and gut feeling reactions was the core focus of many rising investors. However, these subjective tips are full of personal biases and unexplainable mental models that are gained through years of experiences and are difficult to replicate. The recent rise of machine learning and big data has revolutionized the way researchers approach startup analysis. Statistical models are taking center stage in the attempt to find hidden trends that point to unicorns in the sea of startups competing for seed funding. Startup databases such as CrunchBase, Pitchbook, and CB-Insights contain information on general company data, team breakdown, and obtained funding for hundreds of thousands of startups [39]. Previous research using solely this data has had limited success, partially due to the training data containing funding information typically unavailable during the investment stage, and partially because relying on solely hard data ignores the importance of creativity, innovativeness, and group dynamic that is crucial to startup success.

One key difficulty in modeling startup growth is excluding factors that are unknown during the investment stage in the training dataset. Information about a company's funding rounds, especially later funding stages, are typically unknown when deciding whether or not to invest. Additionally, using web sources with publication dates after the funding stage in question results in an inaccurate model for startup success. These biases in the training data result in the prediction model performing significantly worse during testing and real world prediction than during the training stage, which should be minimized as much as possible in this thesis.

The "wisdom of crowds" is another key breakthrough in predictive modeling, employing the knowledge of large amounts of non-expert humans to aggregate subjective and objective knowledge into accurate predictors of success [29]. Employing the wisdom of many reduces individuals' biases and complements the machine's accurate analysis of hard datasets with human evaluation of soft data in risky, uncertain situations. This thesis will look into utilizing internet references and social media mentions to employ the "wisdom of crowds" to measure company hype, complemented by more traditional machine learning on general organizational information to predict startup success.

A total of 9,842 companies were used in this research, resulting in 22,125 distinct datapoints that are used to represent a company's profile at a specific point in time. 6,574 of those datapoints represent companies that had raised funding within the allotted time window, approximately 29.7% of the total dataset. The initial group of 853 "successful" companies that raised funding was collected from Pitchbook, a software firm that delivers data, research, and technology covering private capital markets including venture capital, private equity and MA transactions. These companies received funding between January 1, 2020 and January 13, 2022 and were selected by a machine learning engineer at JP Morgan as good candidates for investment. The additional companies were randomly chosen from Crunchbase, another service containing financial and business data on hundreds of thousands of early-to-later-stage companies. The criteria for selection was a founding date more than 3 and less than 15 years ago, giving the startup enough time to gain traction and raise a round but not too long that the investment patterns have significantly changed.

The goal of this research is to predict the likelihood of a startup raising funding within a fixed time horizon, using available public data to model a company's projected growth. Precision vs recall is a big tradeoff in predicting startup success, with higher precision preferred due to the riskiness of investing in a false positive company [41].

Definition: Throughout this paper, a startup is defined as being "successful" if they raise a funding round within an alloted time frame.

2.3 Data Sources

In this section, we describe the data sources used to generate our feature set. This includes Crunchbase, the Programmable Google Search Engine, and Twitter.

2.3.1 Crunchbase

This paper uses data obtained from one of the largest companies containing business information – Crunchbase [4]. Their platform contains information on private and public companies ranging from early-stage startups to Fortune 500 companies. Their content includes investment and funding information, founding members and individuals in leadership positions, mergers and acquisitions, news, and industry trends. It is widely used by investors for gaining insights on potential investments and by businesses looking to scope out their competition. It is also heavily used in previous research in startup success prediction [25, 41, 50, 55].

Crunchbase has a free tier basic access plan that provides individuals with very limited data on the companies in their dataset. Initially, this list was used to generate a potential list of companies to be included in this study.

Crunchbase also has an academic research access program that provides researchers with free or discounted access to the entire Crunchbase dataset. After applying to the program, we received 6 month access for the research purposes of this thesis.

The dataset provided by Crunchbase contains several tables that holds information about companies, people, funding events, investors, acquistions, and more. The following tables were used in this study:

- Organizations: Contains general company information such as founding date, number of founders, website, and status (active, closed, acquired, or ipo). A full list of collected fields can be viewed in the appendix.
- Press References: Contains major news articles released about the company. These are oftentimes about new funding events, changes in management, new product releases, and location expansions.
- Funding Rounds: Contains amount of funding raised, funding stage (seed, grant, Series A), and investor information for all previous funding rounds.

The Crunchbase Search API was used to collect all the aforementioned information for each company in the training dataset which will be described in more detail in Section 4.3.1.

2.3.2 Pitchbook

Pitchbook is another software firm that delivers data, research, and technology covering private capital markets including venture capital, private equity and MA transactions [17]. While Crunchbase allows the general public to add and update company information, Pitchbook is fully maintained by a research team dedicated to finding and updating information on funding deals. A random sample of companies that had raised funding was collected from Pitchbook to be used in model training and evaluation.

2.3.3 Google Search Engine

Previous researchers found highly useful feature sets by observing a startup's presence on the internet. Sharchilev and his team were the first to use online search results returned from Yandex, a Russian search engine, to measure references to a company from other websites as a feature in startup success prediction [41]. More recently, Garkavenko et al. used the Programmable Google Search Engine to determine a company's presence on the web [30].

The Programmable Google Search Engine allows developers to include Google search engine results on their websites [5]. For this study, the Custom Search JSON API was used to retrieve web search results programatically for each company in the dataset. For each search, you can enter a keyword, in this case the company name, and a date range, which filters out entries with an invalid publication date. Each result returns the top 10 results for the given query and the total number of matches. This paper was limited to the Google Cloud Platform Free Tier, which restricted access to 100 queries a day. Consequently, we were limited from paging through the search results and only include the top 10 results along with the total count in the training and test datasets.

2.3.4 Twitter

Over the past decade, social media data has been increasingly used in machine learning models. Activity and sentiment of the general public has shown to be be highly useful in event detection and prediction. With Twitter containing massive amounts of public data, both from the general public and the company itself, the quantity and content of the tweets is useful in tracking the evolution of a startup. In particular, Twitter has been found to be a news source for 69% of its users and is commonly used to express opinions and thoughts on topics and trends, making it the ideal place to utilize the collective public mind [27].

Twitter selectively grants Academic Research access for non-commercial use to individuals with specific research objectives. Key benefits include access to every tweet published since launch in 2006 and a limit of 10 million tweets per month. We received Academic access in December 2021 and used query filters to extract relevant company data using the Twitter Search API. Each tweet returned contains the text itself, along with the creation date, author id, language, entity data: urls, hashtags, & mentions; and public metric data: number of likes, retweets, replies, and quotes (Table 2.1). For each returned tweet, structural and linguistic features, including word count, punctuation, sentiment, and complexity, were extracted using natural language processing methods. A full list of the features can be viewed in Appendix A and the methodology is described in detail in Section 4.3.3.

Table 2.1: Tweet Response

Tweet fields
Text
Date posted
Author id
Language
URLs, hashtags, & mentions
Retweets, likes, replies, & quotes

2.4 Classification Metrics

The goal of this paper is to identify whether a company at a given point in time will raise funding within a fixed number of years into the future. This is an important milestone in a company's life-cycle and very important for investors to determine because additional funding rounds mean a greater return on investment for early investors. Companies will therefore fall under one of two categories: (a) receives funding and (b) no funding. The various metrics to measure classification performance include accuracy, prediction, recall, and F1, which are described below.

2.4.1 Accuracy

A very common and simple metric is accuracy, measured by the number of labels correctly and incorrectly assigned to the data. However, in classification tasks where the data is unbalanced and favors one class over another, accuracy is unable to properly convey the modelś performance. For example, in fraud detection, a majority of transactions, let's say 99%, are valid while only 1% are fraudulent. However, a model that always returns valid will achieve 99% accuracy while failing to detect a single fraudulent transaction. In startup prediction, the data is also highly biased towards unsuccessful companies, as only 5-10% of startups ever raise a funding round. Consequently, accuracy will not be used for evaluating model performance as the model will be biased towards the negative class.

2.4.2 Precision

Precision measures how many of the positive labels were classified correctly. In other words, out of all the companies that were predicted to raise money, how many actually did. A high precision is important in startup success prediction to avoid incorrectly investing money into false positives.

2.4.3 Recall

Recall measures how many of the positive labels were correctly labeled as the positive class. This is important because if the model incorrectly predicts a company as not receiving funding when in fact it does, the investor looses out on an opportunity for a great investment in a potential unicorn.

2.4.4 F1

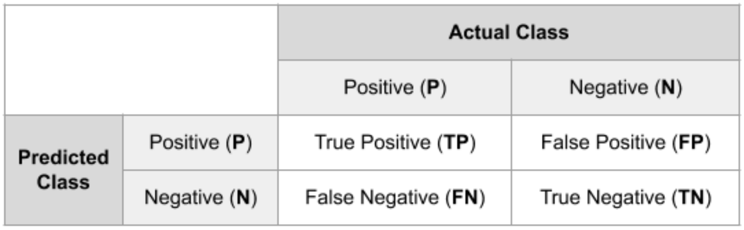
F1 scores combine precision and recall into one harmonic metric. It is possible to adjust the F-score to give more importance to precision over recall, or vice-versa. This paper will primarily use F1 scores to evaluate the performance of the prediction model. This ensures that bad companies aren't excessively misclassified as receiving funding and leading to a bad investment, and that good companies aren't skipped too often, resulting in a lost good investment.

$$Precision = \left(\frac{TP}{TP + FP}\right)$$

$$\text{Recall} = \left(\frac{TP}{TP + FN}\right)$$

$$\text{F1} = 2 * \left(\frac{precision * recall}{precision + recall}\right)$$

Figure 2.2: Precision, Recall, and F1 Formulas



		Actual Class	
		Positive (**P**)	Negative (**N**)
Predicted Class	Positive (**P**)	True Positive (**TP**)	False Positive (**FP**)
	Negative (**N**)	False Negative (**FN**)	True Negative (**TN**)

Figure 2.3: Confusion Matrix

The formulas used to calculate precision, recall, and F1 are shown in Figure 2.2.

2.4.5 Confusion Matrix

Classification tasks usually use the metrics of Precision, Recall, and F1 scores, which utilize the commonly used confusion matrix, see Figure 2.3. This matrix compares the predicted values to the actual values of the dataset. The possible labels for a classification problem are as follows:

- A True Positive (TP) is when the predicted value (positive) matches the actual value (positive). Ex: image of a cat is labeled as a cat.
- A False Positive (FP) is when the predicted value is positive but the actual value is negative. Ex: image of a dog is labeled as a cat.

- A True Negative (TN) is when the predicted value (negative) matches the actual value (negative). Ex: image of a dog is labeled as a dog.
- A False Negative (FN) is when the predicted value is negative but the actual value is positive. Ex: image of a cat is labeled as a dog.

For the topic classification model, the macro F1 score will be used to capture the performance on all seven topic categories. F1 scores can be evaluated on a micro or macro basis, where micro scores give equal importance to each datapoint, while macro scores give equal importance to each class. Macro scores perform better on imbalanced classes since they're not biased towards the dominant class, and will be used to evaluate the topic models.

Since this thesis focuses on predicting successful startups, precision, recall, and F1 scores will be primarily measured for the positive (raised funding) class. The overall (positive class + negative class) scores will not used for model and feature selection.

2.5 Google Cloud Services

Several Google Cloud services were used to automate the data collection and storage process for the startup tracker website.

2.5.1 Compute Engine

Compute Engine lets you create and run virtual machines on Google's infrastructure [6]. In this paper, Computer Engine was used to create a server to run daily jobs to collect and aggregate data for the startup tracker website.

2.5.2 Cloud Firestore

Cloud Firestore is a NoSQL cloud database hosted on Google Cloud [7]. It can be used to store data in a document format with complex nesting across document collections. It is used in this paper to store collected Twitter data for display on the startup tracker website.

2.5.3 Cron Jobs

"The cron command-line utility is a job scheduler on Unix-like operating systems" [2]. Cron expressions can be used to schedule jobs to run jobs at a fixed interval on a minute, hourly, daily, monthly, or yearly basis. Cron expressions follow the following format: <*minute hour day month year*>. The following cron schedule expression *0 0 * * ** can be used to run a specified job every day at midnight, 00:00.

2.6 Machine Learning Models

Ten ML classification models were implemented in this paper: the six traditional supervised learning approaches: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and four boosting algorithms: AdaBoost (SB), Gradient Boost (AB), XGBoost (XGB), and CatBoost (CB).

2.6.1 Boosting Algorithms

While many ML models focus on high quality prediction performed by a single model, boosting algorithms aim to improve performance by combining a sequence of weaker models together. In this ensemble learning approach, a random sample of data is selected, fitted with a model (ex. logistic regression or decision tree) and then trained sequentially [53]. Each model is built on top of the previous one and tries to compensate for the weaknesses of its predecessor by correctly classifying missed data points. The weaker rules from each individual classifier are combined to form one, strong prediction rule. It's important to note that boosting is not a model but rather an algorithm which uses a specified model for each one of its steps.

2.6.1.1 CatBoost

The CatBoost ensemble method, short for Categorical Boosting, supports numerical, categorical, and text features as input. It was developed in 2017 by machine learning researchers and engineers at Yandex, a Russian search engine company [23]. CatBoost uses the gradient boosting technique on decision trees to build its ensemble pipeline. Gradient boosting is a type of boosting algorithm that aims to create the next best model by combining with previous models to minimize the overall prediction error. Each model makes small changes in the prediction of data points aiming to reduce the gradient of the error. CatBoost uses one-hot encoding to handle categorical features, making it the ideal algorithm to use for our complex feature set.

2.7 Natural Language Processing

Natural language processing (NLP) refers to the branch of computer science, particularly in the field of artificial intelligence (AI), focused on teaching machines how to process and understand human text and speech and to respond in an intelligible and understandable way.

2.7.1 NLP Libraries

This paper uses four open-source NLP python libraries, NLTK, Spacy, PassivePy, and readability to process and extract linguistic features from the collected tweets.

2.7.1.1 Natural Language Toolkit (NLTK)

The Natural Language Toolkit (NLTK) is a python platform used in natural language processing (NLP) to work with language data [15]. It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It also includes over 50 corpora and textual sources such as the Penn Treebank for training and testing language models and NLP programs. The NLTK library was used to generate several linguistic features for the Twitter dataset. First, the punctuation package was used to find and count all punctuation characters in the text. Next, the CMU Pronouncing Dictionary (cmudict) was used to determine how many syllables were in each tweet. WordNet, a large lexical English database contains nouns, verbs, adjectives and adverbs was used to calculate syntactic complexity [22]. Each Wordnet part-of-speech category (noun, verb, adj, etc.) is made up of synsets which group together synonymous words that express the same concept. For example, eat and feed make up the same verb synset grouping. These synsets can each have a parent and multiple children, such as devour, consume, gobble, and nibble (children) and consume (parent) for the verb eat. Finally, the VADER SentimentIntensityAnalyzer was used to calculate tweet sentiment towards the company in question which is described in more detail in Section 2.7.2.1.

2.7.1.2 Spacy

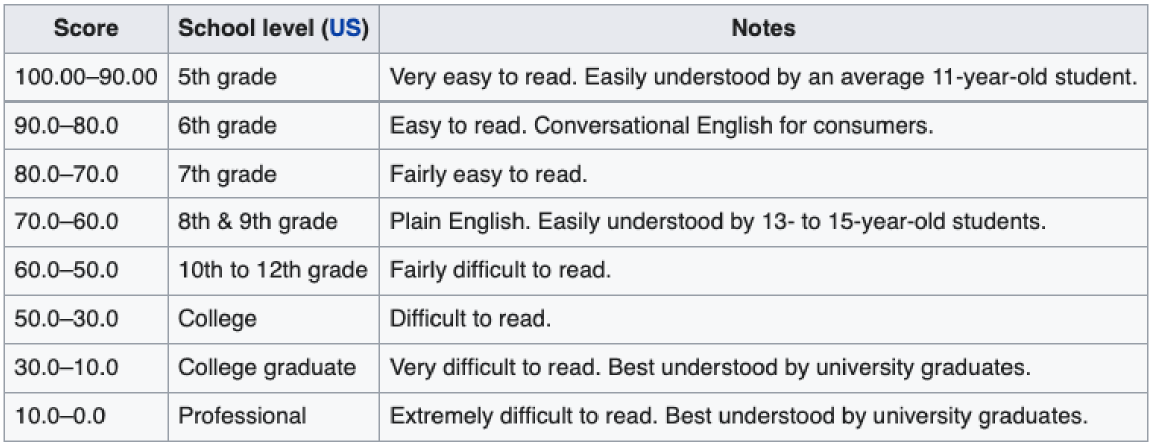
Spacy can be used for many NLP tasks such as learning what a text is about, what the words mean in context, how similar two texts are, and what named entities are mentioned such as companies and people [19]. In this paper, Spacy is used to extract syntactic meaning of tweets. This includes part-of-speech (POS) tagging, determining the textual grammar and labeling each word as a noun, verb, adjective... Additionally, it was used to calculate the shape of each word. For example, the sentence ”I love to eat.” would be represented as ”X xxxx xx xxx.”.

2.7.1.3 PassivePy

PassivePy is a tool developed to identify passive voice in large textual data [16]. Given an input text, it returns a distribution of how many sentences used a passive voice. For context, a passive voice is when the subject is acted on by the verb in the sentence, and is oftentimes used to emphasize the verb or when the subject is vague or unknown. For example, “The paper has been written.” is passive while “I wrote the paper” is not. It was hypothesized that a high occurrence of passive language symbolizes a dissociated between the subject and the topic at hand, and could correlate with a negative financial future.

2.7.1.4 Readability

A readability score measures how easy or difficult it is to understand a piece of text. Several factors go into a text’s readability, such as word choice, sentence length, sentence structure, and the average syllables per word. For example, choosing a word that’s unfamiliar or complex compared to a similar simpler word, such as “jocular” in



Score	School level (US)	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–10.0	College graduate	Very difficult to read. Best understood by university graduates.
10.0–0.0	Professional	Extremely difficult to read. Best understood by university graduates.

Figure 2.4: Flesch Reading-Ease Scores [24]

place of “happy”, can reduce the ability for readers to understand a piece of text. A lower readability score is important in social media, especially Twitter where tweets are limited to 280 characters, to make sure the message is clearly conveyed to readers. Consequently, we predict a high readability score will be correlated with less public engagement and a lower success rate.

One method to measure readability is the Flesch reading-ease test, where higher numbers indicate material is easier to read [24]. The formula for the Flesch reading-ease score (FRES) test is as follows:

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right)$$

These numeric scores are then interpreted as school level readability grade as shown in the Figure 2.4.

2.7.2 Sentiment Analysis

Sentiment Analysis is the process of computationally determining an author’s subjective attitude toward a particular topic. The goal is to determine whether or not their opinion is positive, negative, or neutral and how potent are their feelings are from a given piece of text.

2.7.2.1 VADER

Valence aware dictionary for sentiment reasoning (VADER) is the most common library used for sentiment analysis and is freely available through the NLTK library. It assigns a positive, negative, or neutral score to each word in the input text and outputs a composite score after aggregating the assigned labels. "I love cats!" will have a high positive score while "Ouch the cat scratched me and it hurts" output a negative score. Additionally, Vader is optimized for social media data which makes it a go-to method for analysing Twitter data. However, the rule-based approach has a major drawback in that it ignores the context in which the words were used. For example, "The startup destroyed its competitors" scores -0.49 on a -1 to 1 scale due to the token-based nature of VADER, while a human reader can tell that the outcome was positive for the startup in question.

2.7.2.2 BERT

Bidirectional Encoder Representation for Transformer (BERT) is an NLP model developed by Google Research in 2018, and has achieved state-of-the-art performance on several NLP tasks. A BERT model was trained on approximately 124 million tweets from January 2018 - December 2021 and is openly available on Hugging Face for public use [12]. This roBERTa-base model was fine-tuned for sentiment analysis on the TweetEval benchmark and is optimized for English language used in social media. It uses embeddings to model word meaning, achieving higher performance than rule based sentiment analysis at the cost of a much higher computational and time requirement.

Both of these sentiment analysis models are used on the collected tweets to model public attitudes towards a company.

Chapter 3

RELATED WORK

Considering the vast impact of startups and the huge potential for return on investment, there has been a long history of research on determining what factors are crucial for business success. Initially, researchers relied on questionnaire's, such as in Stuart and Abetti's 1987 paper, where the authors used founder surveys to uncover trends in successful startups [45]. One such finding was that a team's technical and market experience must align with the industry they're entering to be successful. Recently, this research has turned to machine learning models to predict successful ventures amongst the sea of new startups.

Traditionally, hard factual data from CrunchBase, such as the industry sector, company headquarters, and previous funding rounds, has been the principle form of data collected. Recently, research has been turning to capturing and interpreting unquantifiable "soft" features such as founder creativity and innovation and "making predictions in unknowable risk situations of extreme uncertainty" [29]. Employing a wisdom of crowds approach ensures that individual human biases and errors are eliminated through the aggregation of human knowledge. Combining hard and soft features has led to continuous improvements in the startup prediction space.

3.1 Feature Extraction

The work by Xiang and his team is one of the first attempts to use machine learning models to predict startup success [50]. Specifically, they were using Merger & Acquisition as their metric to categorize a company as successful or not. In their work, the authors were the first to use data from CrunchBase, the largest public business database, to build their company profiles. On top of extracting factual company features such as company age and headquarters, they also collected team profiles and insights on company news articles. One of their key contribution was using topic modeling on articles from TechCrunch, the most popular tech news website, to uncover what new events the company was going through. The articles' headlines were used to group the articles into one of five categories under the assumption that words commonly related to acquisition and funding could be learned and used to model positive growth.

The next key milestone in startup success prediction was using online web data to capture a company's growth [41]. This study focused on predicting startup funding of Series A or higher within the upcoming year, with an angel or seed round as a required trigger for prediction. They built on Xiang's work [50] and crawled data from LinkedIn and the web to complement CrunchBase data. Their data was grouped into four categories: 1. general company data such as age, industry, office locations, and number of products. 2. investment data such as founding rounds and investor profiles. 3. founder data such as the number of team members and LinkedIn profile data capturing their previous experiences and attempted/successful startup ventures. 4. web mentions in news articles and links on other websites. To gather these web mentions, a web scraping study counted the number of times a company's website was directly linked on other webpages, and found that this web based startup success prediction (WWSSP) outperforms all other prediction models.

A recent study found that just looking at geographical, demographic, and general company information results in precision, recall, and F1 scores of 57%, 34%, and 43% respectively [25]. This study had the largest training set to-date, consisting of 213,171 companies. They found that location and industry of a company are some of the key indicators of startup success. Additionally, the exclusion of funding data removes the biases found in previous research, which trained data on funding information unavailable during the investment stage in real world scenarios. However, relying heavily on black-box founder demographics could result in one group being favored over another, raising red flags on the ethics of these models [38]. Similarly, focusing on startup location could result in a class imbalance since startups are typically centered in technological regions. The high volume of startups coming from areas such as Silicon Valley could result in only Silicon Valley startups being deemed worthy of investment, ignoring other regions simply because they haven't generated any startups before. Therefore, additional features are necessary to reduce the inherent bias of relying on demographics, geographical, or obtained funding to predict startup success.

The latest work on startup success prediction focused on using free, publicly available web information as their data source [30]. While previous work focused on using structured databases for their machine learning models, building and maintaining these databases requires a tremendous amount of human effort. Garkavenko and her team explored whether freely available data such as the website of a startup, its social media activity, and its web presence can be used to predict funding events within a given time horizon. They started off by gathering 22k startups from hubs, investors, and conferences around the world, primarily focusing on European startups. Then, they used the startup's own website to extract general information such as the country of origin, age, number of employees, and number of offices. Next, they gathered social network data such as the amount of social media accounts the startup had, collected though links on the startup's website, and their activity on Twitter including number of tweets, likes, and received mentions by other users. Then, they summarized the financial history of the startup containing previous funding round, last fundraising amount, and time since last secured round. This data was extracted from tweets and news articles using regular expressions. They had a false positive rate of 8.5% and a false negative rate of around %6 for a small sampled data set of 200 startups. Finally, they used the free Google Search API to retrieve the top 10 results for each startup and gathered the number of relevant results that mentioned the company name, the number of total results, and the number of results from each of the 500 popular domains. They tested model performance using the most widely used machine learning models such as Logistic Regression, Random Forest, and a gradient boosting algorithm supporting categorical variables called CatBoost.

3.2 Stock Market Prediction

Closely related to startup success is stock market prediction. While startup success prediction focuses on early-stage private companies and stock market prediction is centered on later-stage public companies, both aim to predict the financial future of a company to determine whether or not a company is a worthwhile investment.

There are two main types of prediction targets in this field of research: (a) stock market movement, where the goal is to predict whether the stock will go up or down in a specified time window; and (b) price prediction. While the motivation behind the first is to determine whether or not the stocks should be bought or sold, predicting price provides investors with more information and enables them to make informed decisions about the associated risk.

In their paper, Weng et al. aim to predict short-term stock prices using ensemble methods and online data sources [49]. They used historical stock data, well-known technical indicators, value counts and sentiment scores of published news articles, trends in Google searches, and unique Wikipedia visitors for relevant Wiki pages for their feature set. Once the data was collected, they created four machine learning ensemble methods: (a) a neural network regression ensemble; (b) a support vector regression ensemble; (c) a boosted regression tree; and (d) a random forest regression. They collected data for 20 different stock indicators (Amazon, Chevron, Coca-Cola, IBM, McDonald's, Pfizer...) across a range of industries, volatility's, growth patterns, and general conditions to test their ML model. Additionally, they used PCA to limit the feature set to save training time. They were able to achieve a test mean percent error (MAPE) $\leq 0.75\%$ for all but one stock indicator using their best model, the boosted regression tree with no PCA. They found that the MAPE was lower for instances with no PCA and that the support vector regression ensemble had the lowest performance of the four.

Additional research in this field has looked into using social media sources to predict stock movement and prices. Interactions on social media, particularly Twitter, have been successful in predicting several other categories of events, such as sports outcomes and political elections, and this has been used for both startup success and stock prediction as described in the next section.

3.3 Twitter for Prediction

Interactions on social media have been found to reveal remarkably accurate predictions about future events. Twitter is one of the largest social networks globally with 206 million daily active users and approximately 200 billion tweets shared each year. It is frequently used as way to share public opinions and consume news, making it the ideal place to capture user interest and sentiment about a topic.

Previous researchers have used Twitter to (a) pinpoint victims and their locations during natural disasters [43], (b) predict the outcome of English Premier League football matches [34], and (c) predict election results and political candidates' likelihood of election [44, 35].

3.3.1 Election Results

A study conducted on 98 election prediction papers found that Twitter prediction approaches fell into one of three main categories: (a) volumetric data, such as the number of users, tweets, mentions, likes and favorites, (b) sentiment analysis, such as the number of positive/negative tweets about a party or candidate, and (c) social network analysis, using community detection and node (person) importance using graph analysis [35]. Probabilistic classifiers (naive bayes), decision trees, linear classifiers (support vector machines, neural networks, and KNNs), and rule-based classifiers make up the supervised learning approaches, while k-means clustering was used for unsupervised learning and recurrent neural networks and CNNs were used in deep learning approaches. An overwhelming majority, 89% of the papers, used a sentiment analysis approach, since measuring the attitudes of users has proven to be one of the best ways to predict electoral results. A detailed image of the common approaches can be viewed in Figure 3.1.

3.3.2 Stock Market Movement

One of the first works on stock market prediction was done by Zhang and his team in 2010 [35]. Over the course of six months, from March 30, 2009 to Sept 7, 2009, they collected between 8100 to 43040 tweets each day, approximately 1% of the total daily volume. They created several categories of words, such as happy, worried, fear, hope, and identified how many words in each category were found for a given time range. Those word counts, representing the collective mood of the user base, was found to correlate to the movement of the Dow, NASDAQ, and SP 500.

Future work built on this research with advanced sentiment analysis models, with Valence Aware Dictionary and sEntiment Reasoner (VADER) the most popular model.

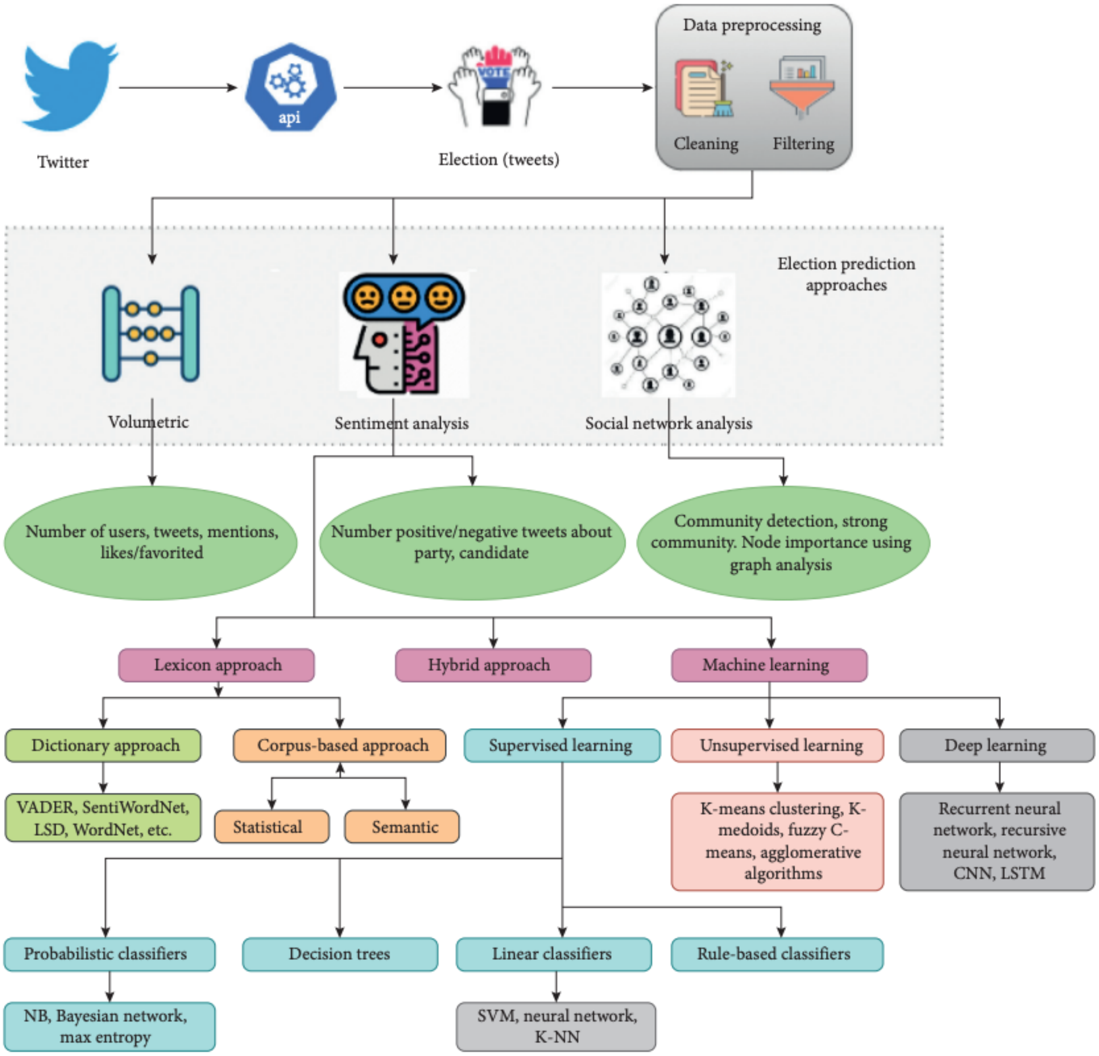

Twitter
api
Election (tweets)
Data preprocessing
Cleaning
Filtering
Election prediction approaches
Volumetric
Sentiment analysis
Social network analysis
Number of users, tweets, mentions, likes/favorited
Number positive/negative tweets about party, candidate
Community detection, strong community. Node importance using graph analysis
Lexicon approach
Hybrid approach
Machine learning
Dictionary approach
Corpus-based approach
Supervised learning
Unsupervised learning
Deep learning
VADER, SentiWordNet, LSD, WordNet, etc.
Statistical
Semantic
K-means clustering, K-medoids, fuzzy C-means, agglomerative algorithms
Recurrent neural network, recursive neural network, CNN, LSTM
Probabilistic classifiers
Decision trees
Linear classifiers
Rule-based classifiers
NB, Bayesian network, max entropy
SVM, neural network, K-NN

Figure 3.1: Overview of election prediction approaches and techniques [35]

Koukaras et. al. worked on predicting whether the Microsoft stock would close higher or lower than the previous day [36]. They tested seven different models: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF) and Multilayer Perceptron (MLP), and achieved an F-score of 76.3% using a SVM model. However, their work was limited to just one company, and they suggested that future researchers explore whether using only verified users to eliminate potential noise from bots, would improve results.

To improve the validity of the collected tweets, it's important to recognize whether or not the information comes from a credible and trustworthy source. Huang and his team investigated four different aspects of a social media user to score his/her trustworthiness to help weigh the user's overall impact on the sentiment score [32]. (a) Expertise, measures a user's involvement in the subject of interest, (b) experience is the difference between a user's expertise and the average expertise in the network, (c) authority is the number and quality of social media links (retweets and quotes) a user receives from recognized hubs [8] and (d) reputation is the number and quality of social media links to the particular user. All these metrics combined can help reduce the influence of scams, bots, and low credibility users in prediction algorithms.

3.3.3 Startup Success

In their paper, Antretter and his team were the first in the entrepreneurship field to show the potential of using data mining, natural language processing, and machine learning to capture online information to predict entrepreneurial outcomes [26]. They relied solely on Twitter data to predict startup survival, whether or not a startup would be alive in another 5 years. They collected Twitter activity, measured through likes, followers, and the sentiment of user comments, for 253 seed or early-stage companies founded between 2006 and 2018, of which 72% survived for at least five years. Using the Twitter REST API, they collected a total of 187,323 tweets, 102,501 retweets, and 441,583 likes to train the model. Survival was characterized by an active company website five years after the founding date.

They used Linguistic Inquiry Word Count (LIWC) to capture positive and negative emotions in seven different languages to associate words most strongly associated with survival or death. Some other key features include the number of tweets, number of likes, retweets, and followers, and the tweet length. The length of tweets, # of likes given, # of likes received, ratio of followers over following, # of followers, negative content retweets, and user engagement revealed to be the most important factors correlated with new venture survival. They achieved an recall of 86% and a precision of 80% in correctly classifying startup survival, indicating that startups classified as surviving have a probability of 83% of actually surviving. While this work shows the importance of Twitter data in startup prediction, survival is rarely used as a success metric in practice. This paper will build upon their findings to show social media's impact on predicting the more popular measurement of startup success: funding events.

In a more recent study from 2021, Tumasjan and his team continued the work in exploring Twitter sentiment's correlation to startup success [47]. Using a sample of over 4,600 VC financing rounds, they collected over 400,000 tweets published a few weeks before these events occurred. The lexicon and rule-based method VADER (Valence Aware Dictionary and sEntiment Reasoner) which was trained on social media data was used to calculate sentiment.

The researchers found that while Twitter sentiment was able to predict valuations, it failed to show a significant relationship to investment success, classified in this paper as an IPO or acquisition of the firm. They found that investors across industries are paying more and more attention to Twitter sentiment, but that "hype" generated by the masses is often due to over-optimism that isn't the best indicator of long term success. On the other hand, they found that patents are a strong indicator for IPO or M&A, and that the effect of Twitter sentiment is stronger if a startup has applied for patents.

While this research proves that Twitter Sentiment is not the best indicator of IPO and M&A than patents, VADER sentiment alone is not enough to rule out social media's signal of startup success. For starters, on a VADER scale from -1 (negative) to 1 (positive), the headlines "Guardian Capital Group buys majority stake in Alta Capital Management" and "VeoRide Announces E-Scooter Durability and Sustainability Breakthrough" both achieve a compound score of 0.0, symbolizing a 100% neutral sentiment. In contrast, a human reader can deduce that both of those headlines are positive events. Additionally, it takes an average of 8-10 years for a company to IPO and 6-10 years for a M&A deal to be reached. The paper doesn't mention if all the companies in the dataset had been around for at least ten years. Finally, it was found that while Twitter Sentiment alone was not enough to accurately signal startup success, it had a much stronger effect when paired with patent data. Consequently, this paper will focus on improving the prediction accuracy by pairing general company data and web activity with a large number of Twitter features such as general company Twitter statistics (likes, followers, retweets), topic modeling, and deep linguistic features.

Chapter 4

IMPLEMENTATION

This section begins by introducing the company selection process used for generating the data set present in this thesis. Next, the data collection pipeline will be described in detail, explaining how Crunchbase, Google Search and Twitter data was collected for 9,842 companies. Then, the natural language processing pipeline will be introduced, explaining how linguistic features were extracted for textual data. Next, the Startup Tracker website will be described in detail, going over how this tool can be used to improve the startup prediction process for VCs and investors. Finally, all the components will be combined into a real-time prediction pipeline, taking in a Crunchbase ID, start date, and time window and outputting the startup's likelihood of success. An overview of the system design can be viewed in Figure 4.1.

4.1 Company Selection

9,842 companies were selected for this study, with 9,380 used for predicting whether or not a company will raise any funding round, and 1,438 used to predict whether or not a company will raise an additional round of funding given they had previously raised an angel or seed round. 6,574 (29.7%) of the datapoints make up the positive class, meaning they raised a funding round in the allotted time window.

The initial collection of 853 successful companies was selected by Dr. Moghaddam, a machine learning engineer at JP Morgan, as prime examples of good investments. They were chosen from a dataset of companies that had raised funding between January 1, 2020 and January 13, 2022 on Pitchbook, an software platform similar to

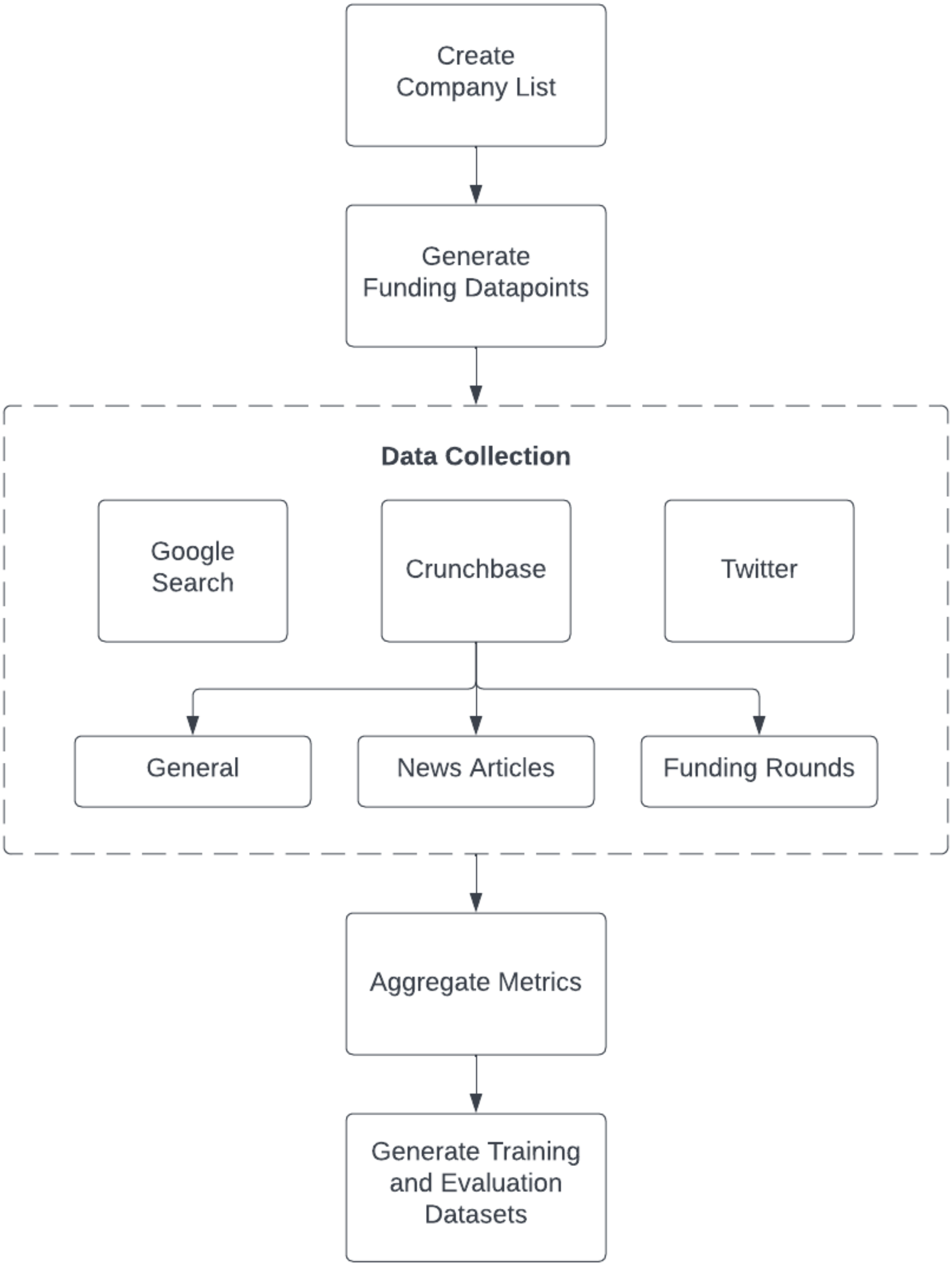


Figure 4.1: System Design

Crunchbase containing information on companies and investments. This is similar to the initial search process that VCs and investors go through when looking for high potential companies to invest in. An additional 7,551 companies were randomly selected from the Crunchbase daily csv export [3]. Crunchbase has a list of approximately 1 million companies they've made publicly available through their Crunchbase Basic access account. This dataset has very minimal data about a company, including their name, description, headquarters, and Crunchbase url. Companies were randomly selected from the list that had existed for more than 2 years but less than 15. This ensures the selected companies had enough time pass that they could have raised a funding round, but not too much time that the investment practices have significantly changed since when the company was founded. Additionally, since Twitter was founded in March of 2006, this ensures Twitter was around when all the startups were created. Finally, randomly choosing companies helps ensure they are from companies a diverse set of industries, locations, and years since founding. Out of the collected companies, 695 had raised funding and 6,856 had not, the 9.2% success rate typical of the industry standard of 5-10%. The final 1,438 companies were selected with the requirement of raising an angel or seed round using Crunchbase Search API. This allows us to compare our final model to previous research in startup success prediction.

Endpoint: https://api.crunchbase.com/api/v4/searches/funding_rounds

Query Parameters:

1. announced_on $\geq$ 2007
2. announced_on $\leq$ 2020
3. investment_type includes (angel, seed)

Table 4.1: Datapoint Distribution

Dataset	Type	Datapoints	Companies	Received Funding
Any Funding	Train	17,233	8,081	5,392 - 31.3%
	Test	3,004	1,541	660 - 22.0%
	All	20,237	9,380	6,052 - 29.9%
Additional Funding	Train	1,369	1,076	408 - 29.8%
	Test	520	362	114 - 21.9%
	All	1,888	1,438	522 - 27.6%
All	—	22,125	9,842	6,574 - 29.7%

The complete breakdown of the datasets and companies used in this thesis can be viewed in Table 4.1.

4.2 Dataset Generation

Many of the companies have been around for around a decade, with the average age being 8.5 years. The oldest company in the training dataset has been around for 15 years and the youngest for 1.5 years.

Since many of the companies had been around for about a decade, they have gone through many different stages of the startup growth cycle. Additionally, out of the 6,574 companies that raised funding, 40.2% raised more than one round, with the average being 1.61 rounds raised and Outreach raising a maximum of 10 different rounds. This means that for any given company, multiple datapoints can be generated that provide a unique insight into the company at that point in time. Additionally, this means that instead of finding new companies to train on, the same companies can be used at different stages of development.

For every company, a list was generated containing all the years since founding up until the current year. Then, four years were randomly selected and January 1 was used to generate a snapshot of the company at that point in time. For example, a company founded on March 16, 2012 might have 1/1/2013, 1/1/2015, 1/1/2020, 1/1/2022 as distinct entries in the generated dataset. These dates were then used to predict whether or not the company raised funding within a fixed time horizon. If the date in question was 1/1/2015 and the year range was 3 years, then the goal is to predict whether or not the company will raise funding between 1/1/2015 and 12/31/2017. A total of 17,233 training and 3,004 evaluation datapoints were generated by selecting several prediction dates for every company.

All collected data for that dataset entry discards any information unknown at that time. News articles, Google Search results, and funding events after that date are filtered out during data collection. Tweets are collected for the 9 months preceding the provided date. More details on data collection are described in Section 4.3.

4.3 Data Collection

In this section, we describe the methods used to collect data from Crunchbase, the Google Search Engine, and Twitter for generating our feature set. An overview of the data sources used can be viewed in Table 4.2.

4.3.1 Crunchbase

After creating the company list, detailed information about each company was collected using the Crunchbase API. In this study, all the available data from the organizations, press references, and funding rounds tables was fetched for every company.

Table 4.2: Data Sources

Data Source	Feature Categories	Description
Crunchbase	General, Funding Events, News Articles	Startup information including company founding date, description, industries, headquarters, founders, funding rounds, etc. https://api.crunchbase.com/api/v4/entities/organizations/
Google Search Engine	Google Search	Captures internet presence by returning total search results and top 10 hits. https://www.googleapis.com/customsearch/
Twitter	Twitter	Twitter captures public awareness and attitude towards the company on social media. https://api.twitter.com/2/tweets/search/all

Crunchbase Endpoint: https://api.crunchbase.com/api/v4/entities/organizations/ORG_ID?field_ids=layout_id,is_locked&layout_mode=view_v2&user_key=API_KEY

The organizations table includes general information about companies such as name, headquarter address, number of employees, website, and social media links. Crunchbase also keeps track of the status of the organization – active, closed, acquired, or ipo (public company). Each organization is also described by the industry categories and subcategories that it operates in.

The press references table includes all the news articles collected for a company. Each entry includes the title of the article, the date it was posted, its publisher, and the publication date.

Detailed information about all the funding events for a company is included in the funding_rounds table. This includes information on the announcement date, amount raised, and investment type (seed, angel funding, series A, B, C, etc.). The name of every investor and their type (venture_capital, early_stage_venture, seed, angel_group) was also collected.

4.3.2 Google Search Engine

Google Search results were collected for each dataset entry consisting of the company and a point in time from which the prediction is made. The free tier Google Search API was used to fetch the top 10 results for every company. For every search, the company name was used as the keyword, and a date range was used to limit look-ahead bias by filtering out search results unavailable at the time of prediction. The start date was set as one year before the company was founded to limit irrelevant results published years before the company was born. The complete query url can be viewed below:

Google Search Endpoint: https://www.googleapis.com/customsearch/v1?key=
API_KEY&cx=SEARCH_ENGINE_KEY&q=COMPANY_NAME&sort=date:r:
START_DATE:END_DATE

Each response returned the total number of results found by Google, and a link, title, and snippet for each one of the top 10 results. It's important to note that the total number of results contains a majority of irrelevant matches that Google returns, similar to the actual search engine behavior.

4.3.3 Twitter

Twitter's Academic Research plan includes access to every tweet published since launch in 2006 with a limit of 10 million tweets per month. We received Academic access in December 2021 and used query filters to extract relevant company data using the Twitter Search API.

Twitter search queries can be constructed to retrieve only relevant tweets. This thesis uses the following query structure:

Table 4.3: Constructing Twitter Query

Tweet filters	Syntax
contains company website	website_url
posted by company	from:*company_username*
repost of company tweet	url:*company_username*
company mentions	@*company_username*
excludes retweets	-is:retweet

– (website_url OR from:*company_username* OR url:*company_username* OR @*company_username*) -is:retweet

All tweets that either contain the company website, are posted by the company Twitter account, are re-posts of a company tweet, or directly mention the company's twitter account are retrieved. [Table 4.1] Retweets are skipped to limit repeat data as the number of retweets is already included in the tweet metadata.

Initially, tweets were queried that also contained the company name, but those were eventually discarded due to the inability to distinguish false matches in the returned results. For example, the name Apple could refer to the company or the fruit, resulting in a large volume of false positives. Tweets that either contain the company website or directly mention the company or reference a company tweet are guaranteed to be postive matches.

Every tweet response contains the text itself, along with the creation date, author id, language, entity data: urls, hashtags, & mentions; and public metric data: number of likes, retweets, replies, and quotes (Table 2.1)

Twitter data was fetched for every company and prediction date pair. The Twitter API was called with the constructed query and a date range. The end date used was the provided prediction date (ex. 1/1/2015), and the start date was set to 9 months before the end date. A maximum of 500 results were returned with each API call, and subsequent calls were made to paginate through additional tweets for a maximum of 5,000 tweets per datapoint. These parameters ensure the API limit isn't exceeded too quickly, the returned results are relevant to the startup's current stage, and that enough results are found to model the growth of the company over time.

Twitter Endpoint: https://api.twitter.com/2/tweets/search/all

A total of 4,355,869 tweets were retrieved for 22,125 distinct training and evaluation datapoints. 13,951 (63.1%) of those datapoints had a company twitter account and 8,640 (39.1%) had at least one tweet result returned for the provided 9 month period. This resulted in an average of 201 tweets per company, but the mean was highly influence by the highly active accounts with over 1,000 collected tweets, 4.0% of the total datapoint set. A default value of 0 was used for all tweet features where no tweets were collected within the time horizon.

Table 4.4: Startup Feature Overview

Category	Name	Description	Type
General	Description	Company description	Text
	Company age	Months since founding	Num.
	Founder count	Number of founders	Num.
	Industries	Distinct industry categories listed on Crunchbase	Num.
	Name length	How long is the company's name	Num.
	Website length	How long is the company's website url	Num.
	Social media	Twitter, Facebook, and LinkedIn accounts	Num.
	Country	Location of company headquarters	Cat.
	State	Location of company headquarters	Cat.
	Headquarter hub	Four fields for CA, NY, TX, or other hub HQ	Cat.
Funding	Num previous rounds	Total number of previous fundraising rounds	Num.
	Last fundraising amount	Money raised (USD) during the last round	Num.
	Months since fundraising	Months since last fundraising round	Num.
	Last fundraising stage	Angel, seed, series A, etc.	Cat.
	Num total investments	Total number of investments received	Num.
	Num distinct investors	Number of distinct investors across all rounds	Num.
	Distinct/total investors	0-1 scale for capturing repeat investors	Num.
News	Num articles	Total number of Crunchbase articles	Num.
	Top publisher count	# of articles written by a top 10/50 publisher	Num.
Google	Total results	Total Google Search results for the company name	Num.
	Own company results	Number of times the company's website, LinkedIn, Twitter, or Facebook appeared in top 10 results	Num.
	Top publisher count	# of results by a top 10/50 publisher website	Num.
	Top google count	# of results by a top 10/50 Google website	Num.
Twitter	Tweet count	Total number of tweets	Num.
	Company count	Total number of tweets by company	Num.
	Unique users	Total number of unique users tweeting	Num.
	Engagement metrics	Likes, retweets, replies, quotes	Num.
	Sentiment scores	VADER and Bert sentiment	Num.
	Tweet contents	% tweets containing company website, account, replies, mentions, hashtags, links, and emojis	Num.
	Linguistic Features	Tweet structure: characters, words, sentences, shape; and deep linguistic: passive voice, parts-of-speech, syllables, reading score, complex words, etc.	Num.

4.4 Feature Extraction

In this section, we describe the feature extraction process used to generate data for prediction. The features can be classified into five broad categories according to the information sources that they capture: general, news, funding, web search, and twitter. See Table 4.4 for an overview of the features used in this thesis. The complete feature list can be viewed in Appendix B.

4.4.1 General Company Data

General company information such as the company name, description, months since founding, number of founders, and website url were collected directly from the Crunchbase Search API.

Additionally, the category groups of the company were collected and a sparse matrix was generated with all 47 possible industry categories, ranging from Software to Health Care to Media and Entertainment. Each company listed between 0-10 industry categories on their Crunchbase profile. To reduce the feature list and increase matches between companies, and the top six categories were chosen from each industry and sorted by how popular that category was across the company list. The complete category list can be viewed in Appendix E.

4.4.2 News Articles

The Crunchbase URL has a press_references table that contains information on recent news articles for a company. For each company in the dataset, press references were fetched in batches of 100, with a maximum of 2000 retrieved per company. For each press release, the title, publisher, and post date were extracted and saved. A total of 38,660 press references were retrieved for the 9,842 companies in our dataset. An additional 20,725 press references were fetched for 12,832 companies outside of the dataset using companies listed in the Crunchbase csv export [3]. This collection of press references was used to determine the top 10 and top 50 most popular publishers for news articles on Crunchbase. The top five publishers on Crunchbase were TechCrunch (1,714 occurrences), PR Newswire (849), Business Wire (753), PRNewswire (696), and PRWeb (528). It is inferred that a publisher who commonly posts about startups across industries and regions is a credible source on new ventures, and it is a high accomplishment to be featured on their site. The amount and percentage of press references written by a popular publisher is used to calculate the top_10 and top_50 publisher count features.

4.4.3 Funding Events

The Crunchbase URL also contains a funding_rounds table with information on a company's fundraising history. For each company in the dataset, all previous funding rounds were collected, and the number of previous rounds, previous fundraising amount, months since last fundraising, the last funding stage, number of total investment and distinct investors, and the ration between distinct and total investors were calculated as the funding event feature set.

4.4.4 Google Search Results

Given a company and prediction date pairing, the top 10 Google Search results and the total result count are returned as raw data. The root urls of the top 10 result links are determined and used to count how many times the company website or social media links (Twitter, Facebook, and LinkedIn) appears in the results. It was hypothesised that a high number of company link matches could positively correlate with funding as the company scores high in SEO. Additionally, the top 10 and top 50 websites appearing in Google Search results are calculated to be top referrers. Similar to press references, it is hypothesised that a mention by a popular startup website means the company is on track for success. Additionally, the number of times the top 10 and 50 news article websites computed in Section 4.4.2 appeared in the search results were as used as features.

4.4.5 Twitter

A total of 4,355,869 tweets were collected that were posted during a 14 year period between April 2007 and December 2021. These tweets were either posted by the company itself or contain the company website, company username, or reference a company tweet (Section 4.3.3). 68 features were calculated for every tweet and included in the final prediction model. A full list of tweets can be viewed in the Appendix A and are described in more detail below.

4.4.5.1 Tweet Engagement

Public metrics are returned for every tweet and contain the (a) number of likes, (b) number of retweets, (c) number of replies, and (d) number of quotes per tweet. These metrics were aggregated for the nine month period and the average count per tweet and total count over the time period are included in the prediction dataset.

4.4.5.2 Sentiment Analysis

Sentiment Analysis is the process of computationally determining whether an author's attitude in a given text is positive, negative, or neutral towards the topic at hand.

The Natural Language Toolkit's (NLTK) VADER sentiment analysis library was used to calculate the tokenized sentiment for each tweet. The raw tweet text was given as input and the positive, negative, neutral, and compound scores were returned per tweet. The final compound score was used as the VADER sentiment feature. Additionally, Google's BERT language model previously trained on 124 million tweets was used to calculate a semantic-based sentiment score for the returned tweets. However, since BERT uses word embeddings to model word meaning, the computation cost (200 tweets per minute) was too high to run on each one of the 732,744 tweets. Therefore, bert sentiment was calculated for the most recent 100 returned tweets from a maximum set of 5000. The average sentiment score for either all (VADER) or first 100 (BERT) tweets was calculated and used in the feature set.

4.4.5.3 Tweet Distribution

Every tweet was either posted by the company, mentioned the company, contained the company's website, or was a reply or reference to a company tweet. This distribution is captured by post processing the tweets and determining: (a) the number of tweets posted by the company / total tweets, (b) the number of tweets containing the company's username / total tweets, (c) the number of tweets containing the company's website / total tweets, (d) the number of tweets replying to a company tweet / total tweets, (e) the number of tweets replying to a company tweet / total tweets, and finally (f) the total number of tweets in the given nine month period, capped at 5,000.

4.4.5.4 Tweet Contents

Regex and string matching was used to determine the number of hashtags, mentions, links, and emojis in each tweets.

1. Hashtags: All words starting with a *‘#‘*.
2. Mentions: All words starting with a *‘@’*.
3. Links: All words starting with *‘http’* or *‘www.’*.
4. Emojis: The emoji python package was used to find all emoji unicodes in the tweet.

All of the above four metrics were calculated for each tweet, and the totals were aggregated to determine the average number of occurrences per tweet (per metric), and a ratio of how many tweets contained the metric / total tweets. For example, if 100 tweets out of 1,000 contained a url link, and 130 total links were found, the avg_has_links score would be 0.1 and the avg_num_links would be 0.13.

Each tweet included metadata on the language it was written in. The number of distinct languages found across collected tweets was used as a tweet feature. Additionally, the topic detection model described in Section 4.5 was used to determine if the tweet was talking about a new funding event, merger and acquisition, geographical expansion, product launch, award received, or management change. The number of times each of those topics was mentioned, along with the distinct number of topics found, were used as features to capture news and hype about major company events.

4.4.5.5 Tweet Structure

Basic natural language processing was performed to calculate metrics on the structure of the tweets. First, the tweet was tokenized to determine the number of words and sentences in the text. The length of the text was used as the character count, and number of punctuation characters was used as the punctuation count. Additionally, Spacy, an open-source software library for advanced natural language processing, was used to calculate the shapes of the tweets. For example, the text "I went to work today with my friend Dennis Smith..." will be encoded as "X xxxx xx xxxx xxxxx xxxx xx xxxxxx Xxxxx Xxxxx...". This shape structure was used to calculate the number of unique word shapes, total word shapes, and the ratio of unique/total word shapes. Additional deep linguistic features that were calculated with Spacy will be described in Section 4.4.5.6.

4.4.5.6 Deep Linguistic

Deep linguistic features were calculated using advanced natural language processing to uncover semantic and syntactic patterns in the text. First, the Spacy library was used to calculate grammatical features for the text. The text was tokenized and tagged with specific and generalize parts-of-speech (POS) labels. For example, the word "cat" would be tagged with a "NOUN" POS label and a "NN" (Common Noun) tag. The proper noun "United States" would be tagged with a "PROPN" POS label and a "NNP" (Proper Noun) tag. The number of total and distinct POS and tag labels was counted and used as a feature, along with the ratio of distinct/total POS and tag labels.

Additionally, PassivePy was used to calculate the amount of times a passive voice was used in collected tweets. The tweet was given as input and the number of sentences containing a passive voice was returned as output. This number was averaged across all tweets for the avg_passive_count score.

4.4.5.7 Complexity

The complexity of the language used in the tweets is the final linguistic feature category. It was hypothesised that a high complexity could correlate to lower success as social media users typically communicate using colloquial language and could be dissuaded by complicated messages. First, the Flesch Reading-Ease score was calculated using the readability library to determine what reading grade-level, from 5th grade to college graduate, the tweet is categorized as. This was the aggregated for the average tweet readability score feature. The readability library also returns the number of (a) long words, characterized by being longer than 7 letters, and (b) complex words, defined as a polysyllabic word make up of more than three syllables. These are averaged across the returned tweets to make up the avg_long_words and avg_complex_words features respectively.

WordNet, a large English database contains nouns, verbs, adjectives and adverbs was used to calculate syntactic complexity. Each part-of-speech category (noun, verb, adj, etc.) is made up of synsets which group together synonymous words that express the same concept. The average number of synsets per tweet and the ratio of unique over total synsets were calculated for the feature set. Additionally, the distance of each synset from the root was calculated to determine the synset

complexity. For the word “puppy”, the synset complexity is 15:
puppy → *dog* → *canine* → *carnivore* → *placental* → *mammal* → *vertebrate* → *chordate* → *animal* → *organism* → *living_thing* → *whole* → *object* → *physical_entity* → *entity*.

The average synset complexity per word and tweet are the final metrics used in the Twitter feature list.

4.5 Topic Modeling

A topic classification model was built to determine whether or not a news article, google search result, or tweet was about one of six topics. This target topic list consists of (a) funding events, (b) merger and acquisitions (c) geographical expansions, (d) new product launches, (e) awards received, and (f) management changes. These six topics were chosen as key milestones typical for companies experiencing growth and expansion, typical of successful companies.

To train the classification model, a dataset of 3196 news article headlines from Crunchbase was collected and manually labeled as one of the above six topics, or other. The topic distribution is shown in Figure 4.2.

The input text was then put through a preprocessing pipeline. This involved (a) converting the text to lowercase, (b) removing all non-alphanumberic characters, (c) removing all stopwords, (d) lemmatizing and stemming the words, and (e) replacing all instances of a dollar amount with MILLION, BILLION, or AMOUNT. The dollar amounts were replaced with the aforementioned labels to reduce the number of potential token inputs and increase the number of matches in the word based prediction model.

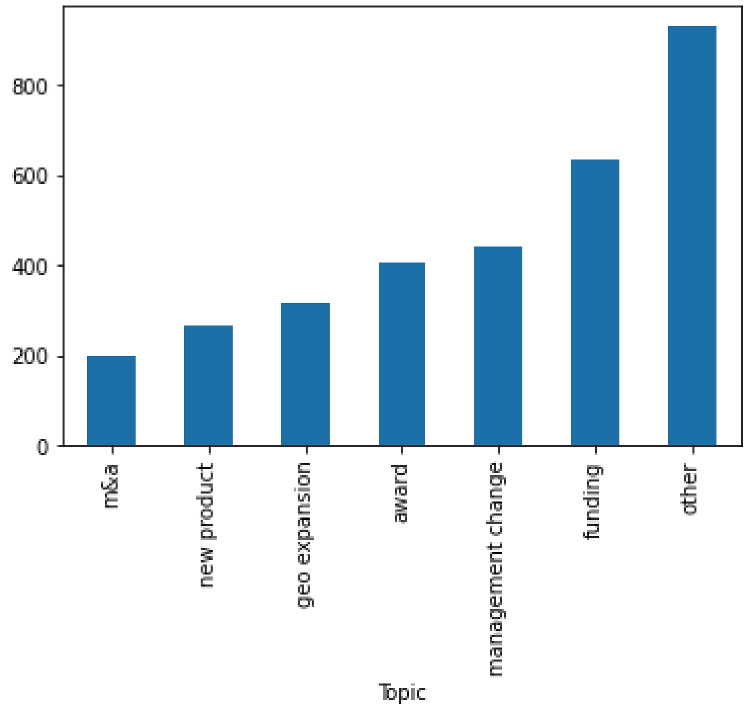


Figure 4.2: Topic Distribution

A bag of words model was used to classify the input text. To convert the tokenized input text into numeric features, the sklearn CountVectorizer was used to create a sparse matrix representation of the token counts. For example, given the two inputs [“I love cats and I love dogs”, “I love my cat”], that will converted to the matrix [(2, 1, 1, 1, 0, 0), (1, 0, 0, 0, 1, 1)] with the numbers representing the number of times [love, cats, and, dogs, my, cat] occurred in the text respectfully. Note, “I” is excluded as the CountVectorizer filters out single character tokens.

The dataset was split 70/30 between the training and test sets, and five different models were trained and evaluated: (a) Logistic Regression, (b) Random Forest, (c) Naive Bayes, (d) Support Vector Machine, and (e) XGBoost. This is described in more detail in Section 5.

4.6 Startup Tracker Website

One key goal of this thesis was to develop a tool for investors and VCs to use to help with the search for high-potential startups. This takes the form of a startup tracker website, where important company information and trends over time can be viewed and the top candidates can be fetched to assist with the search process. The primary focus of the initial prototype was Twitter data, as Crunchbase data can be readily retrieved by investors while no automated process currently exists to fetch and aggregate company Twitter trends.

Eight companies that had raised funding between January 1, 2020 and January 13, 2022 were selected to develop and test the Twitter trend pipeline: Ducalis, Quickframe, PrizePool, Accion Systems, Dataherald, Pentester Academy, Yotascale, and Scienaptic. These specific companies were selected as they all had a Twitter account but ranged in company activity and user engagement, as well as spanned a range of industries and geographic locations. Company Twitter activity, user Twitter activity, and advanced NLP features were calculated and graphed to visualize engagement and trends over time.

4.6.1 Feature Engineering

A sample of the features collected in Section 4.4.5 were collected and displayed on the startup tracker website. They are described in detail below:

1. Company Features: Information about a Twitter account can be retrieved using the user endpoint: https://api.twitter.com/2/users. This endpoint returns the user's follower count, following count, number of tweets, and number of listed tweets along with general account information such as the

user name and website url. However, this information cannot be fetched for past dates, and only returns the current company information. Consequently, additional tweet data was collected using the query described in Table 4.3. This allowed us to calculate the number of tweets posted by the company and the number of likes, quotes, replies, and retweets received for several years prior to the current date.

2. User Features: The query described in Table 4.3 allowed us to fetch all tweets containing the company's Twitter username, website, and all reposts of company tweets. This data was used to calculate the number of users tweeting about the company, along with the number of likes, quotes, replies, and retweets received on those tweets.

3. NLP Features: Several of the top NLP features were displayed on the website to show additional information on Twitter activity. This includes the calculated VADER sentiment, readability scores, synset complexity, and emojis, hashtags, and links identified in the text. Additionally, the most frequently used words, excluding common words such as 'a', 'the', and 'to' were calculated and displayed in a word cloud to diversify the included graphics.

4.6.2 Automating Data Collection

To ensure up-to-date information is displayed on the website, a server was set up on Google Cloud Compute Engine to fetch and aggregate Twitter data [6]. A cron schedule was set to run two jobs daily.

1. Fetch Tweets: A cron job was scheduled at 12am UTC daily to fetch all tweets posted by or about the company within the past 24 hours. The following cron schedule expression was used: *0 0 * * **.
2. Fetch Company Data: A cron job was scheduled at 1am UTC daily to fetch twitter metrics (followers, tweets, etc.) for the company's Twitter account. The following cron schedule expression was used: *0 1 * * **.

The collected tweets and company data were then stored in Google's Cloud Firestore. Every day corresponded to a new document in the tweets collection, with the filename being the date (ex. 2022-12-01) and the contents including all the collected tweets for that day.

4.6.3 Summary Tables

Since millions of tweets are collected and stored in Cloud Firestore, it could take seconds to minutes for all the computations to run and return the correct aggregate data for frontend display. Consequently, summary tables are computed on a daily, weekly, monthly, and quarterly basis to significantly speed up the retrieval time.

For every company, four collections exist containing tweet data grouped by day, week, month and quarter. An additional cron schedule was created to aggregate all the twitter data. First, all the company tweets are retrieved that fit within the provided date frame. For example, if we're computing a weekly summary table, all tweets posted within the past 7 days are returned. Next, for every feature in the summary table, the data is aggregated and the results are stored in Firestore for future retrieval. A few features include the total number of tweets, the number of distinct users, and the average VADER sentiment. On the frontend, the data is then easily retrieved by querying the summary table using a provided date range.

4.6.4 Website Design

The frontend website was created by Kenny Lau for tracking startup trends. It was developed using the React frontend framework and allows visitors to search by name to view company information. Each startup contains pages for viewing Twitter data on company trends, user trends, and NLP features. Additionally, there is a screening option for filtering by the top 25%, 50%, and 75% of startups given a particular combination of startup features. For example, a user may filter to return the top 25% of companies by user tweets and calculated sentiment score.

4.7 Prediction Models

A total of ten ML classification models were implemented in this paper: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Naıve Bayes (NB), Decision Tree (DT), Random Forest (RF), AdaBoost (SB), Gradient Boost (AB), XGBoost (XGB), and CatBoost (CB).

First, the training and test datasets were loaded in from csv files, containing 17,233 datapoints and 3,004 datapoints respectively. There was a higher portion of negative datapoints so the dataset was balanced by upsampling the positive class to reduce bias in the prediction model.

Categorical features such as industry categories (Ex. Software, Health Care, Financial Services) were converted to their respective indices since all of the models except CatBoost require numerical features for prediction. For the CatBoost model, the categorical feature list was passed in along with the training data as input when fitting the model.

Next, the numerical features were scaled using the MinMaxScaler to ensure the model wasn't favoring specific features over others. For example, the amount of money raised is typically in the millions while the number of founders is in the single digits, yet one isn't a million times more important than the other.

Next, each one of the ten models was trained on the pre-processed datasets and evaluated based on criteria described in Chapter 5.

4.8 Real-time Prediction Pipeline (RTPP)

The real-time prediction pipeline combines all the steps described previously in this chapter to output a probability score for a company raising funding within an alloted time frame. The following data is needed to generate a company profile that is used for prediction:

1. Crunchbase ID: The company must have a Crunchbase profile to be assigned a probability of success. Given a Crunchbase URL, the unique organization ID is used for collecting Crunchbase data on the company, including general company data, previous funding round, and published news articles. Given the following Crunchbase link, https://www.crunchbase.com/organization/apple, the company ID is 'apple'.

2. Start Date: The provided start date is used to determine if a company will raise a funding round from that point in time. A past date can be given to test the model's performance by calculating the likelihood of success and then comparing to what actually happened. By default, the current date is used to predict future funding rounds.

3. Year Range: The provided year range can be 1-5 years and will be used to predict whether the specified company will raise a funding round within X years of the given start date.

Given a Crunchbase ID, start date, and year range, the RTPP process retrieves data from Crunchbase, the Google Search API, and Twitter and computes all 171 features as described previously in this section. This can take anywhere between 10 seconds to 3 minutes depending on the number of online activity found for the company. The number of fetched tweets is capped at 6,000, the number of news articles is limited at 2,000, and the Google Search API is called once per company to reduce the computational time and to ensure this research stays within the free-tier API limits.

Chapter 5

EXPERIMENTAL SETUP

The primary goal of this thesis is to predict startup success, measured by the raising of a funding round from investors. This section discusses the processes used for evaluating the prediction models used in achieving that objective. First, it introduces the evaluation metrics used in this paper. Next, it goes over the evaluation setup for the topic classification model. Then, it goes over the steps taken to evaluate the three key components of the funding prediction pipeline: (a) the prediction model used, (b) the feature set, and (c) the time horizon being evaluated. Finally, our best prediction model was compared to related work in the field, with the goal of predicting a Series A or higher round of funding within a one-year horizon.

5.1 Evaluation Metrics

For the topic classification model, the macro F1 score will be used to capture the performance on all seven topic categories. F1 scores can be evaluated on a micro or macro basis, where micro scores give equal importance to each datapoint, while macro scores give equal importance to each class. Macro scores perform better on imbalanced classes since they're not biased towards the dominant class, and will be used to evaluate the topic models.

Since this thesis focuses on predicting successful startups, precision, recall, and F1 scores will be measured for the positive (raised funding) class. The overall (positive class + negative class) scores will be recorded but not used for model and feature selection.

5.2 Topic Modeling

Four machine learning models were created to classify an input text as falling into one of seven categories: (a) funding event (funding), (b) merger and acquisition (m&a) (c) geographical expansions (geo expansion), (d) new product launch (new product), (e) awards received (award), (f) management change, and (g) other. These events were deemed as important milestones for a growing startup and were hypothesised to be helpful in predicting future funding rounds.

To evaluate the performance of the machine learning models, the 3196 hand-labeled news headlines were split 70/30 into training and test datasets. The data was preprocessed as described in Section 4.5 and trained on each one of the five ML models. The macro F1 score for the combined classes was used to determine the best prediction model, and then used to improve performance on each individual class.

After the best prediction model was determined, the precision, recall, and F1 scores for each individual class were calculated and samples of correctly and incorrectly labeled headlines was reviewed. Initially, there was only six classes, excluding m&a, but analysis on the labeled headlines revealed that a high number of merger and acquisition events were classified in the *other* category instead of in *funding.* As those are highly valuable datapoints, the data was then relabeled and a new class was created to capture m&a events. The models were then retrained on the new, seven class dataset and the best model was chosen to label news articles, google search results, and tweets.

5.3 Funding Event Prediction

The process of evaluating the best funding prediction model was fourfold; First, ten different machine learning models were trained and evaluated to determine the highest performer. Next, the feature set was analyzed to determine the best collection of features that yield the highest results. Next, the best model along with the optimal feature set was used to compare how well the model performs on predicting funding events at different ranges into the future. Finally, the dataset was limited to only successful companies and evaluated on how well it could predict a second round of funding for companies that had already raised an initial round.

A total of 9,380 companies were used in this section to generate 20,237 datapoints, approximately 2 per company. These each represented a specific point in time during the company's history; more specifically, January 1st of a given year. Each datapoint contained all the features extracted from Crunchbase, Google, and Twitter and these were used to predict whether or not that company would raise a round of funding within a specified time horizon.

5.3.1 Model Performance

A total of ten different prediction models were evaluated to determine the best predictor for company success, measured by the highest F1 score for the positive (raised funding) class. A total of 171 features were generated for each datapoint, representing general company data, funding events, news articles, Google search results, and Twitter social media activity for the target company.

5.3.2 Feature Engineering

Selecting an optimal feature set is crucial to improve model performance, since including bad features can confuse the models and result in less importance given to features that actually make a difference. The CatBoost algorithm was used to evaluate the feature set as achieved the highest performance out of all ten models.

Feature importance was evaluated in a four step process.

1. Baseline: All the features, 171 in total, were used to establish a baseline metric of model performance.
2. Feature Set: The impact of each feature category: general, funding, news articles, Google Search, and Twitter, was measured by removing those respective fields from the dataset and measuring the difference from the original F1 score. Additionally, the F1 score was calculated using only the feature set, to measure standalone performance.
3. Numeric/Categorical: The impact of each numerical and categorical features was measured by including only the 161 numerical, and then only the 10 categorical features, when training and evaluating the Catboost algorithm.
4. Individual Features: The impact of each feature was measured in two ways: (a) calculating model performance using only that feature and (b) by removing it from the baseline and then capturing the difference between the baseline F1 score and the removed feature F1 score.

The features that had both a high individual score and a high impact when removed from the feature set were added to a reduced feature set. Performance using this new feature list was expected to improve performance, as non-influential and noisy

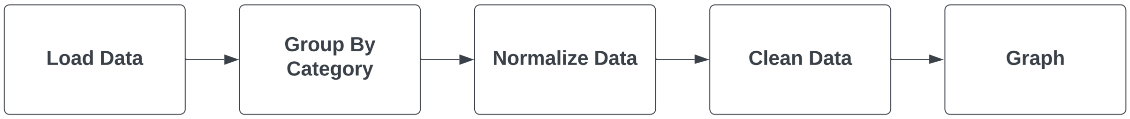


Figure 5.1: Feature Analysis Pipeline

features wouldn't bias the final model and take away focus from features that matter.

5.3.3 Feature Analysis

The features in the final reduced feature set were further analyzed to determine why they performed well on success prediction. This was done by following the steps outlined below (Figure 5.1).

1. Load Data: All 22,125 datapoints were loaded into a Pandas Dataframe.

2. Group by Funding/Funding: The distribution of datapoints across the positive (raised funding) and negative (no funding) class was calculated for the specific feature. For example, for the num_founders feature, the number of companies that raised funding and had 1, 2, etc. founders and the number of companies that didn't raise funding and had 1, 2, etc. founders was computed.

3. Normalize Data: To handle the data being skewed towards the negative class, the data was normalized to a 0-1 scale. This ensures the numbers aren't artificially high due to more companies not raising funding than successfully doing so.

4. Clean Data: The numeric values were optionally rounded to ensure the plotted data is smooth and easily visualized. For example, the sentiment scores ranged from -1 to 1, but having 200 datapoints on a 0.01 scale results in

jagged and unsmooth data. Consequently, sentiment values were rounded to the nearest 0.1.

5. Graph: The final cleaned feature data is plotted as a bar or line graph.

5.3.4 Year Range Comparison

Prior research in startup prediction focused on predicting startup success using a single measure of success, whether that was related to funding rounds, merger and acquisition, or simply surviving for yet another year. However, no one has explored how well prediction models perform on predicting funding a variable number of years into the future.

The same datapoints, each representing a company at a fixed point in time, were used in all forecasting instances. However, the y_value, whether or not the company raised a funding round, was changed to reflect whether or not the company raised money in the upcoming 1, 2, 3, 4, or 5 years. These values are based on whether or not the funding round list collected from Crunchbase contains an entry between the provided point in time and a specified time horizon, given in years. This resulted in five separate datasets that were then trained using a CatBoost model containing the finalized feature set. The percentage of companies that raised funding, along with the precision, recall, and F1 scores of those companies, was reported for each year range.

5.3.5 Probability Cutoff

The Catboost boosting algorithm classifies data into the positive class if it's assigned a probability score of 50% or greater for being in that class. Consequently, precision and recall can be assigned different weights by assigning different probability requirements for being classified as the positive class. In a practice, an investor will only be able to fund a very small number of startups, so they're primarily interested in viewing only the top companies with a very high likelihood of success. Consequently, high precision is typically preferred at the expense of recall, and this ratio can be adjusted by setting the probability cutoff. For example, a cutoff of 80% will only label companies as receiving funding if there's an 80% or greater probability score computed by the model of funding being raised.

5.3.6 Additional Funding Rounds

The top two state-of-the-art prediction models [30, 41] predict whether or not a company will raise a Series A or higher round of funding within the next 12 months. They chose angel and seed rounds as triggers to be included in the dataset, and consequently, all included companies have already raised a funding round.

To measure our model's performance against this existing work, the collected data was filtered to only include companies that had raised an angel or seed round of investment. The y_label was updated to classify a successful funding round as being series A or higher, instead of any funding round. The CatBoost boosting method was trained on the finalized feature set using 0.5 and 0.75 cutoffs, and the precision, recall, and $F_{0.1}$ scores were compared against the WBSSP and FPAWI results.

Chapter 6

RESULTS

This section will report on the performance of the machine learning models used in this paper, finishing with a comparison to prior work in this field. Since we're interested in predicting successful companies, the precision, recall, and F1 metrics for the positive, raised funding class will measured and reported. The overall metrics can be viewed in Table 6.3 but won't be referenced again as we're not interested in identifying unsuccessful companies.

6.1 Topic Modeling

A total of 3,196 hand-labeled news article headlines were used to train and evaluate the best machine learning model for topic prediction. Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine, and the XGBoost algorithm were used, with XGBoost achieving the highest scores of 0.82 for precision, recall, and F1.

Table 6.1: Model Comparison for Topic Prediction

Model	Precision	Recall	F1
Naive Bayes	0.75	0.76	0.75
Support Vector Classifer	0.82	0.78	0.80
Random Forest	0.81	0.80	0.80
Logistic Regression	0.81	0.79	0.80
XGBoost	**0.82**	**0.82**	**0.82**

Table 6.2: Topic Classification Report

Topic	Precision	Recall	F1
new product	0.67	0.61	0.64
m_a	0.75	0.77	0.76
geo expansion	0.70	0.82	0.75
other	0.81	0.75	0.78
funding	0.88	0.90	0.89
award	0.91	0.92	0.91

The XGBoost algorithm was then used to evaluate model performance on an individual topic basis (see Table 6.2). The model performed particularly well on funding and award classification, with F1 scores of 0.90 and 0.92 respectfully. However, it scored low on the new product category, with an F1 score of 0.61. It was able to correctly classify obvious funding news such as "Vannevar Labs Raises $12M in Series A Funding". However, it mislabeled a few ambiguous headlines such as "Mobile.dev launches with $3M seed to catch app issues pre-production", which humans labeled as 'funding' but the model misclassified as 'new product' most likely due to the *launch* keyword.

6.2 Prediction Methods

Scores for each one of the ten prediction methods used can be viewed in Table 6.3. These were trained on the entire feature set of 161 unique numerical fields, and the CatBoost model was trained on an additional 10 categorical fields. The goal was to predict whether or not the company will predict funding 3 years into the future given a starting date.

Table 6.3: Prediction Method Performance

Model	Overall Precision	Overall Recall	Overall F1	Funding Precision	Funding Recall	Funding F1
Naive Bayes	0.6136	0.6399	0.6197	0.3780	0.5212	0.4382
K Nearest Neighbor	0.6407	0.6883	0.6461	0.4009	0.6500	0.4960
Decision Tree Classifier	0.6726	0.6956	0.6815	0.4737	0.5695	0.5172
Random Forest	0.7690	0.7694	0.7692	0.6392	0.6406	0.6399
AdaBoost	0.7435	0.8147	0.7624	0.5452	0.8227	0.6558
Support Vector Classifier	0.7462	0.8090	0.7654	0.5579	0.7955	0.6558
XGBoost	0.7587	0.8027	0.7755	0.5935	0.7500	0.6627
Gradient Boost	0.7479	0.8204	0.7673	0.5513	**0.8312**	0.6628
Logistic Regression	0.7523	0.8130	0.7718	0.5694	0.7955	0.6637
CatBoost	**0.8056**	**0.8545**	**0.8250**	**0.6634**	0.8273	**0.7363**

The CatBoost boosting algorithm achieved the highest precision and F1 scores of 0.663 and 0.736. The Gradient Boost algorithm achieved the highest recall score of 0.831 but its average performance in precision consequently placed it third on the scoring list. Classic machine models such as Naive Bayes, K-Nearest Neighbors, and Decision Trees scored significantly lower than the other methods, achieving F1 scores of 0.438, 0.496, and 0.517 respectively. Logistic Regression achieved comparable performance to the top boosting algorithms, with an F1 score of 0.664.

Since the CatBoost ensemble method outperformed the other classification models, it will be used for all further prediction tasks.

6.3 Feature Engineering

A total of 171 features were collected for each company in the training and evaluation datasets. This consists of 161 numerical and 10 categorical features, and the categorical ones were only used by the CatBoost boosting algorithm. These features are distributed into five distinct categories: (a) 22 general, (b) 6 funding, (c) 10 news articles, (d) 8 Google Search, and (e) 78 Twitter and can be viewed in Table 6.4. The complete feature list can be found in Appendix B along with the precision, recall, and F1 scores for classification using only that individual feature.

The 47 industry features represent a sparse matrix of whether or not the company listed a specific industry category on their Crunchbase profile. These were combined into 6 categorical features that represent the first 6 categories the company listed on their profile to help reduce the vector space.

The five feature sets, industry features were grouped into the general category, were individually evaluated to determine their impact on funding prediction. Results can

Table 6.4: Feature Set Distribution

Category	Numerical	Categorical	Total
General	13	9	22
Industry*	47	0	47
Funding	5	1	6
News Articles	10	0	10
Google Search	8	0	8
Twitter	78	0	78
All	161	10	171

Table 6.5: Performance by Feature Set

Category	# Features	Precision	Recall	F1
categorical	10	0.5216	0.6939	0.5956
numeric	161	0.6895	0.6864	0.6879
google	8	0.4305	0.0985	0.1603
article	10	0.5556	0.1970	0.2908
twitter	78	0.5269	0.2667	0.3541
industries	47	0.4584	0.3091	0.3692
funding	6	**0.7553**	0.3227	0.4522
general	22	0.5747	**0.7288**	**0.6426**
excluding general	149	0.6476	0.4621	0.5393
excluding funding	165	0.6293	0.7409	0.6806
excluding twitter	93	0.6521	0.8121	0.7233
excluding google	163	0.6468	**0.8242**	0.7249
excluding industries	124	0.6546	0.8212	0.7285
excluding article	161	**0.6613**	0.8167	**0.7308**
Top Features (Selected)	18	0.6303	0.8136	0.7103
Top Features (CatBoost)	18	0.6423	0.8136	0.7179
All Features	171	**0.6634**	**0.8273**	**0.7363**

be viewed in Table 6.5. General company data collected from Crunchbase such as months since founding, number of founders, headquarters location, and company name length had the biggest impact on model performance.

The CatBoost boosting algorithm was trained and evaluated using only numerical features and only categorical features. Adding categorical features improved performance by 7.0%, raising the F1 score from 0.6879 to 0.7363. It also achieved a high F1 score of 0.5956 using only the 10 categorical features. This feature set consisted of the company description, headquarters location, industry categories, and last funding stage. This shows that entering the right market in a startup rich zone greatly correlates with company success.

One way to measure a feature's impact on the overall model performance is to exclude that feature from the feature list and measure the effect on classification. This was done using the entire feature list as the baseline. The top three features that resulted in the highest performance drop once removed are as follows: (1) number of founders, (2) the last funding stage, and (3) the number of months since founding. Additionally, the top three features that resulted in the highest performance using only that feature were: (1) number of founders, (2) the ratio between distinct and total investors, and (3) the number of previous funding rounds. The number of founders and the last funding stage were in both the top 10 excluded and solo lists. The entire list of 18 high performing features can be viewed in Table 6.6. Using the final reduced feature set resulted in an F1 score of 0.7103, only a 3.7% drop compared to the entire feature list. This is a very high score considering only 18/171 features were used.

Table 6.6: Top Features

field	category	precision (solo)	recall (solo)	f1 (solo)	precision (excluding)	recall (excluding)	f1 (excluding)	f1 impact (excluding)
num_founders	general	0.6106	0.6818	0.6442	0.6256	0.7697	0.6902	0.0415
distinct_total_investor_ratio	funding	0.5480	0.3894	0.4553	—	—	—	—
num_prev_funding_rounds	funding	0.5480	0.3894	0.4553	—	—	—	—
months_since_last_funding	funding	0.6032	0.3409	0.4356	—	—	—	—
last_funding_stage	funding	0.8194	0.2818	0.4194	0.6329	0.7864	0.7014	0.0303
total_prev_investments	funding	0.5929	0.2803	0.3807	—	—	—	—
num_distinct_investors	funding	0.5890	0.2758	0.3756	—	—	—	—
other_label	articles	0.4790	0.2591	0.3363	—	—	—	—
avg_paragraphs	twitter	0.4680	0.2545	0.3297	—	—	—	—
avg_total_shape	twitter	0.4462	0.2576	0.3266	—	—	—	—
months_since_founding	general	—	—	—	0.6398	0.8045	0.7128	0.0189
description	general	—	—	—	0.6912	0.7530	0.7208	0.0109
country	general	—	—	—	0.6612	0.7924	0.7209	0.0108
Internet Services	general	—	—	—	0.6502	0.8167	0.7240	0.0077
avg_long_words	twitter	—	—	—	0.6476	0.8212	0.7241	0.0076
avg_synset_ratio	twitter	—	—	—	0.6537	0.8121	0.7243	0.0074
management_change_label	twitter	—	—	—	0.6537	0.8121	0.7243	0.0074
name_length	general	—	—	—	0.6548	0.8106	0.7244	0.0073

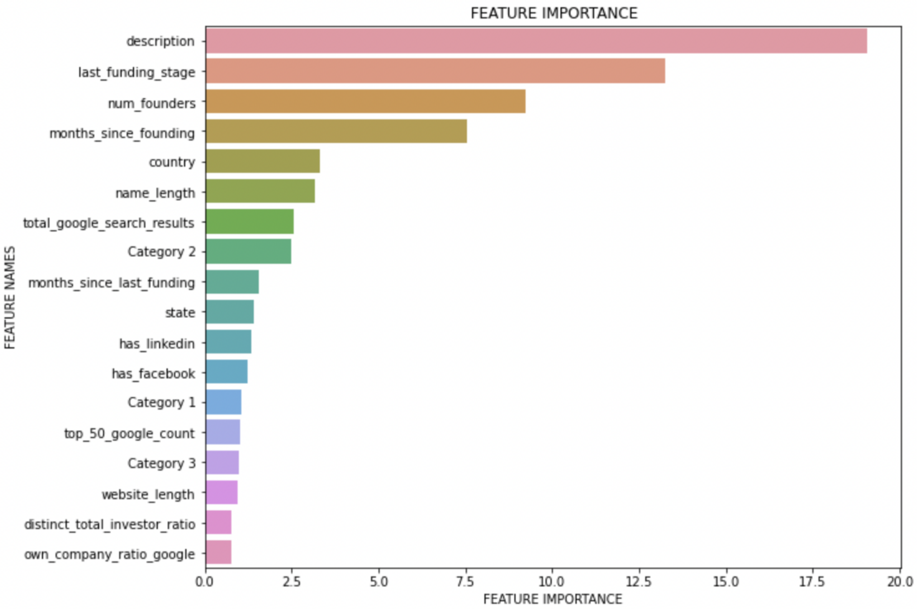


Figure 6.1: Top 18 Catboost Features

Catboost saves the weights of each feature and this can be used as a proxy for feature importance, as a higher weight means the output will be more significantly impacted given an equal difference in the input. A list of the top 18 features by Catboost training weight can be viewed in Figure 6.1. A model was trained using only the aforementioned 18 features and an F1 score of 0.718 was achieved, scoring higher than the manually selected top 18 features but still not as high as using the entire feature set.

Feature engineering wasn't able to find an optimal feature set that performed better than the entire feature list. This is probably due to the smart weighing of features during Catboost training, and the effects of irrelevant features are consequently minimized if not completely eliminated. For future prediction tasks, the entire feature list will be used.

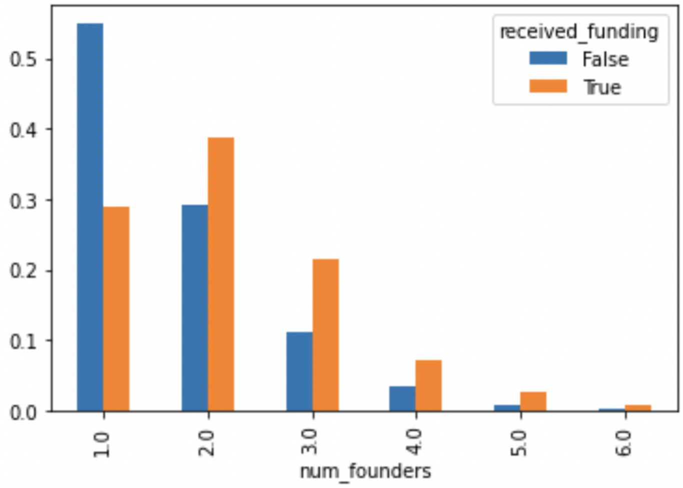


Figure 6.2: Number of Founders

6.4 Feature Analysis

In addition to identifying high-potential companies, investors and founders want to know what factors lead to a company's success. This section dives into the key features that make a difference between successful and unsuccessful companies.

An F1 score of 0.611 can be achieved using only the number of founders listed on a company's Crunchbase profile. As shown in Figure 6.2, a company is pretty much guaranteed to not have raised funding if they have zero listed founders. While the company most certainly has a founder, this could represent a lack of responsibility or initiative from the founder. Additionally, it could mean that the company never did well enough for anyone outside the organization to care enough to input the founders' information.

A majority of companies in our dataset, 70% of unsuccessful companies and 30% of successful companies, had a seed round of funding as their last fundraising stage (Fig. 6.3). This seems to be the bottleneck in startup funding, as most companies, and especially unsuccessful companies, never reach another funding round.

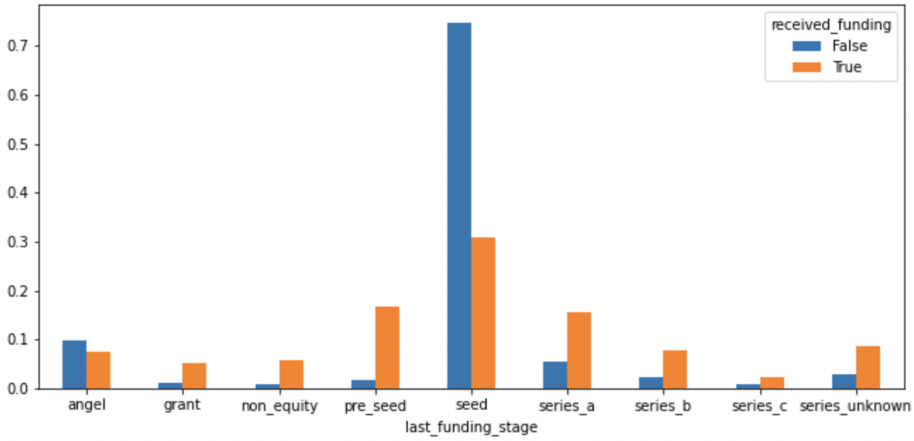


Figure 6.3: Last Funding Stage

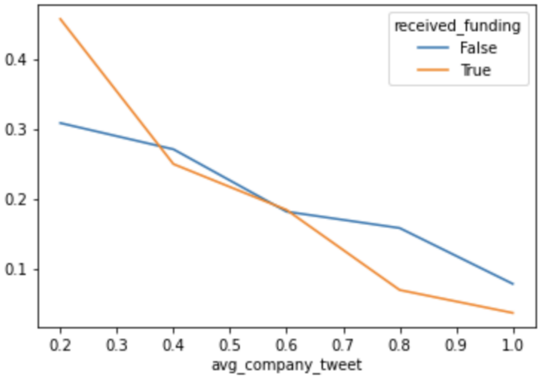


Figure 6.4: Company Tweet Ratio

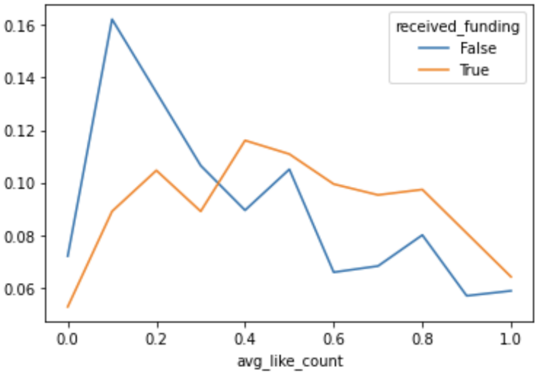


Figure 6.5: Likes Received

However, more analysis is needed to determine if the company didn't raise a funding round because of poor company performance, or because they were already doing well and didn't need to raise any outside investment.

As shown in Figure 6.4, having a lower percentage of company tweets, calculated by determining how many of the total returned tweets were posted by the company's Twitter account, corresponds to a higher likelihood for success. This measures public engagement, meaning more people actively discussing the company leads to a higher chance of raising funding.

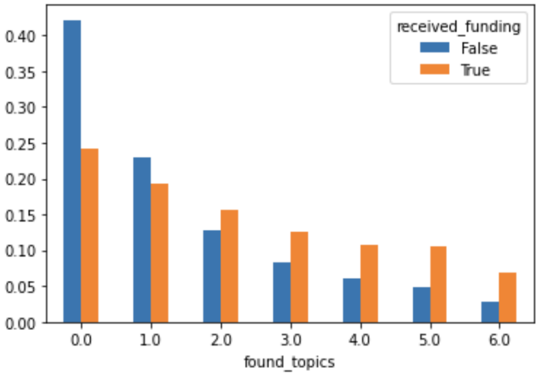


Figure 6.6: Tweet Topics

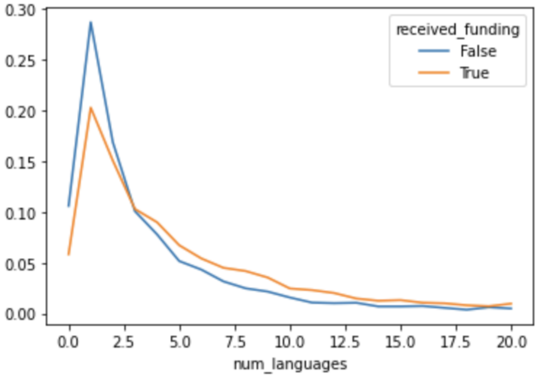


Figure 6.7: Tweet Languages

The average like count, representing how many likes a tweet receives on average, correlates with a higher likelihood of success (Fig. 6.5). Similar to the company tweet ratio, this feature measures user engagement with the company on social media.

A higher number of unique company milestone topics identified in the returned tweets, ranging from 0-6, also correlates with a higher chance of raising a round of funding (Fig. 6.6). This means people are discussing a diverse set of topics, ranging from new product launches, management changes, and funding round, on social media.

The number of languages contained in the returned tweets, while not as significant as the previously mentioned features, did have a small correlation with raising a funding round (Fig. 6.7). This shows that having a diverse group of people knowing about and engaging with the company increases the likelihood of success.

The distribution of company sentiment between companies that raised and didn't raise funding is shown in Figure 6.8. Companies with more positive language on social media were more likely to raise funding than those with more negative and neutral content. However, after a certain limit, being too positive started to decrease chances of success, potentially due to users not believing the over-optimism

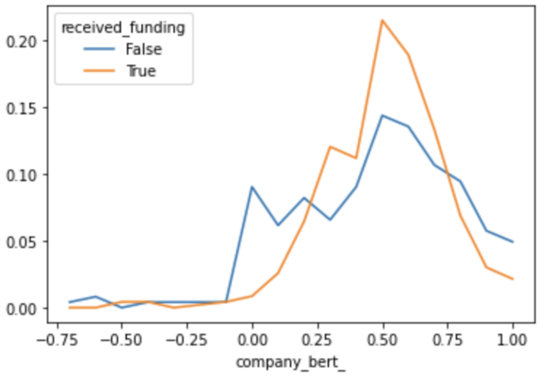


Figure 6.8: Company Sentiment
Bert

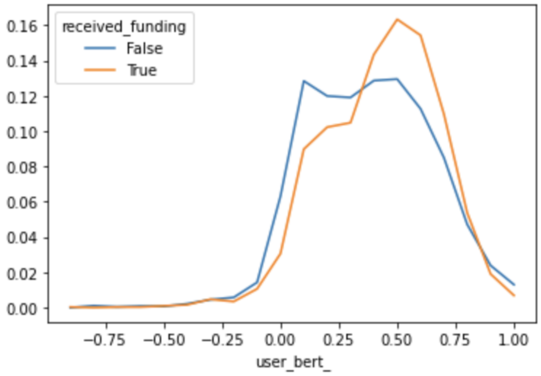


Figure 6.9: User Sentiment
Bert

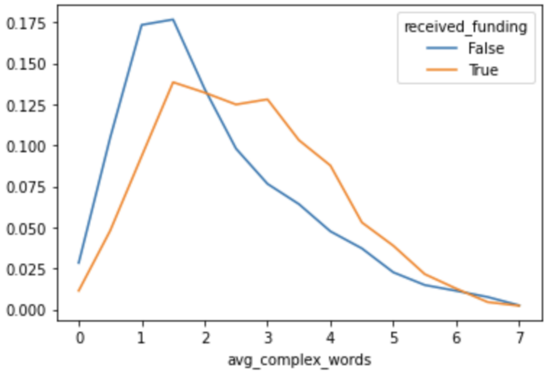


Figure 6.10: Complex Words
3 or more syllables

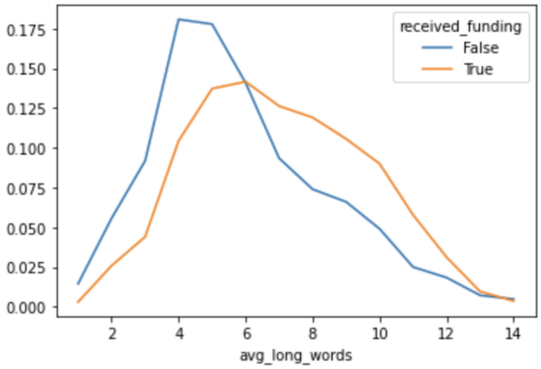
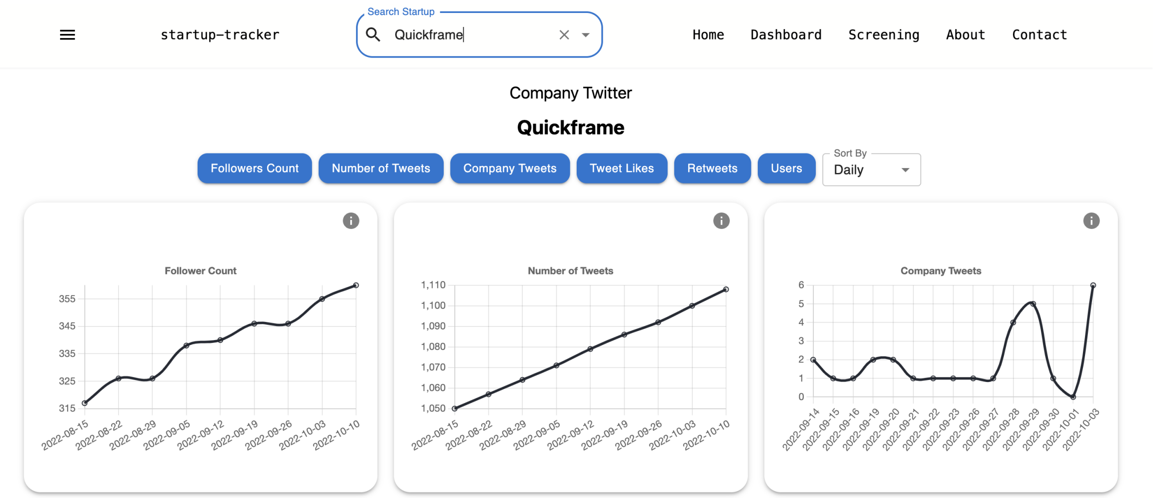


Figure 6.11: Long Words
More than 7 characters

of the startup team. As seen in Figure 6.9, a higher user sentiment also corresponds to a higher likelihood of being funded, but on a smaller extent than company tweets.

Finally, the complexity of tweets is found to correlate with increased funding. The distribution of the average amount of complex words per tweet, words with three or more syllables, and the average number of long words per tweet, characterized by being more than seven characters, can be viewed in Figures 6.10 and 6.11. Having a higher complexity means more work and effort was put into the content shared on social media, both by the company and by other users. It's important to note that these features are highly correlated, meaning they measure the same metric. Consequently, it may result in the model placing too much emphasis on tweet


Search Startup: Quickframe

Sort By: Daily

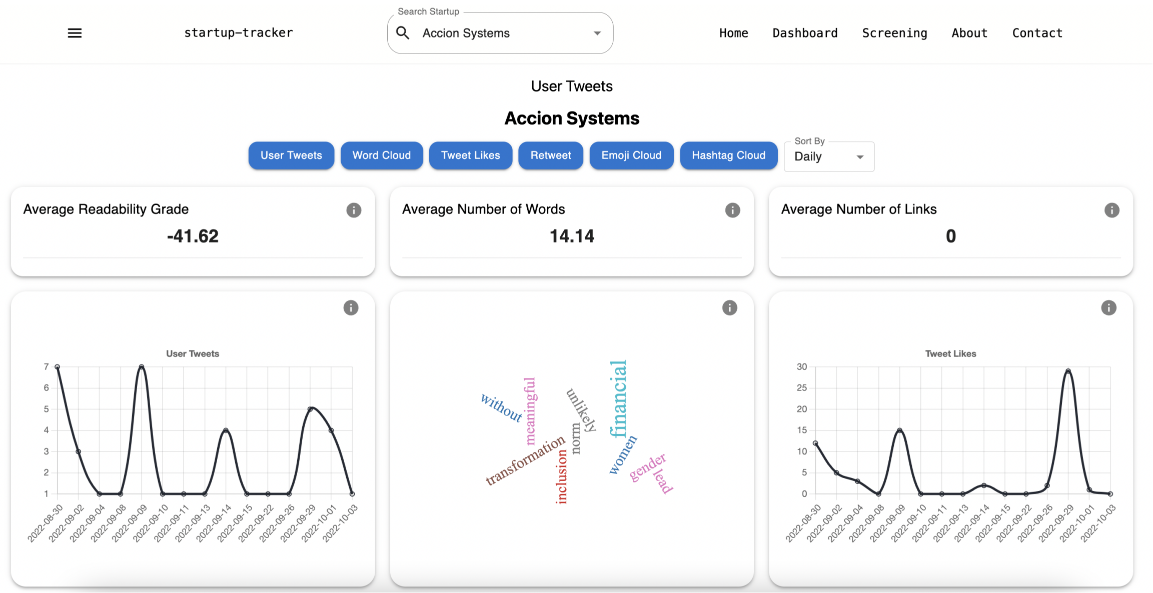
Figure 6.12: Startup Tracker Website - Company Twitter Features

complexity, and future work is needed to remove similar correlated variables in the feature set.

Additional feature graphs can be viewed in Appendix D.

6.5 Startup Tracker Website

The startup tracker website is currently up and running on a Google Cloud Compute Engine server and contains data for eight companies: Ducalis, Quickframe, PrizePool, Accion Systems, Dataherald, Pentester Academy, Yotascale, and Scienaptic. Daily, weekly, monthly, and quarterly summary tables were created for 29 different features, and 21 are currently supported on the site: six under Company Twitter, nine under User Tweets, and six under Advanced NLP. Snapshots of the website can be viewed in Figures 6.12, 6.13, and 6.14.



Search Startup: Accion Systems

User Tweets

Accion Systems

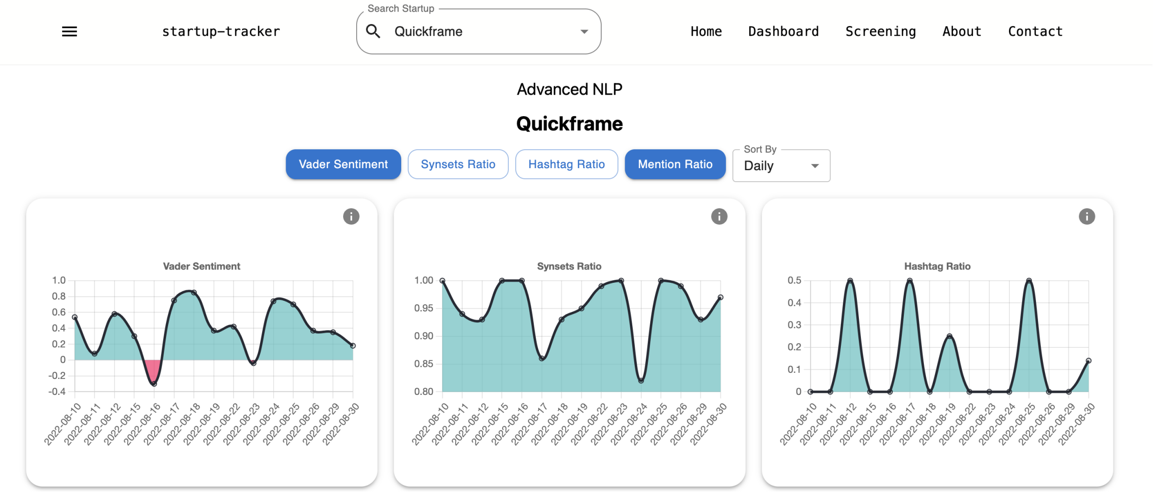
Sort By: Daily

Average Readability Grade: -41.62

Average Number of Words: 14.14

Average Number of Links: 0

Figure 6.13: Startup Tracker Website - User Twitter Features



Search Startup: Quickframe

Advanced NLP

Quickframe

Sort By: Daily

Figure 6.14: Startup Tracker Website - Advanced NLP Features

Table 6.7: **Predicting Funding Variably into the Future**

Range	Precision	Recall	F1	Received Funding
1 year	0.6224	0.4587	0.5282	10.9%
2 years	0.6379	0.7349	0.6830	17.1%
3 years	0.6634	0.8273	0.7363	22.0%
4 years	0.6847	0.8622	0.7633	22.5%
5 years	**0.7042**	**0.8660**	**0.7768**	22.6%

6.6 Year Range Comparison

For the previous two sections of this paper, we focused on predicting whether or not a company will raise a round of funding within 3 years of a given date. This section explores how well the prediction model performs on different date ranges. The CatBoost ensemble method was trained on all 171 collected features with the y_value representing whether or not the company raised money 1, 2, 3, 4, and 5 years into the future. The results can be viewed in Table 6.7.

For starters, the distribution of datapoints in the positive class varied greatly between the different date ranges. Only 10.9% of startups in the evaluation dataset raised funding within a year, compared to 22.6% that raised funding looking 5 years into the future, a 107% increase. This means there are less positive examples in the training set for the model to train on, and the CatBoost method achieved an F1 score of 0.5282 with the limited training data.

The model performance improves as the year range increases, scoring 0.6830, 0.7363, 0.7633, and 0.7768 looking 2, 3, 4, and 5 years into the future. The model experiences the biggest jumps in improvement between year 1 and 2 (29.3%) and year 2 and year 3 (7.8%). After year 3, the rate of improvement drops to 3.7%.

Table 6.8: Performance by Cutoff Threshold

Cutoff	Precision	Recall	F_1	$F_{0.1}$
0.99	**1.0000**	0.0061	0.0120	0.3811
0.97	**1.0000**	0.0424	0.0814	0.8173
0.95	0.9512	0.1182	0.2102	0.8892
0.90	0.9505	0.2621	0.4109	**0.9265**
0.85	0.9039	0.3848	0.5399	0.8920
0.80	0.8676	0.4864	0.6233	0.8609
0.75	0.8283	0.5773	0.6804	0.8247
0.70	0.7792	0.6364	0.7006	0.7775
0.65	0.7562	0.6955	0.7245	0.7555
0.60	0.7269	0.7500	0.7383	0.7271
0.55	0.7009	0.7955	**0.7452**	0.7018
0.50	0.6634	**0.8273**	0.7363	0.6647

Additionally, while the number of companies raising funding more than doubled between year 1 and year 3, there's only a 2.7% increase between year 3 and 5. Since investors want a return on investment as soon as possible, the paper primarily uses the 3 year range to balance real-world usability with model performance.

6.7 Probability Cutoff

Classification scores for cutoff thresholds ranging from 0.5 to 0.95 were calculated in steps of 0.05 (see Table 6.8). Additionally, performance was measured with the cutoff set at 0.97 and 0.99 to capture the highest bias towards precision over recall. Increasing the cutoff threshold for being classified in the positive class steadily improved precision, starting at 0.6634 and reaching the maximum precision of 1.0 at the 0.97 cutoff threshold. In contrast, recall steadily dropped from 0.8273 to 0.0061, meaning less and less companies fit the steep cutoff threshold to be classified as successful. The best F_1 score was achieved at the 0.55 cutoff, where precision and recall were almost equally favored. Setting beta at 0.1 heavily favors precision over recall, and the best $F_{0.1}$ score was achieved at the 0.9 cutoff with a score of 0.9265. At the 0.9 cutoff threshold, a precision of 0.9505 and a recall of 0.2621 were achieved, meaning 175/660 companies of the companies that raised funding were correctly identified and there's a 95% chance that a company that a company with a successful classification will actually raise funding. However, the dataset isn't large enough to fully generalize a 95% probability score across all startups, industries, and countries.

6.8 Comparison to Related Work

To complete the evaluation of our results, our best CatBoost boosting model using the entire feature set was compared to prior work in the startup prediction space.

Sharchilev et al. achieved state-of-the-art performance in 2018 with their Gradient Boosting Decision Tree trained on Crunchbase, LinkedIn, and internet data collected from the Yandex Search Engine [41]. Their model was trained on 21,947 companies collected from Crunchbase through May 2014, and tested on an additional 15,128 companies with snapshots taken between May 2015 and May 2016. They had a total of 49 distinct features, ranging from general company data, previous funding rounds, founder backgrounds, and web presence metrics. Garkavenko et al. further advanced the field of startup success prediction with their 2022 paper using online, publicly available data from the startup's own website, the Google Search API, and Twitter to collect 17 distinct features for each startup [30]. They used the Catboost boosing algorithm trained on 33,165 datapoints to predict future funding rounds.

Table 6.9: Comparison to results reported in [30] and [41]

Model	Abbv.	Description
Our Best Model	$CB_{0.75}$	Catboost with 0.75 confidence cutoff
Default Cutoff Catboost	$CB_{0.50}$	Catboost with default 0.5 confidence cutoff
Predicting Startup Funding From Freely, Publicly Available Web Information [30]	FPAWI	Catboost trained on startup's own website, Twitter API, and Google Search API data
Web-based Startup Success Prediction [41]	WBSSP	Gradient Boosting Decision Tree trained on Crunchbase, Linkedin, and Yandex Search Engine data

Model	Datapoints	Features	Precision	$F_{0.1}$
$CB_{0.75}$	1,888	171	0.744	0.730
$CB_{0.50}$	1,888	171	0.656	0.655
FPAWI	33,165	17	0.640	0.531
WBSSP	15,128	49	0.626	0.383

Although the research papers didn't share the datasets used, care was taken to replicate the experimental setup as best as possible. In particular, an angel or seed round of investment was a required trigger to be considered, and the goal was to predict a Series A or later round of funding within one year of a provided date.

The results are displayed in Table 6.9. Our default CatBoost model trained on the entire feature set achieved state-of-the-art performance with precision and $F_{0.1}$ scores of 0.656 and 0.655 respectively. The WBSSP and FPAWI models achieved $F_{0.1}$ scores of 0.383 and 0.531 at the same task of identifying Series A or later funding rounds. Furthermore, our best Catboost model adjusted with a cutoff threshold of 0.75 achieved precision and recall scores of 0.744 and 0.730, surpassing all other models. While increasing the cutoff threshold to 0.90 further increased the precision and $F_{0.1}$ scores to 1.0 and 0.869, this wasn't used as the final model as the dataset isn't large enough to dismiss overfitting in the evaluation set (see App. C).

6.9 Real-time Prediction Pipeline

The real-time prediction pipeline takes in a company's Crunchbase ID, a start date, and a time window (1-5 years) and outputs a probability score for the company raising funding within that year range.

An output will follow one of two formats:

1. Predicted to raise funding with X% confidence
2. No funding predicted with Y% confidence

Let's take the company Xano for reference. They are creating software for building scalable, no-code backend solutions. Giving our prediction model the following input: ('xano', 2020-11-01, 2 years), it will return the following output:

```
Predicted to raise funding with 68.7% confidence
Company raised 1 funding round in the provided time window.
Model prediction successful
```

Chapter 7

CONCLUSION

This thesis explores whether company information collected from online data sources can be used to predict startup success. More specifically, can we predict whether or not a company can raise a round of funding within an allotted time frame.

Information on general company data, previous funding rounds, published news articles, Google Search results, and Twitter social media activity was collected for 9,842 companies on Crunchbase. These companies were founded between 2007 and 2021 and 4,086 (41.5%) had raised at least one funding round. A total of 171 features was collected for every datapoint composed of 161 numerical and 10 categorical features. Feature engineering did not reveal an optimal reduced feature set, as Catboost significantly if not completely reduces the impact of negative features by adjusting feature weights. Consequently, we found that our best performance was achieved using the Catboost algorithm using the entire feature set.

The CatBoost ensemble method achieved the best performance with precision, recall, and F1 scores of 0.663, 0.827, and 0.736 respectively at the 3 year prediction task. The same ensemble method achieved F1 scores of 0.528, 0.683, 0.736, 0.763, and 0.777 when tasked with predicting a funding round one to five years into the future. The Catboost model was then adjusted to favor precision over recall by increasing the cutoff threshold, and a maximum $F_{0.1}$ score of 0.927 was reached with a threshold of 0.9. The final objective was to predict whether or not a startup that had already raised an angel or seed round of funding would raise another round within one year of the provided date. Our best CatBoost model trained on all 171 features and a 0.75 cutoff threshold achieved precision and $F_{0.1}$ scores of 0.744 and 0.730, beating the results of previous work in this field.

The number of founders, the last fundraising stage, the number of previous funding rounds, the number of months since last funding, the ratio of distinct to total investors, the number of months since founding, and the country where the company is headquartered are some of the top features for startup success prediction. In particular, we were able to predict funding with an F1 score of 0.611 solely using the number of founders, as most companies that didn't have a listed founder didn't raise a funding round. While the company most certainly had a founder, this could correlate with a lack of responsibility and initiative from the founder or a lack of interest from the community to report the founders' information.

The final contributions of this work are tools for investors and VCs to use to assist in their search for high growth startups with a high likelihood of success. The first is a Startup Tracker website that follows trends in portfolio companies, focused initially on monitoring Twitter activity and engagement. The culmination of this thesis is a real-time prediction pipeline that takes in a company's Crunchbase ID, a start date, and a time window (1-5 years) and outputs the probability of the company raising funding within the provided year range.

7.1 Reflection

It's important to note that while several features have been found to highly correlate to startup success, they don't capture the internal workings of a company. The number of founders, the company description, and the complexity of tweets are indicators of the underlying workings of a company, but can be artificially mimicked to fool the prediction model. Simply finding a random person to join your co-founder team and then posting lots of content on social media, while following what successful companies are doing, fails to address the key components that make a startup succeed. Having a co-founder typically brings in an additional set of skills and expertise to a business, while actively engaging on social media may portray a growing business with a large HR team. Consequently, it's advised to use this prediction model as the first step to screening startups, and then doing due diligence to confirm that a business is worth investing in.

7.2 Future Work

Our work has helped improve startup success prediction and increased understanding of what characteristics indicate a successful startup. However, there is more work that can be done in this area to enhance the dataset and to provide individuals with actionable insights using the prediction model.

7.2.1 Data Collection

We were limited in the number of companies contained in our dataset due to the limited duration of this research and the API limits set by the Twitter and Google Search APIs. Future work utilizing a higher volume of companies would be better equipped to avoid over-fitting to the training set. The 300 million startups created each year are made up of a diverse range of industries, geographic regions, and growth profiles, and a larger dataset would better capture the various trends across the startup space.

7.2.2 Founder Profiles

Previous work into startup prediction has found correlation between the founders' background and the success of their company. Information contained on Crunchbase and LinkedIn can be used to collect data on a founder's educational history, previous ventures, and social networks. These additional fields may provide clues into the individuals' professional network, as people often say one's net worth is their network. Additionally, a founder with a history of successful ventures is more likely to start another successful company than a first-time founder.

7.2.3 News Article Publishers

This research hypothesized that startups featured by top startup news sites would be more likely to succeed. However, oftentimes being featured on rare news channels is more indicative of high growth and potential, as it's a lot more impressive to be featured on Fortune or CNN than on TechCrunch, which publishes a much higher volume of startup news. Consequently, future work on identifying publications by these highly elite and sought-out news channels would give a better glimpse on startup achievement.

7.2.4 Company Size

Small, early-stage companies with less than a dozen employees need a lot less funding than huge businesses overseeing hundreds to thousands of people. Consequently, the amount of funding, and the extent to which funding is needed, varies significantly across the stage the company is at and how many people need to be paid. While a partial view of the stage of a business is contained in the months since founding and the last investment stage features, taking into account the company size will help paint a more wholistic view of the business. Additionally, normalizing metrics by company size can help alleviate the huge disparity between early to later-stage startups, as the number of likes received on social media will vary significantly between new ventures and businesses with thousands of employees and millions of dollars in revenue.

7.2.5 State of the Economy

It's important to recognize that the amount of money invested varies year by year across the venture capital space. Some years may be booming times for startup investments, while recessions dry up funds available for growing startups. Consequently, a startup that would have raised a round of funding one year may not during a period of low investment. Taking into account the state of the economy at the time of prediction will help further improve the prediction model.

7.2.6 Twitter-Trained Topic Modeling

The topic model used in this paper was trained on news headlines and wasn't specifically built for social media data. Consequently, it didn't perform as well on Twitter data as on the news article dataset. Further development on the topic model could significantly improve model performance.

7.2.7 Performance Across Company Backgrounds

This paper focused on predicting whether or not a company would raise a funding round, but didn't dive into what companies it was particularly good or bad at predicting. In particular, it would be interesting to see how well the prediction model performs across different industries: does it favor one sector such as Software or Biotech over others. Additionally, further analysis is needed to determine how well the model fares across different states and countries.

7.2.8 Expanding Startup Tracker Website

A key objective of this paper was to provide individuals and investors practical knowledge and tools to assist in the search for early stage startups that have a high likelihood of success. While the Startup Tracker website tested the functionality and feasibility of the data collection and aggregation system, it's currently in the prototype stage and only supports eight companies. Further development is needed to scale the system and increase the supported company list. Additionally, the real-time prediction pipeline is not connected to any user interface, and isn't easily accessible by the public. Connecting it to the Startup Tracker website will enable individuals to run their own analysis on companies of interest, putting the tools and methodologies described in this paper to wide use.

BIBLIOGRAPHY

[1] Cal Poly Github. http://www.github.com/CalPoly.

[2] Cron Scheduling. https://en.wikipedia.org/wiki/Cron.

[3] Crunchbase Daily CSV Export. https://data.crunchbase.com/docs/daily-csv-export.

[4] Crunchbase Homepage. https://about.crunchbase.com/about-us/.

[5] Custom Search JSON API - Programmable Search Engine - Google Developers. https://developers.google.com/custom-search/v1/overview.

[6] Google Cloud Compute Engine. https://cloud.google.com/compute.

[7] Google Firebase - Cloud Firestore. https://firebase.google.com/docs/firestore.

[8] Hits algorithm. https://en.wikipedia.org/wiki/HITS_algorithm.

[9] How long do firms take from founding to IPO? https://lao.ca.gov/LAOEconTax/Article/Detail/685.

[10] How Startup Funding Works – Infographic. https://blog.adioma.com/how-funding-works-splitting-equity-infographic/.

[11] How to start and successfully sell a startup. https://www.forbes.com/sites/abdoriani/2019/12/13/how-to-start-and-successfully-sell-a-startup/?sh=71df99942cc0.

[12] Hugging face bert sentiment. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest.

[13] Initial public offering (IPO). https://www.investopedia.com/terms/i/ipo.asp.

[14] Introducing the inc. 5000 fastest-growing private companies in America. https://www.inc.com/inc5000/2021.

[15] Natural language toolkit (NLTK). https://www.nltk.org/.

[16] Passivepy: a tool to automatically identify passive voice in big text data. https://github.com/mitramir55/PassivePy.

[17] Pitchbook Homepage. https://pitchbook.com/.

[18] Series a, b, c funding: How it works. https://www.investopedia.com/articles/personal-finance/102015/series-b-c-funding-what-it-all-means-and-how-it-works.asp.

[19] Spacy: industrial-strength natural language processing in python. https://spacy.io/.

[20] Topic modeling starter code. https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/.

[21] Venture capitalist vs. angel investor | who should you pitch to? https://www.patriotsoftware.com/blog/accounting/venture-capitalist-vs-angel-investor/.

[22] Wordnet. https://wordnet.princeton.edu/.

[23] How catboost algorithm works in machine learning, Jan 2021. https://dataaspirant.com/catboost-algorithm/.

[24] Flesch–Kincaid readability tests, Jul 2022. https://en.wikipedia.org/wiki/FleschKincaid_readability_tests#Flesch_Reading_Ease.

[25] P. Antosiuk. Predicting startup success with machine learning methods, 2021.

[26] T. Antretter, I. Blohm, D. Grichnik, and J. Wincent. Predicting new venture survival: A twitter-based machine learning approach to measuring online legitimacy. 2019.

[27] S. Atske. News on twitter: Consumed by most users and trusted by many, Apr 2022.

[28] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation, 2003.

[29] D. Dellermann, N. Lipusch, P. Ebel, K. M. Popp, and J. M. Leimeister. Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method.

[30] M. Garkavenko, E. Gaussier, H. Mirisaee, C. Lagnier, and A. Guerraz. Where do you want to invest? predicting startup funding from freely, publicly available web information.

[31] K. Hu, D. Grimberg, and E. Durdyev. Twitter sentiment analysis for predicting stock price movements.

[32] T.-C. Huang, R. Zaeem, and S. Barber. Stock price prediction using trust filters and enhanced user sentiment on twitter, Nov 2019.

[33] N. ISP. How many tech startups are created each year?

[34] S. Kampakis and A. Adamides. Using twitter to predict football outcomes.

[35] A. Khan, H. Zhang, N. Boudjellal, A. Ahmad, J. Shang, L. Dai, and B. Hayat. Election prediction on twitter: A systematic mapping study, Apr 2021.

[36] P. Koukaras, C. Nousi, and C. Tjortjis. Stock market prediction using microblogging sentiment analysis and machine learning, May 2022.

[37] R. A. Mendoza-Urdiales, J. A. Núñez-Mora, R. J. Santillán-Salgado, and H. Valencia-Herrera. Twitter sentiment analysis and influence on stock performance using transfer entropy and EGARCH methods.

[38] B. Murray. Overcoming AI bias in predicting startup success.

[39] A. Retterath. What‘s the best startup/VC database?

[40] Salton and McGill. Introduction to Modern Information Retrieval, 1983.

[41] B. Sharchilev, M. Roizner, A. Rumyantsev, D. Ozornin, P. Serdyukov, and M. de Rijke. Web-based startup success prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, 2018.

[42] J. Shepherd. 22 essential twitter statistics you need to know in 2022, Nov 2022.

[43] J. Singh, Y. Dwivedi, N. Rana, A. Kumar, and K. Kapoor. Event classification and location prediction from tweets during disasters, Mar 2017.

[44] J. Soler, F. Cuartero, and M. Roblizo. Twitter as a tool for predicting elections results, 2012.

[45] R. Stuart and P. Abetti. Start-up ventures: Towards the prediction of initial success, Nov 1987.

[46] S. Tomy and E. Pardede. From uncertainties to successful start ups: A data analytic approach to predict success in technological entrepreneurship.

[47] A. Tumasjan, R. Braun, and B. Stolz. Twitter sentiment as a weak signal in venture capital financing. 2021.

[48] I. Vayansky and S. A. Kumar. A review of topic modeling methods. 94:101582.

[49] B. Weng, L. Lu, X. Wang, F. Megahed, and W. Martinez. Predicting short-term stock prices using ensemble methods and online data sources, 2018.

[50] G. Xiang, Z. Zheng, M. Wen, J. Hong, C. Rose, and C. Liu. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch. 2012.

[51] B. Yoo and J. T. Rayz. Understanding emojis for sentiment analysis.

[52] X. Zhang, H. Fuehres, and P. Gloor. Predicting stock market indicators through twitter “i hope it is not as bad as i fear”, 2011.

[53] Z. Zhang. Boosting algorithms explained, Aug 2019. https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30.

[54] C. Ünal and I. Ceasu. A machine learning approach towards startup success prediction.

[55] K. Żbikowski and P. Antosiuk. A machine learning, bias-free approach for predicting business success using crunchbase data.

APPENDICES

Appendix A

TWEET FEATURES

Category	Name	Description
Engagement	retweet_count	Total retweets received
	reply_count	Total replies received
	like_count	Total likes received
	quote_count	Total quotes received
	avg_retweet_count	Avg # of retweets per tweet
	avg_reply_count	Avg # of replies per tweet
	avg_like_count	Avg # of likes per tweet
	avg_quote_count	Avg # of quotes per tweet
Sentiment	avg_vader_sentiment	Avg VADER sentiment score
	avg_bert_sentiment	Avg BERT sentiment score

Category	Name	Description
Distribution	num_tweets	Total # of tweets for the past 9 months, max 5000
	avg_company_tweet	Ratio of tweets from the company
	avg_contains_username	Ratio of tweets containing the company's username
	avg_contains_website	Ratio of tweets containing the company's website
	avg_is_reply	Ratio of tweets replying to company tweet
	avg_is_reference	Ratio of tweets referencing a company tweet
Contents	avg_has_hashtags	Ratio of tweets containing hashtags
	avg_has_mentions	Ratio of tweets containing mentions
	avg_has_links	Ratio of tweets containing links
	avg_has_emoticons	Ratio of tweets containing emojis
	avg_num_hashtags	Avg # of hashtags per tweet
	avg_num_mentions	Avg # of mentions per tweet
	avg_num_links	Avg # of links per tweet
	avg_num_emoticons	Avg # of emojis per tweet

Category	Name	Description
Structure	avg_num_chars	Avg # of characters
	avg_num_words	Avg # of words
	avg_num_punct	Avg # of punctuation
	avg_characters_per_word	Avg # word length
	avg_words_per_sentence	Avg # of words per sentence
	avg_distinct_shape	Avg # of unique word shapes
	avg_shape_ratio	Avg # of unique shapes over total shapes
	avg_total_shape	Avg # of total word shapes
Deep Linguistic	avg_passive_count	Avg # of passive sentences
	avg_total_pos	Avg # of part-of-speech tags
	avg_distinct_pos	Avg # of distinct part-of-speech tags
	avg_pos_ratio	Avg ratio of distinct/total part-of-speech tags
	avg_total_tags	Avg # of Spacy grammar tags
	avg_distinct_tags	Avg # of distinct Spacy grammar tags
	avg_tag_ratio	Avg ratio of distinct/total grammar tags
	avg_syll_per_word	Avg syllables per word
	avg_syllables	Avg syllables per tweet
	avg_type_token_ratio	Avg ratio of unique verbs over total words

Category	Name	Description
Complexity	avg_flesch_reading_ease	Avg tweet readability score
	avg_long_words	Avg # of long words per tweet
	avg_complex_words	Total # of complex words per tweet
	avg_num_synsets	Avg # of synsets per tweet
	avg_synset_ratio	Avg unique synsets over total synsets
	avg_synset_complex	Avg total distance to synset root
	avg_synset_complex_per_word	Avg dist. to synset root per word

Table A.1: Twitter Features

Appendix B

FEATURE PERFORMANCE

Feature	Category	Type	Precision (solo)	Recall (solo)	F1 (solo)	Precision (w/out)	Recall (w/out)	F1 (w/out)	F1 Impact
all	all	all	0.6618	0.8182	0.7317	—	—	—	—
num_founders	general	numeric	0.6106	0.6818	0.6442	0.6256	0.7697	0.6902	0.0415
investor_ratio	funding	numeric	0.5480	0.3894	0.4553	0.6577	0.8152	0.7280	0.0037
num_funding_rounds	funding	numeric	0.5480	0.3894	0.4553	0.6487	0.8227	0.7255	0.0062
months_since_funding	funding	numeric	0.6032	0.3409	0.4356	0.6654	0.8227	0.7358	-0.0041
last_funding_stage	funding	categorical	0.8194	0.2818	0.4194	0.6329	0.7864	0.7014	0.0303
total_investments	funding	numeric	0.5929	0.2803	0.3807	0.6566	0.8227	0.7303	0.0014
distinct_investors	funding	numeric	0.5890	0.2758	0.3756	0.6587	0.8303	0.7346	-0.0029
other_topic	article	numeric	0.4790	0.2591	0.3363	0.6675	0.8242	0.7376	-0.0059
avg_paragraphs	twitter	numeric	0.4680	0.2545	0.3297	0.6578	0.8242	0.7317	-0.0000
avg_total_shape	twitter	numeric	0.4462	0.2576	0.3266	0.6585	0.8152	0.7285	0.0032
avg_total_tags	twitter	numeric	0.4462	0.2576	0.3266	0.6763	0.8197	0.7411	-0.0094
avg_like_count	twitter	numeric	0.5064	0.2409	0.3265	0.6654	0.8197	0.7346	-0.0029
avg_quote_count	twitter	numeric	0.4586	0.2515	0.3249	0.6531	0.8242	0.7287	0.0030
avg_retweet_count	twitter	numeric	0.4393	0.2576	0.3247	0.6630	0.8197	0.7331	-0.0014
num_tweets_format	twitter	numeric	0.4282	0.2576	0.3217	0.6597	0.8076	0.7262	0.0055
avg_sent_per_parag	twitter	numeric	0.4488	0.2455	0.3173	0.6566	0.8258	0.7315	0.0002
avg_total_pos	twitter	numeric	0.4699	0.2364	0.3145	0.6561	0.8182	0.7283	0.0034
avg_characters	twitter	numeric	0.4602	0.2364	0.3123	0.6538	0.8212	0.7280	0.0037
avg_wordtypes	twitter	numeric	0.4572	0.2348	0.3103	0.6622	0.8167	0.7313	0.0004

Feature	Category	Type	Precision (solo)	Recall (solo)	F1 (solo)	Precision (w/out)	Recall (w/out)	F1 (w/out)	F1 Impact
like_count	twitter	numeric	0.4477	0.2333	0.3068	0.6569	0.8182	0.7287	0.0030
avg_has_emoticons	twitter	numeric	0.4551	0.2303	0.3058	0.6578	0.8212	0.7305	0.0012
found_topics	twitter	numeric	0.4817	0.2197	0.3018	0.6577	0.8121	0.7268	0.0049
num_articles	article	numeric	0.5391	0.2091	0.3013	0.6704	0.8167	0.7363	-0.0046
quote_count	twitter	numeric	0.4846	0.2152	0.2980	0.6564	0.8106	0.7254	0.0063
retweet_count	twitter	numeric	0.4362	0.2227	0.2949	0.6663	0.8197	0.7351	-0.0034
num_tweets	twitter	numeric	0.4269	0.2212	0.2914	0.6691	0.8212	0.7374	-0.0057
num_users	twitter	numeric	0.4350	0.2182	0.2906	0.6613	0.8167	0.7308	0.0009
avg_long_words	twitter	numeric	0.4605	0.2121	0.2905	0.6476	0.8212	0.7241	0.0076
avg_distinct_shape	twitter	numeric	0.4505	0.2136	0.2898	0.6646	0.8258	0.7365	-0.0048
avg_passive_count	twitter	numeric	0.4191	0.2197	0.2883	0.6589	0.8167	0.7294	0.0023
avg_is_reply	twitter	numeric	0.4222	0.2136	0.2837	0.6585	0.8121	0.7273	0.0044
avg_num_emoticons	twitter	numeric	0.4000	0.2182	0.2824	0.6654	0.8106	0.7309	0.0008
Science and Engineering	industries	numeric	0.4340	0.2091	0.2822	0.6638	0.8167	0.7323	-0.0006
avg_sentences	twitter	numeric	0.4444	0.2061	0.2816	0.6658	0.8182	0.7342	-0.0025
avg_synset_complex	twitter	numeric	0.4573	0.2030	0.2812	0.6565	0.8167	0.7279	0.0038
avg_is_reference	twitter	numeric	0.3972	0.2167	0.2804	0.6550	0.8197	0.7281	0.0036
avg_num_synsets	twitter	numeric	0.4602	0.2015	0.2803	0.6613	0.8167	0.7308	0.0009
award_topic	twitter	numeric	0.4482	0.2030	0.2795	0.6646	0.8167	0.7328	-0.0011
avg_syllables	twitter	numeric	0.4555	0.2015	0.2794	0.6557	0.8167	0.7274	0.0043
avg_num_words	twitter	numeric	0.4272	0.2045	0.2766	0.6578	0.8212	0.7305	0.0012
avg_words	twitter	numeric	0.4415	0.2000	0.2753	0.6622	0.8167	0.7313	0.0004
top_50_publishers	article	numeric	0.5634	0.1818	0.2749	0.6687	0.8167	0.7353	-0.0036
avg_reply_count	twitter	numeric	0.3977	0.2091	0.2741	0.6622	0.8227	0.7338	-0.0021

Feature	Category	Type	Precision (solo)	Recall (solo)	F1 (solo)	Precision (w/out)	Recall (w/out)	F1 (w/out)	F1 Impact
avg_distinct_tags	twitter	numeric	0.4635	0.1924	0.2719	0.6570	0.8212	0.7300	0.0017
reply_count	twitter	numeric	0.4092	0.2015	0.2701	0.6610	0.8242	0.7336	-0.0019
state	general	categorical	0.3452	0.2212	0.2696	0.6614	0.8197	0.7321	-0.0004
Category 2	general	categorical	0.4565	0.1909	0.2692	0.6589	0.8167	0.7294	0.0023
avg_num_chars	twitter	numeric	0.4549	0.1909	0.2689	0.6642	0.8273	0.7368	-0.0051
avg_num_mentions	twitter	numeric	0.3970	0.2015	0.2673	0.6601	0.8182	0.7307	0.0010
avg_complex_words_dc	twitter	numeric	0.4286	0.1909	0.2642	0.6577	0.8152	0.7280	0.0037
avg_complex_words	twitter	numeric	0.4340	0.1894	0.2637	0.6626	0.8212	0.7334	-0.0017
avg_type_token_ratio	twitter	numeric	0.4116	0.1939	0.2636	0.6562	0.8212	0.7295	0.0022
avg_contains_website	twitter	numeric	0.4118	0.1909	0.2609	0.6618	0.8212	0.7329	-0.0012
funding_topic	twitter	numeric	0.4494	0.1818	0.2589	0.6539	0.8303	0.7316	0.0001
avg_num_syllables	twitter	numeric	0.4207	0.1848	0.2568	0.6638	0.8258	0.7360	-0.0043
avg_synset_complex per_word	twitter	numeric	0.4164	0.1848	0.2560	0.6646	0.8167	0.7328	-0.0011
avg_synset_ratio	twitter	numeric	0.3877	0.1909	0.2558	0.6537	0.8121	0.7243	0.0074
avg_num_hashtags	twitter	numeric	0.4470	0.1788	0.2554	0.6574	0.8258	0.7320	-0.0003
avg_chars_per_word	twitter	numeric	0.4152	0.1818	0.2529	0.6606	0.8227	0.7328	-0.0011
avg_has_mentions	twitter	numeric	0.3799	0.1894	0.2528	0.6626	0.8273	0.7358	-0.0041
avg_other_user	twitter	numeric	0.3961	0.1848	0.2521	0.6606	0.8227	0.7328	-0.0011
avg_has_hashtags	twitter	numeric	0.4212	0.1742	0.2465	0.6570	0.8212	0.7300	0.0017
english_vader	twitter	numeric	0.4034	0.1773	0.2463	0.6553	0.8152	0.7265	0.0052
Category 4	general	categorical	0.4561	0.1652	0.2425	0.6634	0.8212	0.7339	-0.0022
mgmnt_change_topic	twitter	numeric	0.4862	0.1606	0.2415	0.6537	0.8121	0.7243	0.0074
avg_has_links	twitter	numeric	0.3898	0.1742	0.2408	0.6650	0.8212	0.7349	-0.0032
avg_words_per_sent	twitter	numeric	0.4413	0.1652	0.2404	0.6527	0.8258	0.7291	0.0026

Feature	Category	Type	Precision (solo)	Recall (solo)	F1 (solo)	Precision (w/out)	Recall (w/out)	F1 (w/out)	F1 Impact
user_vader	twitter	numeric	0.3904	0.1727	0.2395	0.6634	0.8121	0.7302	0.0015
new_product_topic	twitter	numeric	0.5124	0.1561	0.2393	0.6634	0.8212	0.7339	-0.0022
ca_hq	general	numeric	0.3380	0.1848	0.2390	0.6712	0.8197	0.7381	-0.0064
avg_distinct_pos	twitter	numeric	0.4225	0.1652	0.2375	0.6547	0.8273	0.7309	0.0008
Category 3	general	categorical	0.4320	0.1636	0.2374	0.6558	0.8258	0.7311	0.0006
geo_expansion_topic	twitter	numeric	0.4725	0.1561	0.2346	0.6642	0.8182	0.7332	-0.0015
num_languages	twitter	numeric	0.4106	0.1636	0.2340	0.6650	0.8121	0.7312	0.0005
avg_vader_sentiment	twitter	numeric	0.3452	0.1758	0.2329	0.6538	0.8212	0.7280	0.0037
avg_syll_per_word	twitter	numeric	0.3776	0.1682	0.2327	0.6598	0.8227	0.7323	-0.0006
top_10_publishers	article	numeric	0.5975	0.1439	0.2320	0.6700	0.8242	0.7391	-0.0074
avg_pos_ratio	twitter	numeric	0.3899	0.1636	0.2305	0.6581	0.8167	0.7289	0.0028
avg_num_links	twitter	numeric	0.4244	0.1530	0.2249	0.6708	0.8212	0.7384	-0.0067
avg_bert_sentiment	twitter	numeric	0.3977	0.1561	0.2242	0.6659	0.8303	0.7390	-0.0073
contains_website	twitter	numeric	0.3962	0.1561	0.2239	0.6565	0.8167	0.7279	0.0038
name_length	general	numeric	0.4213	0.1500	0.2212	0.6548	0.8106	0.7244	0.0073
avg_english	twitter	numeric	0.3723	0.1545	0.2184	0.6611	0.8333	0.7373	-0.0056
avg_num_punct	twitter	numeric	0.3992	0.1500	0.2181	0.6634	0.8273	0.7363	-0.0046
english_bert	twitter	numeric	0.4000	0.1485	0.2166	0.6610	0.8242	0.7336	-0.0019
avg_syllables_word	twitter	numeric	0.3579	0.1545	0.2159	0.6590	0.8288	0.7342	-0.0025
funding_topic	article	numeric	0.6897	0.1212	0.2062	0.6562	0.8212	0.7295	0.0022
contains_username	twitter	numeric	0.4485	0.1318	0.2037	0.6626	0.8152	0.7310	0.0007
avg_shape_ratio	twitter	numeric	0.3689	0.1364	0.1991	0.6601	0.8182	0.7307	0.0010
description	general	categorical	0.4067	0.1288	0.1956	0.6912	0.7530	0.7208	0.0109
Artificial Intelligence	industries	numeric	0.5232	0.1197	0.1948	0.6617	0.8152	0.7305	0.0012
avg_has_username	twitter	numeric	0.3836	0.1273	0.1911	0.6618	0.8212	0.7329	-0.0012

Feature	Category	Type	Precision (solo)	Recall (solo)	F1 (solo)	Precision (w/out)	Recall (w/out)	F1 (w/out)	F1 Impact
avg_flesch_reading_ease	twitter	numeric	0.3584	0.1227	0.1828	0.6594	0.8242	0.7327	-0.0010
avg_tag_ratio	twitter	numeric	0.3348	0.1167	0.1730	0.6671	0.8227	0.7368	-0.0051
user_bert	twitter	numeric	0.3226	0.1061	0.1596	0.6519	0.8227	0.7274	0.0043
Category 6	general	categorical	0.4014	0.0864	0.1421	0.6750	0.8182	0.7397	-0.0080
Category 5	general	categorical	0.5521	0.0803	0.1402	0.6602	0.8212	0.7319	-0.0002
top_10_article_count	google	numeric	0.3750	0.0773	0.1281	0.6562	0.8273	0.7319	-0.0002
m&a_topic	twitter	numeric	0.4375	0.0636	0.1111	0.6687	0.8227	0.7378	-0.0061
Biotechnology	industries	numeric	0.3878	0.0576	0.1003	0.6593	0.8182	0.7302	0.0015
num_categories	general	numeric	0.4459	0.0500	0.0899	0.6544	0.8318	0.7325	-0.0008
top_50_article_count	google	numeric	0.4844	0.0470	0.0856	0.6510	0.8197	0.7257	0.0060
Payments	industries	numeric	0.4182	0.0348	0.0643	0.6577	0.8182	0.7292	0.0025
avg_company_tweet	twitter	numeric	0.6774	0.0318	0.0608	0.6646	0.8197	0.7341	-0.0024
new_product_topic	article	numeric	0.5263	0.0303	0.0573	0.6609	0.8152	0.7300	0.0017
geo_expansion_topic	article	numeric	0.8182	0.0273	0.0528	0.6598	0.8227	0.7323	-0.0006
Agriculture and Farming	industries	numeric	0.4500	0.0273	0.0514	0.6609	0.8121	0.7288	0.0029
award_topic	article	numeric	0.7083	0.0258	0.0497	0.6671	0.8258	0.7380	-0.0063
num_category_groups	general	numeric	0.4211	0.0242	0.0458	0.6550	0.8258	0.7306	0.0011
company_bert	twitter	numeric	0.7143	0.0227	0.0441	0.6581	0.8167	0.7289	0.0028
company_vader	twitter	numeric	0.5357	0.0227	0.0436	0.6601	0.8182	0.7307	0.0010
mgmnt_change_topic	article	numeric	0.7778	0.0212	0.0413	0.6558	0.8258	0.7311	0.0006
months_since_founding	general	numeric	0.4333	0.0197	0.0377	0.6398	0.8045	0.7128	0.0189
num_company_results	google	numeric	0.2273	0.0076	0.0147	0.6535	0.8288	0.7308	0.0009
own_company_ratio	google	numeric	0.2273	0.0076	0.0147	0.6618	0.8182	0.7317	-0.0000
top_10_google_count	google	numeric	0.2500	0.0030	0.0060	0.6541	0.8167	0.7264	0.0053

Feature	Category	Type	Precision (solo)	Recall (solo)	F1 (solo)	Precision (w/out)	Recall (w/out)	F1 (w/out)	F1 Impact
top_50_google_count	google	numeric	0.2500	0.0015	0.0030	0.6507	0.8242	0.7273	0.0044
m&a_topic	article	numeric	0.2000	0.0015	0.0030	0.6589	0.8167	0.7294	0.0023
Content and Publishing	industries	numeric	0.0000	0.0000	0.0000	0.6582	0.8227	0.7313	0.0004
Messaging and Telecommunications	industries	numeric	0.0000	0.0000	0.0000	0.6550	0.8227	0.7293	0.0024
Energy	industries	numeric	0.0000	0.0000	0.0000	0.6561	0.8152	0.7270	0.0047
Lending and Investments	industries	numeric	0.0000	0.0000	0.0000	0.6691	0.8182	0.7362	-0.0045
Sports	industries	numeric	0.0000	0.0000	0.0000	0.6691	0.8182	0.7362	-0.0045
Clothing and Apparel	industries	numeric	0.0000	0.0000	0.0000	0.6622	0.8197	0.7326	-0.0009
Gaming	industries	numeric	0.0000	0.0000	0.0000	0.6597	0.8167	0.7299	0.0018
Sustainability	industries	numeric	0.0000	0.0000	0.0000	0.6577	0.8152	0.7280	0.0037
Natural Resources	industries	numeric	0.0000	0.0000	0.0000	0.6613	0.8167	0.7308	0.0009
Travel and Tourism	industries	numeric	0.0000	0.0000	0.0000	0.6671	0.8288	0.7392	-0.0075
Video	industries	numeric	0.0000	0.0000	0.0000	0.6610	0.8182	0.7312	0.0005
Events	industries	numeric	0.0000	0.0000	0.0000	0.6638	0.8197	0.7336	-0.0019
Music and Audio	industries	numeric	0.0000	0.0000	0.0000	0.6550	0.8227	0.7293	0.0024
Government and Military	industries	numeric	0.0000	0.0000	0.0000	0.6642	0.8182	0.7332	-0.0015
Platforms	industries	numeric	0.0000	0.0000	0.0000	0.6663	0.8167	0.7338	-0.0021
Consumer Goods	industries	numeric	0.0000	0.0000	0.0000	0.6622	0.8227	0.7338	-0.0021
Navigation and Mapping	industries	numeric	0.0000	0.0000	0.0000	0.6658	0.8212	0.7354	-0.0037
Category 1	general	categorical	0.0000	0.0000	0.0000	0.6585	0.8152	0.7285	0.0032

Feature	Category	Type	Precision (solo)	Recall (solo)	F1 (solo)	Precision (w/out)	Recall (w/out)	F1 (w/out)	F1 Impact
website_length	general	numeric	0.0000	0.0000	0.0000	0.6654	0.8227	0.7358	-0.0041
ny_hq	general	numeric	0.0000	0.0000	0.0000	0.6622	0.8167	0.7313	0.0004
tx_hq	general	numeric	0.0000	0.0000	0.0000	0.6609	0.8152	0.7300	0.0017
other_hub_hq	general	numeric	0.0000	0.0000	0.0000	0.6562	0.8242	0.7307	0.0010
country	general	categorical	0.0000	0.0000	0.0000	0.6612	0.7924	0.7209	0.0108
has_facebook	general	numeric	0.0000	0.0000	0.0000	0.6613	0.8167	0.7308	0.0009
total_search_results	google	numeric	0.0000	0.0000	0.0000	0.6602	0.8212	0.7319	-0.0002
Privacy and Security	industries	numeric	0.0000	0.0000	0.0000	0.6626	0.8182	0.7322	-0.0005
Consumer Electronics	industries	numeric	0.0000	0.0000	0.0000	0.6654	0.8106	0.7309	0.0008
num_other_sources	google	numeric	0.0000	0.0000	0.0000	0.6663	0.8288	0.7387	-0.0070
Software	industries	numeric	0.0000	0.0000	0.0000	0.6634	0.8212	0.7339	-0.0022
Information Technology	industries	numeric	0.0000	0.0000	0.0000	0.6589	0.8136	0.7281	0.0036
Health Care	industries	numeric	0.0000	0.0000	0.0000	0.6601	0.8182	0.7307	0.0010
Internet Services	industries	numeric	0.0000	0.0000	0.0000	0.6502	0.8167	0.7240	0.0077
Data and Analytics	industries	numeric	0.0000	0.0000	0.0000	0.6590	0.8227	0.7318	-0.0001
Financial Services	industries	numeric	0.0000	0.0000	0.0000	0.6590	0.8197	0.7306	0.0011
Other	industries	numeric	0.0000	0.0000	0.0000	0.6618	0.8242	0.7341	-0.0024
Commerce and Shopping	industries	numeric	0.0000	0.0000	0.0000	0.6553	0.8182	0.7278	0.0039
Hardware	industries	numeric	0.0000	0.0000	0.0000	0.6557	0.8167	0.7274	0.0043
Professional Services	industries	numeric	0.0000	0.0000	0.0000	0.6511	0.8258	0.7281	0.0036
Sales and Marketing	industries	numeric	0.0000	0.0000	0.0000	0.6642	0.8212	0.7344	-0.0027
Media and Entertainment	industries	numeric	0.0000	0.0000	0.0000	0.6654	0.8167	0.7333	-0.0016

Feature	Category	Type	Precision (solo)	Recall (solo)	F1 (solo)	Precision (w/out)	Recall (w/out)	F1 (w/out)	F1 Impact
has_linkedin	general	numeric	0.0000	0.0000	0.0000	0.6558	0.8197	0.7286	0.0031
Apps	industries	numeric	0.0000	0.0000	0.0000	0.6741	0.8242	0.7416	-0.0099
Design	industries	numeric	0.0000	0.0000	0.0000	0.6667	0.8212	0.7359	-0.0042
Mobile	industries	numeric	0.0000	0.0000	0.0000	0.6667	0.8182	0.7347	-0.0030
has_twitter	general	numeric	0.0000	0.0000	0.0000	0.6667	0.8212	0.7359	-0.0042
Manufacturing	industries	numeric	0.0000	0.0000	0.0000	0.6716	0.8212	0.7389	-0.0072
Transportation	industries	numeric	0.0000	0.0000	0.0000	0.6622	0.8227	0.7338	-0.0021
Community and Lifestyle	industries	numeric	0.0000	0.0000	0.0000	0.6634	0.8273	0.7363	-0.0046
Real Estate	industries	numeric	0.0000	0.0000	0.0000	0.6590	0.8227	0.7318	-0.0001
Advertising	industries	numeric	0.0000	0.0000	0.0000	0.6679	0.8167	0.7348	-0.0031
Education	industries	numeric	0.0000	0.0000	0.0000	0.6634	0.8182	0.7327	-0.0010
Administrative Services	industries	numeric	0.0000	0.0000	0.0000	0.6630	0.8227	0.7343	-0.0026
Food and Beverage	industries	numeric	0.0000	0.0000	0.0000	0.6630	0.8288	0.7367	-0.0050

Table B.1: Complete Feature Set

Appendix C

PREDICTING ADDITIONAL FUNDING ROUNDS

Cutoff Threshold	Precision	Recall	F_1	$F_{0.1}$
0.50	0.6559	**0.5351**	**0.5894**	0.6545
0.55	0.7042	0.4386	0.5405	0.7000
0.60	0.7143	0.3947	0.5085	0.7086
0.65	0.7115	0.3246	0.4458	0.7032
0.70	0.7174	0.2895	0.4125	0.7070
0.75	0.7436	0.2544	0.3791	0.7297
0.80	0.7500	0.1842	0.2958	0.7279
0.85	0.7778	0.1228	0.2121	0.7388
0.90	**1.0000**	0.0614	0.1157	**0.8686**
0.95	**1.0000**	0.0088	0.0174	0.4720

Table C.1: Received Funding by Cutoff Threshold

Appendix D

FEATURE ANALYSIS

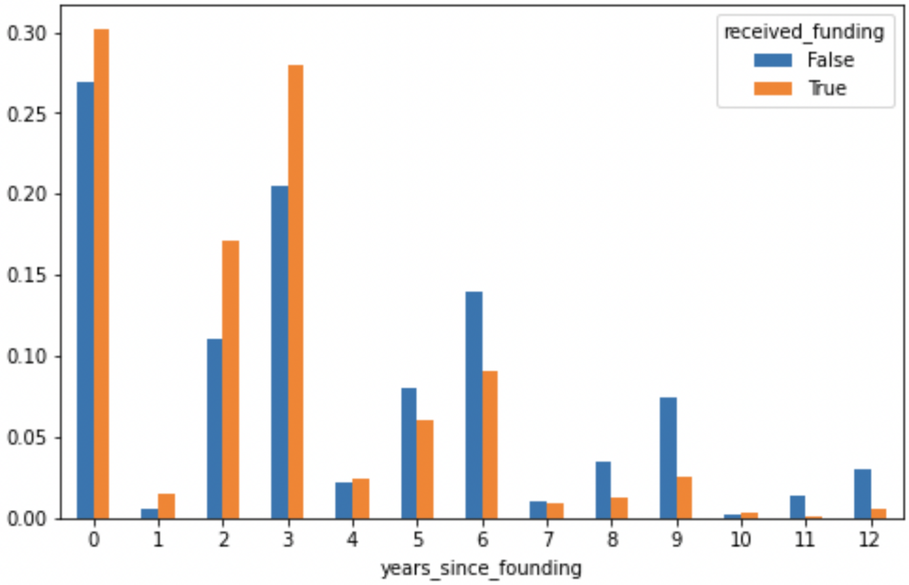


Figure D.1: Years Since Founding

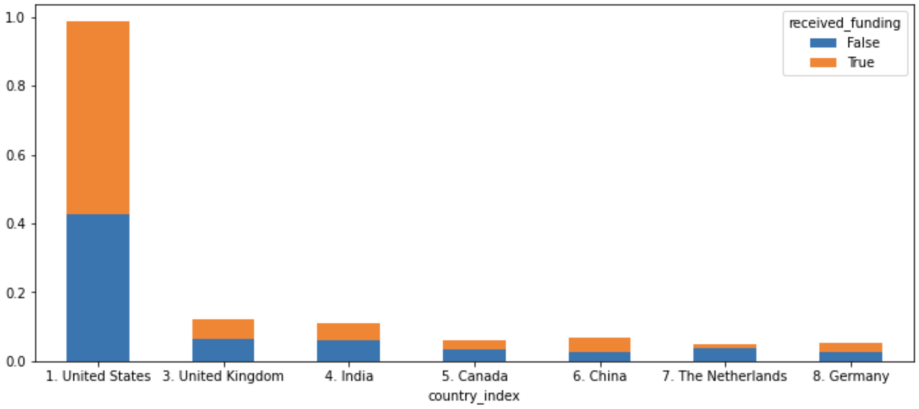


Figure D.2: Top Countries

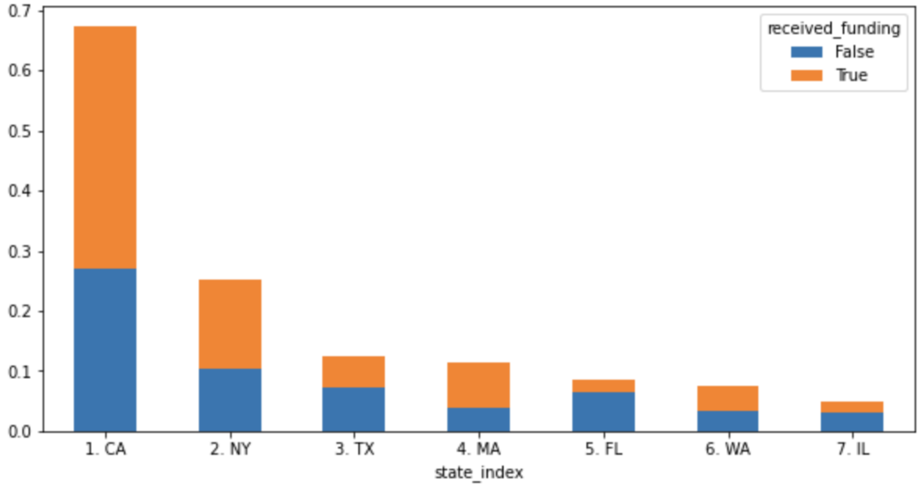
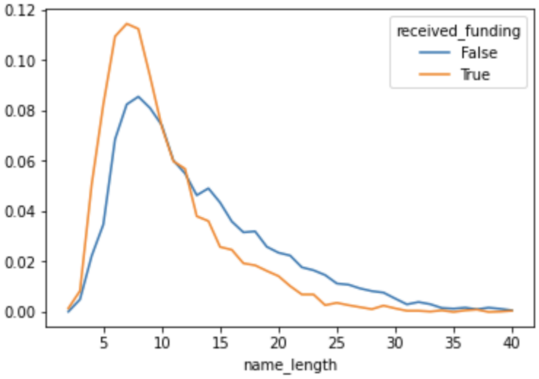


Figure D.3: Top States



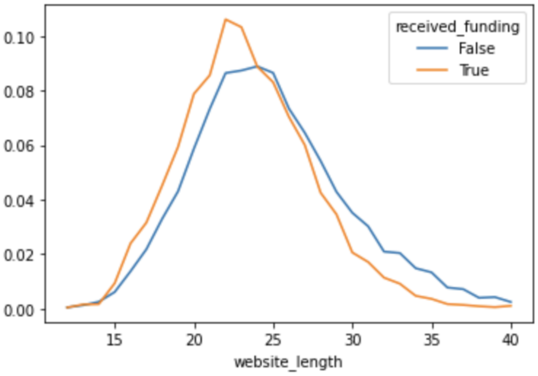


Figure D.4: Name Length

Figure D.5: Website Length

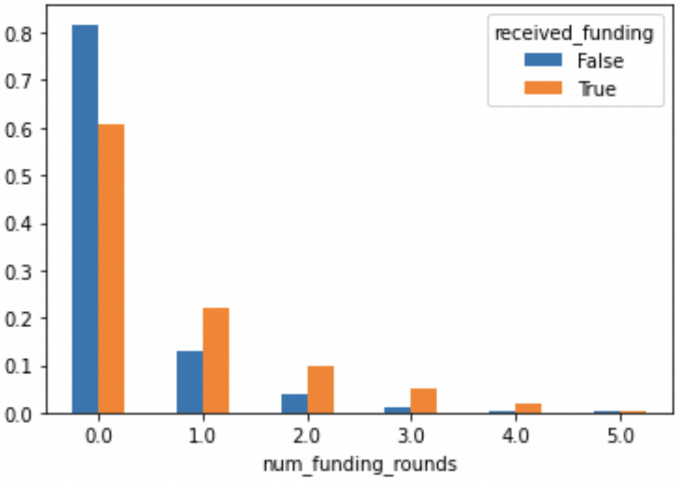


Figure D.6: Funding Rounds

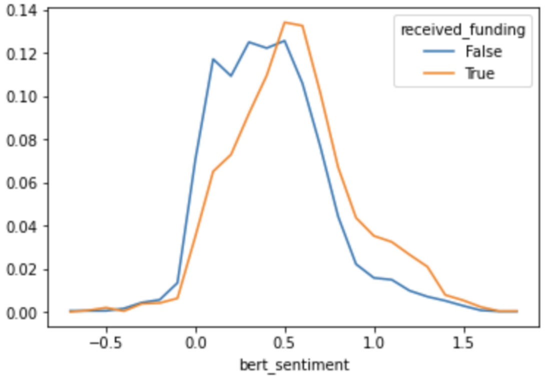


Figure D.7: Overall Sentiment
Bert

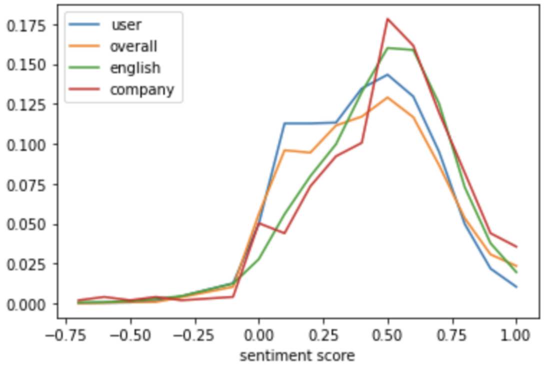


Figure D.8: Sentiment Comparison
Bert

Appendix E

INDUSTRY CATEGORY GROUPS

Category	Count
Software	742
Information Technology	349
Health Care	296
Science and Engineering	293
Internet Services	261
Data and Analytics	258
Financial Services	228
Other	221
Commerce and Shopping	216
Hardware	204
Professional Services	194
Sales and Marketing	194
Media and Entertainment	172
Artificial Intelligence	148
Apps	121
Design	120

Category	Count
Mobile	117
Biotechnology	105
Manufacturing	97
Transportation	95
Community and Lifestyle	93
Real Estate	91
Advertising	86
Education	78
Food and Beverage	78
Administrative Services	76
Consumer Electronics	73
Privacy and Security	71
Consumer Goods	63
Payments	59
Sustainability	56
Content and Publishing	53
Energy	50
Lending and Investments	50
Sports	47
Clothing and Apparel	44
Gaming	39

Category	Count
Natural Resources	33
Travel and Tourism	32
Video	30
Messaging and Telecommunications	28
Music and Audio	22
Government and Military	20
Platforms	20
Agriculture and Farming	18
Events	18
Navigation and Mapping	13

Table E.1: Industry Category Groups