Media Engineering and Technology Faculty
German University in Cairo

Engineering a Data Benchmark from Textual Posts for Entrepreneurial Personality Analysis

Bachelor Thesis

Author: Joy Emad Kamel Labib

Supervisors: Dr. Mervat Mustafa Fahmy Abuelkheir

Dr. Nourhan Ehab Abdelhamid Azab

Submission Date: 2 May, 2024

Media Engineering and Technology Faculty
German University in Cairo

Engineering a Data Benchmark from Textual Posts for Entrepreneurial Personality Analysis

Bachelor Thesis

Author: Joy Emad Kamel Labib

Supervisors: Dr. Mervat Mustafa Fahmy Abuelkheir

Dr. Nourhan Ehab Abdelhamid Azab

Submission Date: 2 May, 2024

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowledgment has been made in the text to all other material used

---

Joy Emad Kamel Labib
2 May, 2024

Acknowledgments

Acknowledgment goes here.

Abstract

Abstract text goes here.

Contents

Chapter 1

Introduction

In a fast-paced changing world, where everyone strives to enter the business market and be updated with the latest trends in business, only few people take the risk and start their businesses. Those are called to have an entrepreneurial soul, which permits a person's innovation to have a great impact on the community around them, as well as the economy of their country.

Defining the word entrepreneur is as simple as describing an instinctive personality filled with identifying opportunities, taking calculated risks, and mobilizing resources. This results in the creation of innovative solutions, products, services, or businesses. Entrepreneurs are characterized by their vision, creativity, resilience, and willingness to positively influence everyone around them.

Since entrepreneurs play a crucial role in driving economic growth and shaping industries through their ability to transform their ideas into real businesses, identifying an entrepreneurial personality becomes more important every day to the economy. It has always been a challenge to recognize the different personalities that exist and define the traits of each one. The entrepreneurial personality is still a complex one, which is linked to a group of specific traits describing the behavior and explaining the decisions they make.

Entrepreneurial personality recognition is currently the concern of various discussions, whether we can classify persons who have an entrepreneurial personality or not. This concerns a lot of businessmen who are willing to invest in small businesses when they don't know whether this person deserves their investment. Additionally, many companies would benefit from such recognition in the recruitment process and selecting the right employee for their industry.

1.1 Motivation

Recognizing a personality depends on identifying the existence of some specific traits with a certain percentage. This inquiry shows these traits through the lifestyle of the person,

starting from their daily decisions, their textual posts, the way they communicate with others, and how they approach showing on social media.

As Artificial Intelligence now shaped our usage of technology, it was based on the enormous collected data from diverse resources. This made it possible to train machine learning models that can come up with a clear classification of whether this person has an entrepreneurial personality or not. Unfortunately, no clear standardized dataset was found to train the models, therefore, the classification does not exist yet.

The ability to use text as a main source for this classification, as well as the Artificial Intelligence Analysis would make this classification and help the society to discover the entrepreneurial personalities living within it. This is the main motivation to create such a data benchmark to bring the classification into reality.

1.2 Objectives

The aim of this project is to engineer a data benchmark that can be used to analyze what makes people successful entrepreneurs. This will be done through these objectives:

1. Investigate diverse sources of textual posts written by entrepreneurs, whether their personal way of expressing themselves in their daily life or the formal way of sharing their experience in life.

2. Extract the textual posts of the entrepreneurs for these sources, accompanied with their meta-data.

3. Design a clear data frame to store all the collected data, along with enriching the data with the textual features for each record.

4. Define a clear pipeline for textual data cleaning from any inconsistencies that could interfere with the classification.

5. Analyze the extracted textual features to reach some validating information of the characteristics that entrepreneurs have in common.

6. Deploy the dataset to be accessible for any consumer working on the analysis of the entrepreneurial personality using a classification of text.

In pursuit of these objectives, our thesis aims not only to contribute to the scholarly discourse surrounding entrepreneurship but also to offer practical insights that may inform and empower aspiring entrepreneurs on their journey to success.

1.3 Outline

This Thesis consists of 5 chapters including the "Introduction". The second chapter includes the background which explains the entrepreneurial personality and its traits, and examines the role of textual data in personality analysis. It also states the different methodologies used in previous studies and identifying the gaps in current research. The third chapter is the methodology that demonstrates the approach of textual data collection from different sources, the selection criteria, the description of the data processing steps, the labeling of the data, and finally the validation of the dataset. Chapter four represents the findings from the evaluation of the benchmark data set, and the interpretation in the context of existing literature on entrepreneurial personality analysis. As well as, a discussion of the limitations of the study and opportunities for future research. The thesis in concluded in chapter 5 that summarizes the study's contributions to the field of entrepreneurial personality analysis, and a reflection on the importance of standardized benchmarks. It also suggests future directions in the research and implementation of the benchmark.

Chapter 2

Background

This chapter includes the background research that founded this bachelor thesis. It is divided into 5 main categories: The Human Personality, The Entrepreneurial Personality, Prediction of the Entrepreneurial Personality, Datasets and lastly, Applications of a Data Benchmark that serves as a high-level introduction to this research.

2.1 The Human Personality

A personality refers to the unique set of individual traits, behaviors, attitudes, and characteristics that distinguish one person from another [14]. It encompasses the way a person thinks, feels, and behaves across various situations and contexts . Personality is believed to be relatively stable over time but can also be influenced by environmental factors [19], experiences, and personal development. Psychologists often study personality to understand how it shapes an individual's thoughts, emotions, motivations, and interactions with others. Personality is closely linked to a broad spectrum of human actions and circumstances, including shopping habits, well-being, social connections, and potentially criminal activities.

The fundamental principle of personality psychology is that stable individual traits lead to consistent behavioral patterns that people tend to show, regardless of the circumstances [36]. Trait models in psychology are constructed based on how people perceive similarities and connections between descriptive words they use to characterize themselves and others. Despite the abundance and diversity of these terms, they generally are reduced to only a few major dimensions [36].

There are several personality models used in predicting personality, such as Big Five Personality, Myers-Briggs Type Indicator [14] (MBTI) or Dominance Influence Steadiness Conscientiousness[14] (DISC). However, after some considerations and literature review process, Big Five Personality is the most popular and precise in telling someone's personality traits [14], [35]. The model describes human personality by five traits/factors,

popularly referred to as the Big Five or OCEAN: Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These traits are today's dominant paradigm in personality research, and one of the most influential models in all of psychology [22].

The Big Five traits are as follows [5], [3]:

1. *Openness to Experience:*
   Individuals high in openness tend to have broad interests and are often imaginative, creative, and intellectually curious, and may be more receptive to change. However, they may also be prone to overthinking, indecision, and distraction.

2. *Conscientiousness:*
   Conscientious individuals are characterized by their strong sense of responsibility, self-discipline, and reliability, and also tend to be organized, and detail-oriented. Though, they may be perceived as rigid, perfectionist, and overly cautious.

3. *Extraversion:*
   Extraverts are sociable, outgoing individuals and derive energy from interacting with others. They tend to be assertive, and often taking on leadership roles [36]. However, they may also experience struggle with introspection, and be prone to risk-taking behaviors.

4. *Agreeableness:*
   Agreeable individuals are compassionate, empathetic, and cooperative. They tend to be trusting, forgiving and considerate of others' feelings and needs. However, they may also be vulnerable to exploitation, overly trusting.

5. *Neuroticism:*
   Neuroticism reflects the tendency to experience negative emotions such as anxiety, depression, and moodiness [36]. Individuals high in neuroticism may be prone to worry, and emotional volatility, struggling with feelings of insecurity, self-doubt, and fear of failure.

2.2 The Entrepreneurial Personality

Moving on to a more specific type of personalities that is the focus of this study, which is the entrepreneurial personality. The term "entrepreneur" originates from the French word "entreprendre" and the German word "unternehmen," both of which convey the concept of "undertaking." As per the Webster dictionary, an "entrepreneur" is someone who organizes, manages, and takes on risks of a business or enterprise [32]. Enterprising behaviour plays an important role in the modern economy which is characterized by instability and rapid change, obliging people and organizations to be in a process of constant innovation [27].

A person may be enterprising at the personal level (personal entrepreneur), which involves demonstrating a strong sense of control and the ability to manage difficult situations and steer their own life [11]. There are two enterprising personalities we can mention, The intra-entrepreneurs and the extra-entrepreneurs. The intra-entrepreneur refers to people who produce changes and innovation within their positions in a company, adding something creative in their projects that are already in progress [23]. The extra-entrepreneur is a person whose goal is the development of new external projects and businesses creation [27]. The focus of this thesis will be the extra-entrepreneur, or as known as the "general entrepreneur", a person who chooses to work for themselves, not for others.

Additionally, as there are well known standardized traits that measures the different dimensions of a personality, there are specific traits that provides a more precise description of how entrepreneurs and non-entrepreneurs differ in specific behavioral dimensions [27]. These traits are measured using the Measure of Entrepreneurial Tendencies and Abilities Measure of Entrepreneurial Tendencies and Abilities [27] (META), which demonstrated that there is a stronger evidence of predictive accuracy with these traits than the Big Five personality traits. META is a state-of-the-art psychometric test that identifies entrepreneurial potential in order to help businesses nurture and retain their entrepreneurial talent. It assesses four aspects of entrepreneurial personality, namely, entrepreneurial awareness, entrepreneurial creativity, opportunism, and vision [1].

Furthermore, after the development of the META, it was discovered that this measuring instrument has focused on a specific trait in an entrepreneur's personality, so there are no comprehensive, exhaustive, systematic analyses of entrepreneurial personality. Then, new eight dimensions of entrepreneurial personality, for young people and adolescents was developed to include all different aspects [33]. The eight specific traits of an entrepreneurial personality are:

1. *Self-efficacy:* It refers to a person's ability to organize and carry out actions effectively, and their persistence when encountering obstacles.
2. *Autonomy:* It refers to the motivation for entrepreneurial creation to achieve a certain individual freedom.
3. *Innovativeness:* It is the will and interest in finding new ways to do usual daily tasks.
4. *Internal locus of control:* It is about taking responsibility and attributing the consequences of one's actions to their causes.
5. *Achievement motivation:* It is the desire to achieve standards of excellence.
6. *Optimism:* It refers to the beliefs a person has about good things happening more than bad things in their life.
7. *Stress tolerance:* It can be defined as the resistance to perceiving environmental stimuli as stressors due to the appropriate coping strategies.

8. *Risk-taking:* It is people's tendency and will to assume certain levels of risk or change to achieve an objective.

Figure 2.1 demonstrates the Big Five Personalities and how can we relate them to the eight specific entrepreneurial traits. As each trait of the Big Five traits can be an umbrella for multiple specific traits.

Openness	Conscientious	Extroversion	Agreeableness	Neuroticism
Imaginative, Wide interest, Curious, Intelligent, Artistic, Unconventional	Organized, Disciplined, Planner, Goal oriented, not impulsive	Energetic, Forceful, Adventurous, Enthusiastic	Sympathetic, Straight forward, Compliance, Generous	Anxious, Tense, Worried, irritable, impulsive, shy

Figure 2.1: Big Five Personality Traits[33]

2.3 Prediction of the Entrepreneurial Personality

In recent years, there has been a significant evolution in how we approach the analysis and prediction of individuals' personalities. Traditional methods heavily relied on standardized questionnaires and self-report measures. Questionnaires where people rate their own behavior with Likert scales are the instrument most commonly adopted for such a purpose [36]. The most popular questionnaires for predicting the general personality traits include the NEO-Personality-Inventory Revised (NEO-PI-R, 240 items), the NEO Five Factor Inventory (NEO-FFI, 60 items), and the Big-Five Inventory (BFI, 44 items). Short questionnaires (5-10 items), much faster to fill, were built by retaining only those items that best correlate with the results of the full instruments [36].

Further, for the eight specific entrepreneurial personality traits, the Battery for the assessment of the enterprising personality (BEPE) was developed. The items making up the battery follow a Likert-type format with five answer categories (1 totally disagree, 5 totally agree), in line with established psychometric literature which indicates that between four and six answer categories produce better psychometric indicators [8]. The main limitation of self-assessments is that the subjects might tend to bias the ratings towards socially desirable characteristics, especially when the assessment can have negative consequences [36].

However, with the rise of technological advancements and social media platforms, there has been a notable shift towards extracting more data from naturalistic data sources. Several works investigate the interplay between personality and computing by measuring the link between traits and use of technology. Online social networks like Twitter, Google+ and Facebook contain much information that can potentially reveal many traits, preferences and opinions of the profile owner. This resulted in research on personal analytic – automatically inferring such latent author attributes in social media [9], [35], [37], [36].

Given its central importance in capturing the essential aspects of human life, increasing attention is being paid to the development of models that can use behavioral data to automatically predict personality. Affective computing focuses on introduces novel techniques that develop and apply affective reasoning tools for personality prediction in multiple modalities and different languages [9]. They infer personality traits from audio, static image, video, or audio visual clip recorded from various scenarios, such as dyadic dialogue, self-evaluating surveys and self-interviews [19]. Data obtained from verbal behavior is one of the key types of such data. Even in the early years of psychology, a person's use of language was seen as a distillation of his or her underlying drives, emotions, and thought patterns [14].

With the aid of machine learning models, researchers and practitioners can extract valuable insights from unstructured textual data, such as emails, social media posts, and online forums. By analyzing language patterns, word choices, and linguistic styles, these models can discern underlying personality traits with a higher degree of accuracy than traditional questionnaires [14], [37], [2]. These measurements capture the within-text distributions of scores for a given psycholinguistic feature, referred as "text contours". Figure 2.2 lists all features with their examples. There are four main groups for these features [14], [35], [36], [9]:

1. *Features of morpho-syntactic complexity:* This group includes surface features such as the average length of clauses and sentences, the features of the type and frequency of embeddings like the number of dependent clauses or the verb phrases, and finally the frequency of particular structure types.

2. *Features of lexical richness, diversity and sophisticated:* This group includes the lexical density features, like the ratio of lexical words to the total number of words in a text. Also, it includes the lexical variation such as the vocabulary used, the lexical sophistication, for the unusual or advanced words in a text, the psycholinguistic norms of words, and lastly, how much people know this word.

3. *Readability features:* This group combines a word familiarity variable from a prespecified vocabulary resource, along with a syntactic variable, such as average sentence length.

4. *Lexicon features:* This last group is derived from a total of ten lexicons that have been successfully used in personality detection, emotion recognition and sentiment analysis research, such as The Affective Norms for English Words (ANEW), DepecheMood++, The Linguistic Inquiry Word Count [26] (LIWC) dictionary, etc [26].

LIWC FEATURES (Pennebaker et al., 2001):

· **Standard counts:**
- Word count (WC), words per sentence (WPS), type/token ratio (Unique), words captured (Dic), words longer than 6 letters (Sixltr), negations (Negate), assents (Assent), articles (Article), prepositions (Preps), numbers (Number)
- Pronouns (Pronoun): 1st person singular (I), 1st person plural (We), total 1st person (Self), total 2nd person (You), total 3rd person (Other)

· **Psychological processes:**
- Affective or emotional processes (Affect): positive emotions (Posemo), positive feelings (Posfeel), optimism and energy (Optim), negative emotions (Negemo), anxiety or fear (Anx), anger (Anger), sadness (Sad)
- Cognitive Processes (Cogmech): causation (Cause), insight (Insight), discrepancy (Discrep), inhibition (Inhib), tentative (Tentat), certainty (Certain)
- Sensory and perceptual processes (Senses): seeing (See), hearing (Hear), feeling (Feel)
- Social processes (Social): communication (Comm), other references to people (Othref), friends (Friends), family (Family), humans (Humans)

· **Relativity:**
- Time (Time), past tense verb (Past), present tense verb (Present), future tense verb (Future)
- Space (Space): up (Up), down (Down), inclusive (Incl), exclusive (Excl)
- Motion (Motion)

· **Personal concerns:**
- Occupation (Occup): school (School), work and job (Job), achievement (Achieve)
- Leisure activity (Leisure): home (Home), sports (Sports), television and movies (TV), music (Music)
- Money and financial issues (Money)
- Metaphysical issues (Metaph): religion (Relig), death (Death), physical states and functions (Physcal), body states and symptoms (Body), sexuality (Sexual), eating and drinking (Eating), sleeping (Sleep), Grooming (Groom)

· **Other dimensions:**
- Punctuation (Allpct): period (Period), comma (Comma), colon (Colon), semi-colon (Semic), question (Qmark), exclamation (Exclam), dash (Dash), quote (Quote), apostrophe (Apostro), parenthesis (Parenth), other (Otherp)
- Swear words (Swear), nonfluencies (Nonfl), fillers (Fillers)

MRC FEATURES (Coltheart, 1981):

Number of letters (Nlet), phonemes (Nphon), syllables (Nsyl), Kucera-Francis written frequency (K-F-freq), Kucera-Francis number of categories (K-F-ncats), Kucera-Francis number of samples (K-F-nsamp), Thorndike-Lorge written frequency (T-L-freq), Brown verbal frequency (Brown-freq), familiarity rating (Fam), concreteness rating (Conc), imageability rating (Imag), meaningfulness Colorado Norms (Meanc), meaningfulness Paivio Norms (Meanp), age of acquisition (AOA)

UTTERANCE TYPE FEATURES:

Ratio of commands (Command), prompts or back-channels (Prompt), questions (Question), assertions (Assertion)

PROSODIC FEATURES:

Average, minimum, maximum and standard deviation of the voice's pitch in Hz (Pitch-mean, Pitch-min, Pitch-max, Pitch-stddev) and intensity in dB (Int-mean, Int-min, Int-max, Int-stddev), voiced time (Voiced) and speech rate (Word-per-sec)

Figure 2.2: Description of all features, with feature labels in brackets[21]

The aim of this thesis is to gather all the textual characteristics of texts entrepreneurs write and post on their different platforms, and to distinguish if all entrepreneurs write in a unique way, more sophisticated one than non-entrepreneurs.

2.4 Datasets

In order to predict the entrepreneurial personality, and achieve high accuracy in whatever deep learning model used, there has to be a standardized benchmark data to train these models with. Depending on this dataset, the accuracy of the prediction can change and achieving a state-of-the-art models. In this part, we will discover the process of engineering such a dataset, through studying how different benchmark data used to train models to predict the human personality like well-known datasets: The Stream-of-consciousness Essays Database [26], myPersonality [16], and built datasets for specific researches.

Starting with the Essays Dataset [26], the database encompasses a diverse range of texts, including novels, short stories, and essays, from various literary periods and cultural contexts. On their first Phase, they defined all the reliability of the language use like the text analysis procedures using the LIWC program, which includes words and categories like emotion category and a sub-category whether it is a negative or positive emotion. Also, they specified the language composition, such as the total number of words, the number of words per sentence, and the number of questions, percentage of unique words, etc. Secondly, they gathered 3 samples of essays from three different sources. The first sample was in a form of daily writings from 15 residential patients in a substance abuse and addiction treatment center in England. The second sample was in a form of daily class assignments by Taos summer school students in New Mexico. The last sample was a group of published abstracts by 40 prominent social psychologists. Each writing sample for each participant was transcribed into a computer text file and analyzed with LIWC program, and the result is described in figure 2.3.
For their second phase, the 1203 essays were from original psychology student samples as assignments of their fall semester in university. These students also completed occasional questionnaires such as the Five Factor Inventory, these answers were used as a labeling for all the essays. Figure 2.4 shows the correlations between the language dimensions and the Five Factor answers. This dataset is considered as a benchmark data, used till this day to train models to predict personality traits.

Study characteristic	Sample 1: Inpatients	Sample 2: Taos summer school	Sample 3: SESP abstracts
Number of writers	15	34	40
Writing samples per writer	18	10	15
Words per sample			
M	166.2	516.7	118.3
SD	114.2	265.5	35.3
Mean correlation	.15	.16	.12[a]
Mean α	.64	.56	.56[a]
Percentage of language categories with Cronbach alphas $\geq .60$[b]			
Overall	69% (70%)	49% (50%)	47% (59%)
Language composition	83% (83%)	83% (88%)	39% (69%)
Psychological process	72% (72%)	48% (48%)	60% (60%)
Relativity	60% (60%)	40% (40%)	40% (40%)
Current concerns	58% (57%)	21% (22%)	47% (70%)

Figure 2.3: Summary of Reliability Studies[21]

LIWC factor	Five-factor dimension				
	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
Immediacy	**.10***	**.04**	**−.16****	**.07***	**−.02**
First-person singular	.13**	.04	−.13**	.07*	.01
Articles	−.09*	−.09*	.13**	−.15**	−.04
Words of more than 6 letters	−.03	−.04	.16**	−.03	.06
Present tense	.06	.01	−.15**	.04	.00
Discrepancies	.05	−.03	−.01	−.02	−.07*
Making Distinctions	**.05**	**−.14****	**.06**	**−.05**	**−.13****
Exclusive	.00	−.08*	.10*	−.06	−.08*
Tentativity	.06	−.14**	.11**	−.02	−.06
Negations	.05	−.12**	.00	−.04	−.15**
Inclusive	−.01	.07*	.01	.03	.06
The Social Past	**.04**	**.00**	**.08***	**−.02**	**−.04**
Past tense	.03	.04	−.03	.06	−.06
Social	−.01	.12**	.02	.00	.02
Positive emotion	−.13**	.15**	−.06	.07*	.07*
Rationalization	**−.06**	**.02**	**−.03**	**.07**	**.04**
Insight	.03	−.02	.07*	.05	−.01
Causation	.03	−.08*	−.08*	.00	−.07*
Negative emotion	.16**	−.08*	.05	−.07*	−.15**

Figure 2.4: LIWC factors and Simple Correlations With Five-Factor Scores[26]

Furthermore with the myPersonality Database [16], the paper discusses the distinct opportunities and challenges offered by Facebook to researches. It permitted users to complete genuine psychometric assessments and receive immediate results. Alongside test data, approximately 40% of participants also chose to share information from their Facebook profiles, resulting in one of the largest social science research databases in history. Respondents came from various age groups, backgrounds and cultures. The Database myPersonality offered a 360-degree assessment feature, encouraging users to invite their friends to judge their personality. This resulted in a database of cross-ratings,4 while also helping to increase the virality of the application; those who were invited to rate their friends often proceeded to take a test themselves. A sample of how the dataset was build is shown in Figure 2.5. The app remained operational until 2012, amassing data from over 6 million volunteers in 4 years. In April 2018, the researchers decided to stop sharing the data with other scholars.

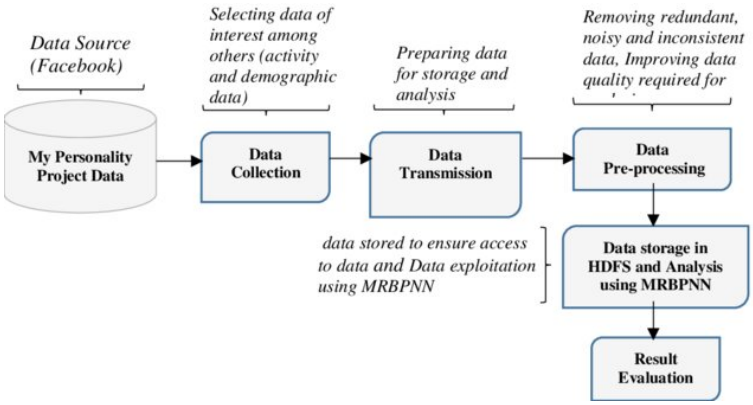

Figure 2.5: Process to engineering a dataset[16]

Moving on to a database that was built to predict the Big Five personality traits on a scale from -3 to 3, rather than a binary classification [2]. The researchers required

training data, to represent the whole data range for each trait in Swedish language. They proceeded by retrieving data from four different Swedish discussion forums and news sites with authors of different personalities. Web spiders were used to download the texts, and they were 70 million texts in total, but they ended up with only Thirty-nine thousand texts that could be annotated. The texts were annotated by 18 psychology students, each annotated a random text with a specific trait from the Big Five personality traits on a discrete integer interval from -3 to 3 as shown in Figure 2.6 . Due to the small number of annotators and the big amount of texts, each text approximately was annotated once. So, it was decided that a smaller subset of the large database would be taken and annotated furthermore, which resulted in a smaller dataset with two thousand texts, with on average 4.5 annotations each as shown in Figure 2.7.

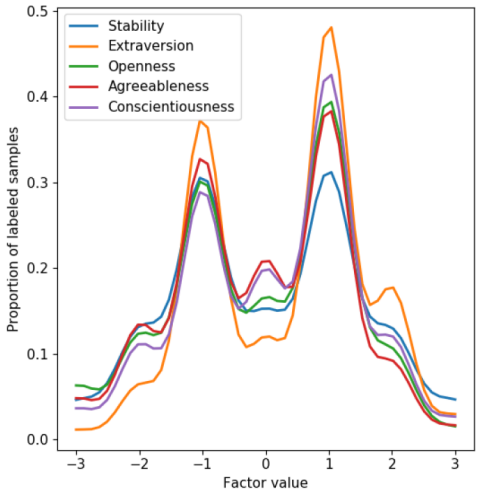

Figure 2.6: Distribution of labeled samples for each of the factors of the large dataset[2].

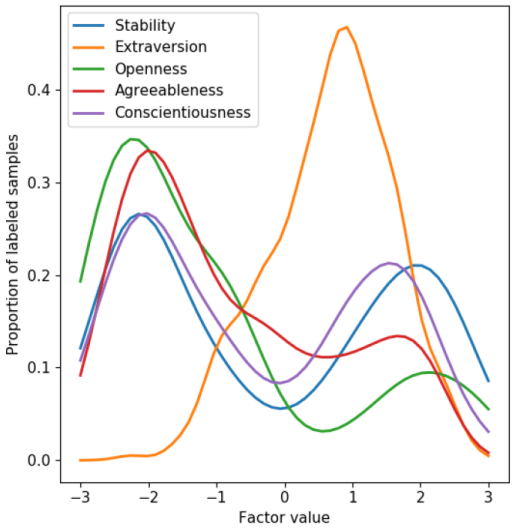

Figure 2.7: Distribution of labeled samples for each of the factors of the small dataset[2].

For feature extraction, ther used Term Frequency-Inverse Document Frequency (TF-IDF) to construct features from labeled data. It was used on word and character level with bi-gram for words and quad-grams for characters. Several regression models were tested, and they used ULMFiT as their Natural Language Processing (NLP). The performance of the models was evaluated with 5-fold cross validation test, as well as a binary classification test, and they did test their model on wild data like Cover Letters Dataset and Self-Descriptions Dataset. At the end, they were able to create models with reasonable performance, with accuracy in-line with the state-of-the-art models. And, it was found that using a smaller amount of high-quality training data with multi-annotator assessments resulted in models that outperformed models based on a large amount of solo-annotated data. Their results showed that extracting personality traits from a text remained a challenge.

2.5 Applications of The Benchmark Data

Imagine yourself as a recruiter standing in your company's booth in a career fair, meeting all the undergraduates and the graduates, and you are searching to recruit someone for a position with a high leadership skills, innovative, independent, and most importantly a risk taker. You have a new tool at the reach of your hands, to just let the student write a paragraph about himself, and from this paragraph, you will be provided with a direct response if this person has the potential of being an entrepreneur or not, and a detailed explanation on whether or not he should proceed to the next recruitment phase.

Analyzing a simple textual post or comment to predict of the author has the potential to be an entrepreneur or not, or he is already an entrepreneur and owns a business, would be a great asset for companies, for a country's economy and for this person's own self-development. Predicting the entrepreneurial personality can have several application and benefits. Firstly, in the realm of recruitment and selection, companies can harness personality assessments to pinpoint individuals possessing traits conducive to entrepreneurship. By identifying candidates with characteristics such as risk-taking propensity, creativity, and proactiveness, organizations can assemble teams better equipped to innovate and drive business growth [9].

On an individual level, personality assessments provide valuable insights into one's entrepreneurial strengths, weaknesses, and areas for development. By understanding their entrepreneurial personality profile, individuals can focus on honing relevant skills and attributes to enhance their chances of success in entrepreneurial endeavors. Also, it can help social network users to understand how others may perceive them based on how they communicate in social media, in addition to its evident applications in online sales and marketing, targeted advertising, large scale polling and healthcare analytics [37].

Similarly, organizations providing business incubation and support services to start-ups can utilize personality assessments to tailor their assistance to the specific needs of

entrepreneurs. By understanding the personality profiles of their clients, these organizations can offer targeted mentoring, resources, and networking opportunities to support the start-ups' entrepreneurial journey and maximize their chances of success [3].

Lastly, at a broader level, policymakers and economic development agencies can leverage personality analysis to understand the entrepreneurial culture within a region or community. By identifying individuals with entrepreneurial potential and providing support and resources to foster their endeavors, policymakers can cultivate a vibrant entrepreneurial ecosystem conducive to economic development and job creation. Policymakers might like to consider promoting and enhancing entrepreneurship predictive personality factors (particularly openness) early on in the education system among children, teens and students who have the potential to become entrepreneurs [3].

Chapter 3

Methodology

Engineering a data benchmark has been always a confusing challenge to achieve due to the fact of multiple pipelines to be followed and the different structures of the data. The aim of the thesis is to engineer a data benchmark of textual posts, written by entrepreneurs in different industries. For the purpose of this paper, an entrepreneur is a person who founded a completely independent company, whether on his own or having a co-founder, most importantly, he is working for himself. The benchmark will be a useful tool for the analysis of the entrepreneurial personalities among our communities, through the prediction of the machine learning models. This prediction will be a great asset on the personal level of the person to be an entrepreneur and be validated to have the entrepreneurial characteristics as mentioned before. Though, this prediction is crucially important for the companies looking for the right person with an independent mindset that will help the company reach a new limit for its target, which will indirectly have an impact on the country's economy.

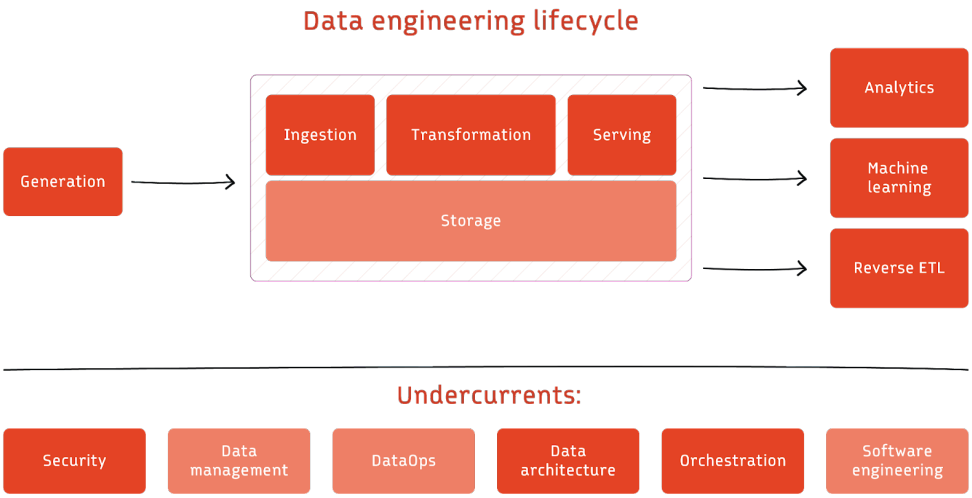

Figure 3.1: Data Engineering Pipeline [28]

Building a dataset requires to follow a particular process as shown in figure 3.1 from the two different processes Extract Transform Load (ETL), Extract Load Transform (ELT), and figure 3.2 illustrates the difference between the two processes. The one used in this thesis is the ELT, extracting the relevant information needed and transform this data through cleaning and labeling to be more relevant to the thesis' aim. And finally, to validate this data and load it and make it accessible for those who will need it. This chapter includes the different phases followed and the methodology to achieve results in my bachelor thesis. It is divided into 5 main categories: The Data Collection, The Data Transformation, The Textual Features Extraction, The Data Labeling, The Data Validation, and lastly, The Deployment of the data.

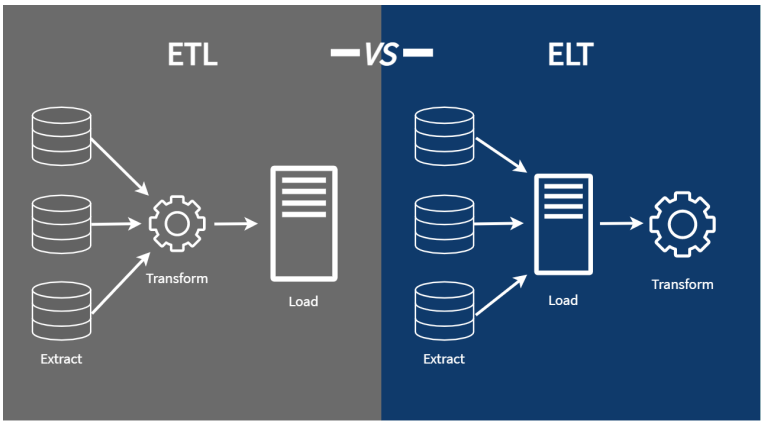

Figure 3.2: Difference between ETL and ELT [18]

3.1 The Data Collection

The data collection phase is divided into three main parts, according the pipeline followed in this thesis. A research was done to investigate the possible sources of the textual data by entrepreneurs. Afterwards, specifying all the entrepreneurs which will use their texts as the main source of our data, and lastly extracting these texts using various tools.

3.1.1 The sources of data

In contemporary times, anyone can write whatever comes their mind on whichever platforms they prefer on the Internet and publish it to reach the biggest audience. Unfortunately, not all platforms are trustworthy of the published texts and can give multiple false data to the researchers. The investigation of the sources of data phase is a very crucial step to maintain a high quality and comprehensive dataset.

While exploring the different platforms on the Internet that offers people to post their writing and express their opinions for specific topics or for the humanly right for freedom, an important consideration was maintained is to find sources that offers two

types of textual posts for our entrepreneurs. The first type focused on the use of the free speech of the entrepreneurs through their social media accounts and the conducted interviews done to discuss their daily lifestyle and their companies establishment journey. The second type focused on the formal use of language through typed articles and blog posts, where entrepreneurs share some advice, lessons learned through their journey to success, and their thoughts on specific topics regarding the industry of their business, or on the general business insights. Having both types of written data is a major key for the entrepreneurial personality analysis, since humans always use the two types on a daily basis, and both types reflect the author's personality characteristics.

After conducting a research on which platforms entrepreneurs usually use to express themselves, several sources and websites [24], were found to publish entrepreneurs' written contributions. For the free speech textual data type, three main sources were used:

1. **X (previously Twitter)**: It is a free social networking site [17] where users broadcast short posts known as tweets. These tweets can contain text, videos, photos or links. Twitter is known to be the platform where entrepreneurs can build and market for their brand, engage with customers, network with industry professionals, and most importantly stay updated on trends.

2. **LinkedIn**: It is the world's largest professional network on the internet. It is used to find the right job, connect and strengthen professional relationships, and learn the skills needed to succeed in a career. Entrepreneurs use it as a networking, personal branding, content marketing, industry insights, and partnership opportunities [29], [13].

3. **Mixergy**[1]: Founded by Andrew Warner in 2006, Mixergy offers a variety of content formats, including courses, podcasts, and an extensive library of interviews with founders, CEOs, and other influential figures in the business world.

Furthermore, the formal well-structured posts from entrepreneurs were drawn out from these four sources:

1. **The Entrepreneurs Library**[2]: It is a podcast, blog, and community for Business proprietors and individuals engaged in small-scale enterprise who loves to read books.

2. **Entrepreneur Media**[3]: It is a multimedia company that provides content, resources, and support for entrepreneurs and small business owners.

3. **StartupNation**[4]: Founded by Jeff Sloan, a multimedia company crafted by entrepreneurs for entrepreneurs, offering necessary insights for personal growth.

[1] https://mixergy.com/
[2] https://www.theelpodcast.com/
[3] https://www.entrepreneur.com/
[4] https://startupnation.com/

4. **Seth's Blog**[5]: The personal blog of Seth Godin, he is a prominent figure in the fields of marketing, entrepreneurship, and leadership.

After an exhaustive research on the sources of texts that will be used for the dataset. We gather all the websites Uniform Resource Locator (URL)(s) in a Google Sheet, with all the sub-pages that will be accessed to collect the data, too. Putting all the URL(s) in a data sheet will ease our automation process of transforming the rows of the sheet to a list, separated with commas, which will be used to loop on the list and access the content of each webpage. For example, we created two sheets, one for the list of all entrepreneurs we will extract their LinkedIn posts or their tweets from X (known as Twitter). The other sheet has all the specific webpages of a given website that will be accessed individually with the format of its pagination.

3.1.2 Web Scraping

Web scraping [15] is a technique used to extract data from websites. As shown in figure 3.3 It involves automated processes that navigate through web pages, gather information, and store it for analysis or other purposes. When applied to a list of web pages, web scraping can efficiently collect all the needed data across multiple sites. By writing scripts or using specialized software, users can specify the data they want to extract, such as text, images, or links, and define the pages to scrape. This method is particularly useful for tasks like market research, competitive analysis, and data aggregation, as it allows users to gather large amounts of information quickly and systematically. Additionally, web scraping can be customized to extract structured data from unstructured sources, making it a powerful tool for extracting insights from the vast expanse of the internet.

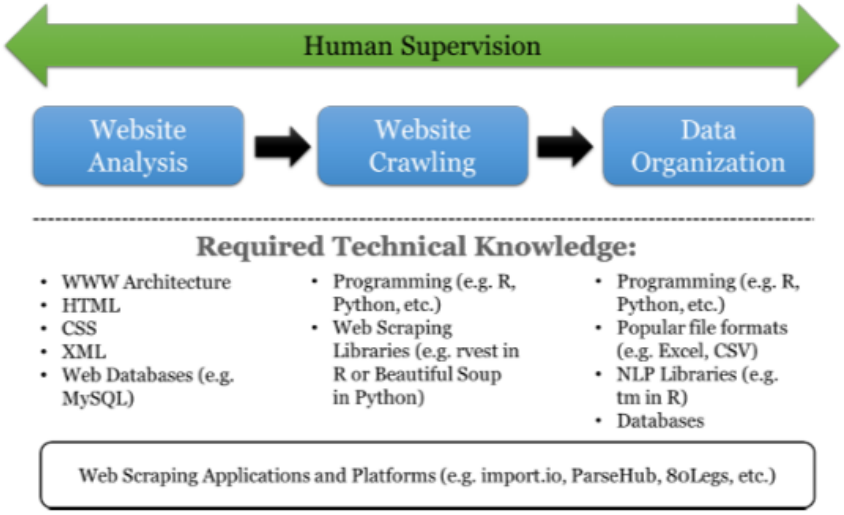

Figure 3.3: Web Scraping Process [15]

[5] https://seths.blog/

For each website in our sources, a Python Jupyter Notebook was created using its relative Python packages to extract the data as mentioned in table 3.1. The choice of these libraries depended on the source and the way we want to deal with it to be able to extract all data. For Twitter, it was straight forward to use *Nitter Scraper* since it is an open-source alternative for Twitter where the tweets of a user are seen and iterated through each page without any authentication. For LinkedIn, it wasn't the same easy case like Twitter, we used a webdriver to control the webpage with our code to natigate through the pages and buttons and get the results we want. And finally, for all the other websites, we used requests to access the URL and get their Hypertext Markup Language (HTML) page source.

Table 3.1: Sources and their respective Python Libraries

Source	Python libraries
Twitter	Nitter Scraper (ntscraper), Pandas
LinkedIn	Selenium, BeautifulSoup, Pandas
Other	Requests, BeautifulSoup

Starting of with Twitter, figure 3.4 below shows the simple python function used to extract all the tweets of a given username to the function responsible for fetching all the data. Along with specifying the fields necessary to form the meta data about our tweet as mentioned in the figure, we construct a dataframe to gather all data and structure it in a tabular form.

```
scraper = Nitter()
def create_tweets_dataset(user, no_of_tweets):
    tweets = scraper.get_tweets(user, mode = "user", number = no_of_tweets)
    if tweets is not None:
        data = {
        'link':[],
        'text':[],
        'name':[],
        'username':[],
        'likes':[],
        'quotes':[],
        'retweets':[],
        'comments':[],
        'date':[],
        'is_retweet':[]
        }

        for tweet in tweets["tweets"]:
            data['link'].append(tweet['link'])
            data['text'].append(tweet['text'])
            data['name'].append(tweet['user']['name'])
            data['username'].append(tweet['user']['username'])
            data['likes'].append(tweet['stats']['likes'])
            data['quotes'].append(tweet['stats']['quotes'])
            data['comments'].append(tweet['stats']['comments'])
            data['retweets'].append(tweet['stats']['retweets'])
            data['date'].append(tweet['date'])
            data['is_retweet'].append(tweet['is-retweet'])
        df = pd.DataFrame(data)
        df.to_csv(user+"_tweets_data.csv")
```

Figure 3.4: Twitter Scraping function using Nitter

Moving on to LinkedIn, the process of posts extraction is done through two main libraries, Selenium is a powerful tool for automating web browsers. It allowed us to interact with LinkedIn in the same way a user would, and BeautifulSoup to parse the HTML and Extensible Markup Language (XML) documents to extract HTML elements as demonstrated in figure 3.5.

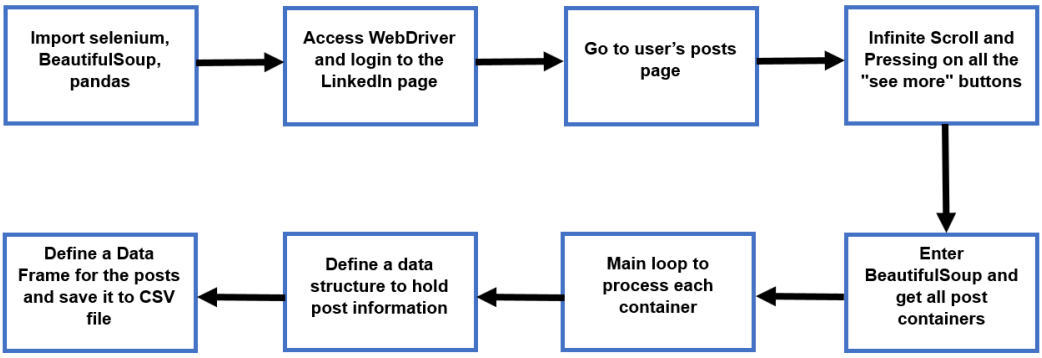

Figure 3.5: LinkedIn Scraping Process

At last, all the other sources only needed a great depth inspection to the HTML page source to be able to get all the HTML elements we are interested in, along with their class names, and a unique identifier to collect all the instances of the wanted elements in a page.

3.2 The Data Transformation

Data Transformation is our next step in the data analysis process, aimed at improving the quality and reliability of datasets. It involves integrating the different Comma-Separated Values file (CSV) files together, to come with one file, and begin the cleaning phase with all its modifications and replacemets within the data to reach a well-balanced and structured dataset.

3.2.1 Data Integration

A dataset has to have one structure for all the derived data from the different sources, that by default has different field names for our records. First of all, we defined a clear structure and the name of fields and its type which we want to appear in the final version of the data benchmark. Afterwards, we order and name all the CSV file headers with the same name specified earlier. Then, we add a column for each source's CSV file with the name of the source that the data was collected from to ensure that while cleaning the data, it would be easier to filter the sources and do a true inspection of the data. And the last step, we will gather all the files in one Excel Sheet with all the data from

the various sources. The last step is very important, since transforming and cleaning the data phase should be done once for all the data. Skipping this step will require doing the data cleaning on each CSV file produced from each source, each is very time consuming to do.

3.2.2 Data Cleaning

Data cleaning is a critical prerequisite for any meaningful data analysis and plays a fundamental role in ensuring the reliability and validity of findings in various fields. It involves identifying and correcting errors, inconsistencies, and inaccuracies within the data to ensure its integrity and usefulness for analysis. By conducting thorough data cleaning, analysts can avoid the risk of drawing incorrect conclusions or making flawed decisions based on flawed data. The data cleaning steps starts with an inspection of all the possible errors existing, then the actual cleaning.

Inspection

The inspection of data involves detecting the presence of any errors within the CSV file, which is done using Microsoft Excel automatically. The error can be a result of a non-consistent record value with the auto-generated type of the column by Excel. It draws our attention to any issues in the actual data before starting the cleaning process.

Cleaning

This step targets specifically the textual post, not the meta data and the other columns we have in the dataset. It includes a number of modifications to be done on the whole dataset as well as on the text, which will make it more readable and relevant to our dataset definition. These modifications are done through Microsoft Excel and Python libraries as Pandas and NumPy, and they are:

1. **Missing Data Handling**
   Missing values are data points that are absent for a specific variable in a dataset. They can be represented in various ways, such as blank cells, null values, or special symbols like "NA" or "unknown". There are a lot of strategies for missing values, like simply removing the whole row with a missing value, by imputation methods or by the forward or backward filling. We will use the mean imputation method to avoid reducing the sample size or losing the accuracy of the data and introducing certain bias. We filter the texts of the same author and get the mean of any numerical missing values to get the missing one. The figure 3.6 shows an example of doing such a technique with is done with the following formula:

$$MeanValue = \frac{\sum_{i=1}^{n} P_i}{n}$$

where P_i is the value of an i cell and n is the total number of values.
For the textual posts from the sources that hadn't any numerical meta data, we inserted a default value "N/A", to specify that these texts didn't have a count for readers' reactions.

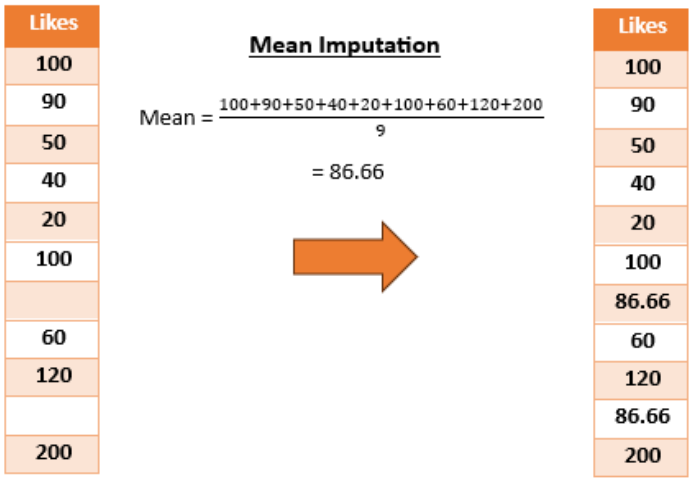

Figure 3.6: Example of mean imputation on the likes column

2. **Duplicates Removal**
   It is very possible to have a lot of duplicate rows in any dataset, even if is collected through an automated process. Removing duplicates from a dataset ensures data integrity and quality, optimizing storage, and facilitating more accurate analyses. This process enhances overall efficiency and reliability in data management and analysis. Microsoft Excel has a function that can be called on any number of rows and it automatically removes duplicates in our data. The duplicates removal was done based on the textual post only, not any other field, as it has to be the only unique value, at least within all the texts of the same author. As illustrated in figure 3.7, the cells with the same color represents the same content, so we removed the same color cells for each author.

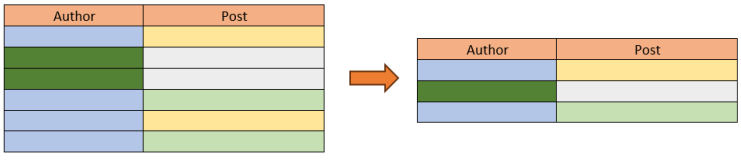

Figure 3.7: Example of Removing Duplicate Posts

3. **Irrelevant Data Removal**
   For the entrepreneurial personality analysis, it is very important to have some criteria on the text, to be able to reach validated insights on the data. After a

deeper looking into the length of the text that could be analyzed, we came to a conclusion that any text cell value having a number of characters less than 35 characters, the whole row shall be removed. This will be done with Microsoft Excel after adding a column for the length of each text record and removing the rows that applies.

4. **Data Type Conversion**
   Microsoft Excel has an auto-detection of the data type in a column based on the type of the majority of the value. Normally, it directly assigns the type Number to any numerical value, and a string to any textual value. When, it comes to dates, it has only some acceptable formats to write a date, and all the values on a column should hold the same date format, so it is a crucial step to insure that all the dates are written in the same way as depicted in figure 3.8.

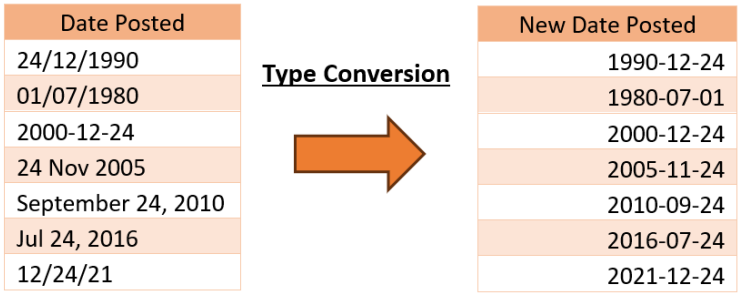

Figure 3.8: Example of Date Format Conversion

5. **Language Translation**
   To have consistent data, we need to have everything in the same language, because the NLP models that analyzes the data are monolingual, they don't process multiple languages at once. We choose English to be our main language. As their are entrepreneurs all around the world, so their was some posts written in Indian and Spanish and other languages. So the languages that we were able to translate to English, we definitely did and for those who we couldn't we had to remove these records.

6. **Replacing Emojis and Special Characters**
   Since our data has a lot of free speech textual posts like tweets. Emojis frequently appear in tweets, adding expressive elements to messages and enhancing communication on social media platforms. Unfortunately, emojis can produce unreliable data analysis, and we can't just remove them because nowadays, people tend to express their feelings using emojis since it is easier. Using Demoji library from Python, emojis will be replaced by their textual meeting as illustrated in figure 3.9 to keep what the entrepreneur really intended to tweet and helps us analyze better their personality characteristics.

```
import demoji

text="IN 📚 ❤️ 🌺 🌹 👶"
demoji.findall(text)

#clcoding.com
```

```
{'📚': 'books',
 '👶': 'baby',
 'IN': 'flag: India',
 '🌺': 'hibiscus',
 '🌹': 'rose',
 '❤️': 'red heart'}
```

Figure 3.9: Example of Emoji Replacement[30]

Minimal cleaning was done on the textual posts, we avoided modifying the text or any specific pre-processing data method to ensure that the text remains in its original format with all the punctuation and the capitalization. The original form will be more useful in the data analysis and the feature engineering of the text that will be later used in the author's personality analysis.

3.3 Textual Features Extraction

Textual features encompass various attributes and characteristics present within written content, serving as key elements for analysis and interpretation. Figure 3.10 illustrates the two main branches of the textual features, and we extracted some of the elements of both[20].

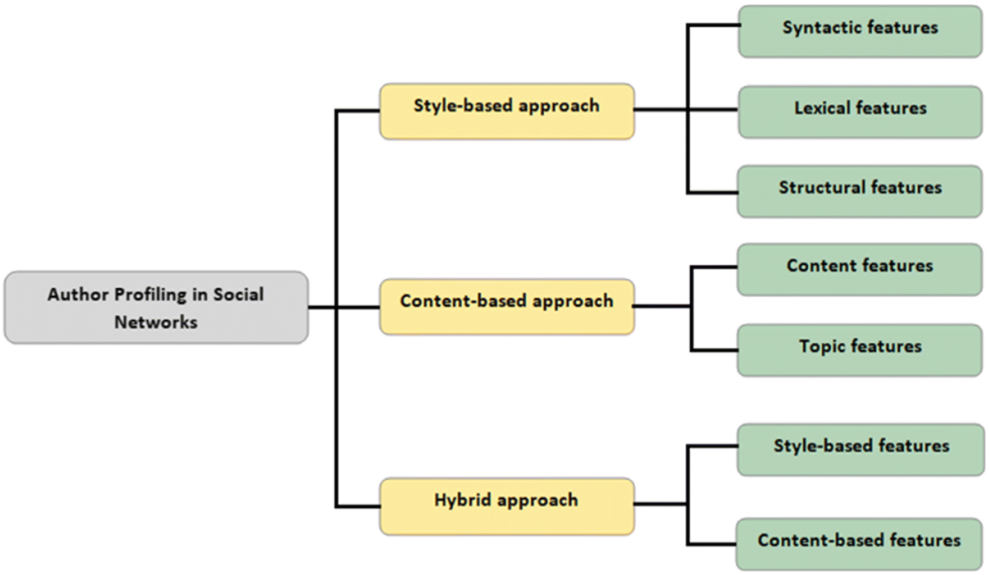

Figure 3.10: Text Analysis

3.3.1 Content-based Features

In NLP, feature extraction [34] is a fundamental task that involves converting raw text data into a format that can be easily processed by machine learning algorithms. There are various techniques available for feature extraction in NLP, each with its own strengths and weaknesses. There are multiple available techniques and each has a different use case as demonstrated in table 3.2. After studying each use case for the mentioned techniques, we will be applying the CountVectorizer technique, which will provide us with enough information on the most used words in each text or in the whole CSV. This technique will be applied using Python, as captured in figure 3.11.

```python
# CountVectorizer
count_vec = CountVectorizer()
X_count = count_vec.fit_transform([text])
print('CountVectorizer:')
print(count_vec.get_feature_names_out()[:10])
print(X_count.toarray()[0][:10])
```

```
CountVectorizer:
['amounts' 'analyze' 'and' 'artificial' 'between' 'computational'
 'computer' 'computers' 'concerned' 'data']
[1 1 3 1 1 1 1 2 1 1]
```

Figure 3.11: Example of CountVectorizer with Python[10]

The application of CountVectorizer in our case, is to have the words with the biggest number of occurrences within a text. Therefore, the returned words will help to better understand the most used words by entrepreneurs.

Table 3.2: Comparison of Text Feature Extraction Techniques[10]

Technique	**Main Features**	**Use Cases**	**Size and Complexity**
CountVectorizer	Converts text to Matrix of word count	Text classification, topic modeling	Simple and fast, suitable for small to medium-sized datasets
TF-IDF	Assigns weights towards based on importance	Information retrieval, text classification	More complex and computationally expensive suitable for medium to large-sized datasets
Word embeddings	Vector representation of words based on semantics and syntax	Text classification, information retrieval	Can handle large datasets computationally expensive to train
Bag of Words	Represents text as a vector of word frequencies	Text classification, sentiment analysis	Simple and fast suitable for small to medium-sized datasets
Bag of n-grams	Captures frequency of sequences of n words	Text classification, sentiment analysis	It depends on the size of the n-grams and the datasets
Hashing Vectorizer	Maps words to fixed-size features space using hashing function	Large scale text classification, online learning	Suitable for large datasets memory efficient May suffered from hash collisions

3.3.2 Style-based Features

The style-based features of a text focus on linguistic attributes such as tone, writing style, frequency of use of punctuation, and in case of the tweets, the number of hashtags and emojis. The following list shows all the features we derived from the textual posts, that will be used in the later section:

1. *Use of Pronouns:*
   The conversational tone was measured with the frequency of occurrences of specific pronouns and its instances as listed in table 3.3. This feature will be extracted through Python by specifying the pronouns as a dictionary and searching through the CountVectorizer result, we will get the actual frequency.

Table 3.3: Pronouns for Tone Analysis

Nominative	**Objective**	**First Possessive**	**Second Possessive**
I	Me	My	Mine
You	You	Your	Yours
He	Him	His	His
She	Her	Her	Hers
We	Us	Our	Ours
They	Them	Their	Theirs

2. *The average word length:*
   The length of a word tends to determine the tone of the author. As shorter words tends to be punchier and harder, in opposition to longer words that give a softer effect [6]. So it will be beneficial to calculate the average word length used in the text through Python.

3. *The average sentence length:*
   Same as for the word length, the length of a sentence decides the tempo of the long text. Shorter sentences, less than or equals to 10 words give a concise style. While longer ones can be a little confusing.

4. *The text length by word count and sentence count:*
   Text length can significantly impact communication effectiveness by influencing readability and comprehension. It balances conveying sufficient information without overwhelming the reader. We used Python to count the number of words and number of sentences in the text.

5. *Polarity:*
   The polarity score is calculated to assess whether the overall text leans towards being positive or negative. The positive and negative score in calculated based on a positive or negative dictionary that when a word is found the score increase in either side. Table 3.4 shows an example of the negative and positive dictionary. The formula used to compute the polarity score is as follows, illustrated in figure 3.12:

$$PolarityScore = \frac{PositiveScore - NegativeScore}{PositiveScore + NegativeScore + 0.000001}$$

Table 3.4: Example of Positive and Negative Dictionary[4]

Serial No.	Positive words	Negative words
1	Amazing	Avoid
2	Authentic	Mistakes
3	Best	Bad
4	Benefits	Complicated
5	Better	Error
6	Great	Fail
7	Happy	Sad
8	Inspiring	Unhappy
9	Productive	
10	Thankful	

6. *Subjectivity:*
The subjective score is used to assess the level of subjectivity or opinion expressed in a text. Figure 3.12 illustrates the relationship between the polarity and the subjectivity of a text. It is calculated using the formula:

$$SubjectiveScore = \frac{PositiveScore + NegativeScore}{TotalWordsaftercleaning + 0.000001}$$

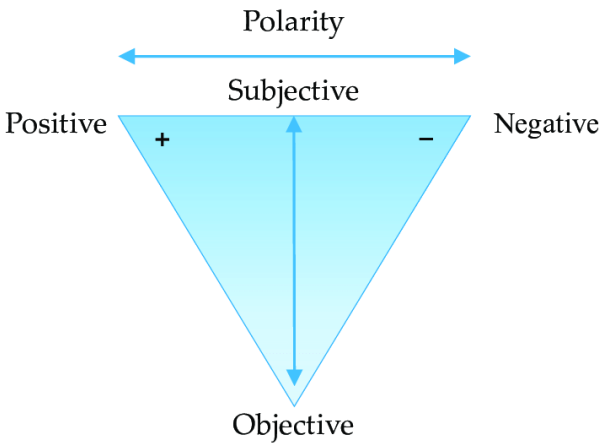

Figure 3.12: The relationship between Polarity and Subjectivity [12]

7. *Complex Word Count:*
The complex word count refers to the number of words in the text that have more than two syllables [7].Using Python's library *cmudict*, we will build a function that counts the syllables of a word and then check if the word in complex and finally, return the complex words count.

8. *Use of Punctuation:*
   Punctuation is essential in writing because it helps to convey meaning and clarify the structure of sentences. Without punctuation, sentences can become confusing or ambiguous, making it harder for readers to understand the intended message. Punctuation marks as demonstrated in figure 3.13.

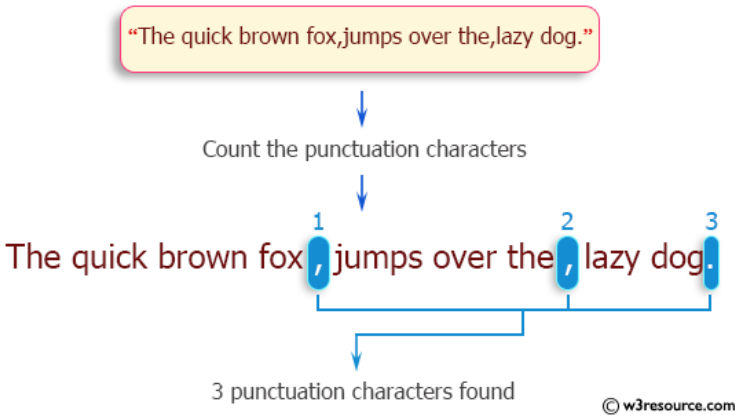

Figure 3.13: The punctuation count in a sentence with Python

9. *Uppercase and Lowercase words:*
   Identifying the frequency of the using Uppercase words in opposition to the lowercase words. We will extract two values, one for the whole uppercase words and one for only the words having their first letter an uppercase. And, same goes for the lowercase words.

3.4 The Data Labeling

Data labeling is the process of adding tags or labels to raw data, in our case, the raw data will be the textual posts or tweets of entrepreneurs. Since, the aim of this thesis in to analyze the entrepreneurial personality through verbal behaviour, data labeling will help to come up with conclusions on how entrepreneurs write. And, for future work we wish to be make the prediction if a given textual input has the entrepreneurial personality or not, the labels would represent an object class to help Machine Learning models learn to recognize specific classes within the data without labels.

Using all the extracted features mentioned in the previous section, a column will be made for each feature with the header as the feature name and the value or score it got on this feature. As demonstrated in figure 3.14 we added to the rows additional information regarding their features.

Figure 3.14: Example of data labeling

Since, the focus of the thesis is about the characteristics of the entrepreneurial personality and this is the first dataset of entrepreneurs' writings, we focus on the LIWC features more than labeling the dataset with the personality traits of the text. To have a ready dataset as input for the Machine Learning prediction models, we aspire to have a labeled dataset with the Big Five Personality variables or MBTI variations, to better understand the different aspects of the entrepreneurial personality. Unfortunately, this aspiration is not possible at the moment, since only two state-of-the-art models, the LIWC and the IBM Watson Personality Insights, offer this feature to extract the personality traits from text. And, these two models have a very limited free access. This task can be done as a new labeled version of this dataset with the traits in the future.

3.5 The Data Validation

To ensure that data is accurate, complete, and consistent, data validation is the necessary process for this step. It involves checking data for errors, inconsistencies, and anomalies to maintain data quality and reliability. This process is crucial, especially when dealing with large datasets, as errors can propagate quickly and impact downstream analyses or decisions.

Data can be examined as part of a validation process in a variety of ways, including data type, constraint, structured, consistency and code validation. Each type of data validation is designed to make sure the textual posts meet the requirements to be useful for analysis.

Among the most basic and common ways that data is used is within a spreadsheet program such as Microsoft Excel or Google Sheets. In both Excel and Sheets, the data validation process is a straightforward, integrated feature. Excel and Sheets both have a menu item listed as Data ¿ Data Validation. By selecting the Data Validation menu, a user can choose the specific data type or constraint validation required for a given file or data range.

Figure 3.15 illustrates the automatic process of data validation that is done. The types of data validation we did on our dataset:

- **Data type validation:**
  It confirms that the data in each field or column matches a specified data type and format.
- **Constraint validation:**
  It checks to see if a given data field input fits a specified requirement within certain ranges. For example, it verifies that a textual post field has the minimum number of characters.
- **Structured validation:**
  It ensures that the data is compliant with the specified data schema we set at the data transformation phase.
- **Consistency validation:**
  It makes sure data styles are consistent. For example, it confirms that all values of the feature extraction scores are listed to two decimal points.

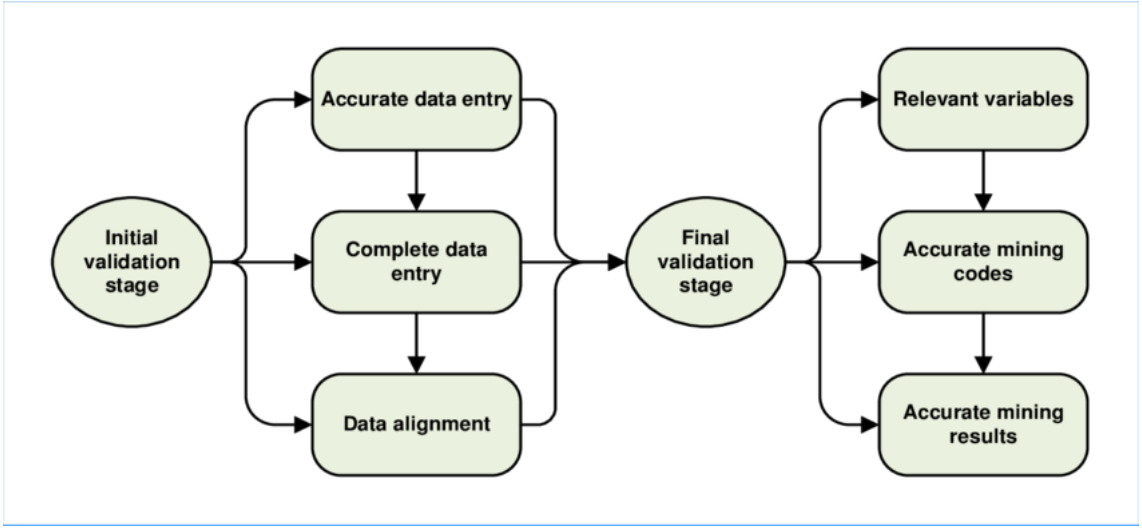

Figure 3.15: Data validation process stages[25]

The output of any given system can only be as good as the data the operation is based on. These operations can include machine learning or artificial intelligence models, data analytic reports and business intelligence dashboards. These reports and data visualization can be a great asset in validating the data and discover any outliers in the columns.

3.6 Data Visualization

Data visualization is the visual presentation of data or information. The goal of data visualization is to communicate data or information clearly and effectively to readers.

The field of data visualization combines both art and data science. While a data visualization can be creative and pleasing to look at, it should also be functional in its visual communication of the data.

Data visualization can be used for:

- Making data engaging and easily digestible
- Identifying trends and outliers within a set of data, as mentioned before in the data validation phase
- Telling a story found within the data
- Highlighting the important parts of a set of data

Data can be represented in a lot of forms, but most importantly, is the choice of the form and how it serves our use case and the data type we want to visualize. Figure 3.16shows some of these forms in a very artistic way.

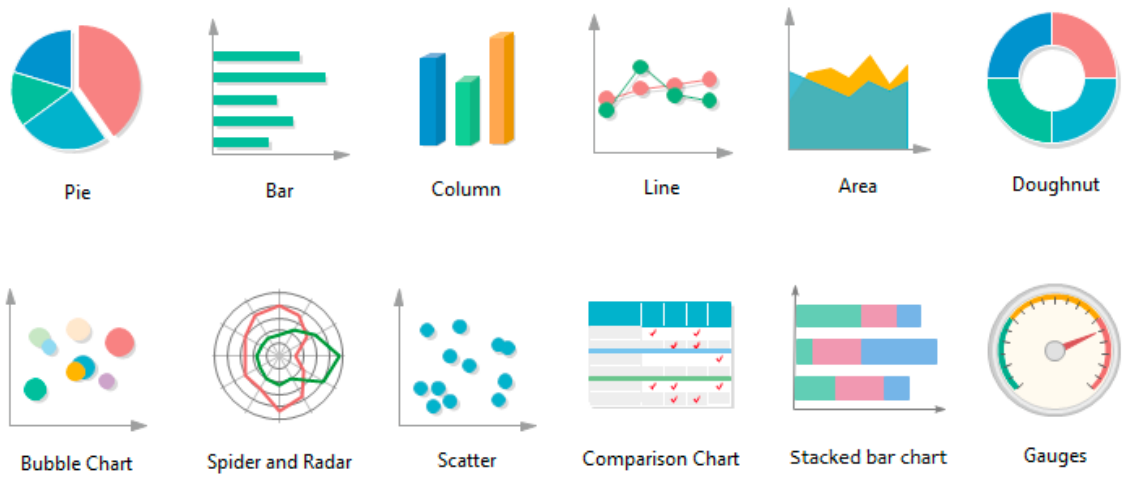

Figure 3.16: Data Visualization[31]

Chapter 4

Results & Limitations

4.1 Experiments Setup

Section text.

4.2 Data Description

Section text.

4.3 Results

Section text.

4.3.1 Experiment 1

Subsection text.

4.3.2 Experiment 2

Subsection text.

4.4 Results Analysis and Discussion

Section text.

Chapter 5

Conclusion & Future Work

5.1 Conclusion

Section text.

5.2 Future Work

Section text.

Appendix

Appendix A

Lists

MBTI	Myers-Briggs Type Indicator [14]
DISC	Dominance Influence Steadiness Conscientiousness[14]
META	Measure of Entrepreneurial Tendencies and Abilities [27]
LIWC	Linguistic Inquiry Word Count [26]
NLP	Natural Language Processing
ETL	Extract Transform Load
ELT	Extract Load Transform
URL	Uniform Resource Locator
HTML	Hypertext Markup Language
XML	Extensible Markup Language
CSV	Comma-Separated Values file

List of Figures

List of Tables

Bibliography

[1] Gorkan Ahmetoglu, Franziska Leutner, and Tomas Chamorro-Premuzic. Eq-nomics: Understanding the relationship between individual differences in trait emotional intelligence and entrepreneurship. *Personality and individual differences*, 51(8):1028–1033, 2011.

[2] Nazar Akrami, Johan Fernquist, Tim Isbister, Lisa Kaati, and Björn Pelzer. Automatic extraction of personality from text: Challenges and opportunities. In *2019 IEEE international conference on big data (big data)*, pages 3156–3164. IEEE, 2019.

[3] Bostjan Antoncic, Tina Bratkovic Kregar, Gangaram Singh, and Alex F DeNoble. The big five personality–entrepreneurship relationship: Evidence from slovenia. *Journal of small business management*, 53(3):819–841, 2015.

[4] Shilpa Balan and Janhavi Rege. Mining for social media: Usage patterns of small businesses. *Business Systems Research Journal*, 8, 01 2017.

[5] Seren Başaran and Obinna H Ejimogu. A neural network approach for predicting personality from facebook data. *Sage Open*, 11(3):21582440211032156, 2021.

[6] Charlotte Baxter-Read and Charlotte Baxter-Read. The 12 elements of tone, October 2023.

[7] Celestial. Text analysis - celestial - medium. *Medium*, June 2023.

[8] Marcelino Cuesta, Javier Suárez-Álvarez, Luis M Lozano, Eduardo García-Cueto, and José Muñiz. Assessment of eight entrepreneurial personality dimensions: Validity evidence of the bepe battery. *Frontiers in Psychology*, 9:375594, 2018.

[9] Kamal El-Demerdash, Reda A El-Khoribi, Mahmoud A Ismail Shoman, and Sherif Abdou. Deep learning based fusion strategies for personality prediction. *Egyptian Informatics Journal*, 23(1):47–53, 2022.

[10] Sahel Eskandar. Exploring feature extraction techniques for natural language processing. *Medium*, April 2023.

[11] Michael Frese and Doris Fay. 4. personal initiative: An active performance concept for work in the 21st century. *Research in organizational behavior*, 23:133–187, 2001.

[12] Usman Ghani, Imran Sarwar, and Aimen Ashfaq. A fuzzy logic based intelligent system for measuring customer loyalty and decision making. *Symmetry*, 10:761, 12 2018.

[13] Alexandra Ioanid, Cezar Scarlat, and Gheorghe Militaru. How managers and entrepreneurs use the innovative social technologies. In *European Conference on Innovation and Entrepreneurship*, page 298. Academic Conferences International Limited, 2015.

[14] Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features. *arXiv preprint arXiv:2204.04629*, 2022.

[15] Moaiad Ahmad Khder. Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, 13(3), 2021.

[16] Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist*, 70(6):543, 2015.

[17] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

[18] Nicholas Leong. How i redesigned over 100 etl into elt data pipelines. *Medium*, May 2023.

[19] Rongfan Liao, Siyang Song, and Hatice Gunes. An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition. *IEEE Transactions on Affective Computing*, 2024.

[20] Tatiana Litvinova, Pavel Seredin, Olga Litvinova, and Olga Zagorovskaya. Profiling a set of personality traits of text author: what our words reveal about us. *Research in Language*, 14(4):409–422, 2016.

[21] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500, 2007.

[22] Robert R McCrae. The five-factor model of personality traits: Consensus and controversy. *The Cambridge handbook of personality psychology*, pages 148–161, 2009.

[23] Michael D Mumford, Samantha Elliott, and Robert W Martin. Intrapreneurship and firm innovation: Conditions contributing to innovation. In *The Psychology of Entrepreneurship*, pages 97–117. Routledge, 2020.

[24] Martin Obschonka, Christian Fisch, and Ryan Boyd. Using digital footprints in entrepreneurship research: A twitter-based personality analysis of superstar entrepreneurs and managers. *Journal of Business Venturing Insights*, 8:13–23, 2017.

[25] Oluwatimilehin Oluwaseun Okeowo. *Investigating quality of data and the need for the restructuring of accident report form in South Africa*. PhD thesis, Stellenbosch: Stellenbosch University, 2018.

[26] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[27] Álvaro Postigo, Marcelino Cuesta, Eduardo García-Cueto, Francisco Prieto-Díez, and José Muñiz. General versus specific personality traits for predicting entrepreneurship. *Personality and Individual Differences*, 182:111094, 2021.

[28] Redpanda. Fundamentals of data engineering—life-cycle, best practices, and emerging trends.

[29] Chris J Reed. *Linkedin Mastery for Entrepreneurs*. Evolve Global Publishing, 2018.

[30] SarahDev. Convert emoji into text in python - sarahdev - medium. *Medium*, September 2023.

[31] Sdhglobal. Ai data visualization: Types, examples, and tools - sdhglobal - medium. *Medium*, November 2023.

[32] H Ramananda Singh and Habib Rahman. Entrepreneurs' personality traits and their success: An empirical analysis. *Research Journal of Social Science and Management*, 3(7):99–104, 2013.

[33] Javier Suárez-Álvarez, Ignacio Pedrosa, Eduardo García-Cueto, and José Muñiz. Screening enterprising personality in youth: An empirical model. *The Spanish Journal of Psychology*, 17:E60, 2014.

[34] Ayisha Tabassum and Rajendra R Patil. A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, 7(06):4864–4867, 2020.

[35] Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, et al. Personality prediction system from facebook users. *Procedia computer science*, 116:604–611, 2017.

[36] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[37] Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. Inferring latent user properties from texts published in social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.