algorithms

MDPI

Article

Do What You Say—Computing Personal Values Associated with Professions Based on the Words They Use

Aditya Jha and Peter A. Gloor *

MIT System Design Management, 77 Massachusetts Avenue, Cambridge, MA 02142, USA; jhaaditya1707@gmail.com
* Correspondence: pgloor@mit.edu

Abstract: Members of a profession frequently show similar personality characteristics. In this research, we leverage recent advances in NLP to compute personal values using a moral values framework, distinguishing between four different personas that assist in categorizing different professions by personal values: "fatherlanders"—valuing tradition and authority, "nerds"—valuing scientific achievements, "spiritualists"—valuing compassion and non-monetary achievements, and "treehuggers"—valuing sustainability and the environment. We collected 200 YouTube videos and podcasts for each professional category of lawyers, academics, athletes, engineers, creatives, managers, and accountants, converting their audio to text. We also categorize these professions by team player personas into "bees"—collaborative creative team players, "ants"—competitive hard workers, and "leeches"—selfish egoists using pre-trained models. We find distinctive personal value profiles for each of our seven professions computed from the words that members of each profession use.

Keywords: NLP; machine learning; HR analytics; job search; virtual tribes

1. Introduction: Predicting Professions from Text Using NLP and ML

In this research project, we apply recent advances in natural language processing (NLP) and machine learning (ML) techniques to personality prediction. In particular, we try to predict professions from transcripts of YouTube videos created by members of these professions. Using video transcripts, we develop models that infer professional roles with high accuracy.

Zhou et al. [1] surveyed the state-of-the-art deep learning methods for ABSA, categorizing approaches into lexicon-based, traditional machine learning, and deep learning methods. Their work provides a foundation for exploring advanced techniques, which we extend by applying contextual embeddings and hybrid models to profession prediction.

Combining different methods, including unsupervised clustering techniques, Support Vector Machines (SVMs), BERT, and Bi-LSTM, allows us to categorize professionals into distinct career classes based on the words they use in these videos.

Furthermore, we also demonstrate how the combination of deep learning models and personality extraction techniques provides a robust framework for identifying personality attributes of professions. Key performance metrics, including 85% precision, 80% recall, and an F1 score of 0.77, highlight the efficacy of our approach. Alongside this, we introduce "personas", groups of personality characteristics, inspired by real-world behaviors (e.g., "Beeflow", "Antflow", and "Leechflow") to further understand the intersection of personality and profession [2].

In the following sections, we will first outline the classical models explored, detail the application of unsupervised and supervised learning techniques, and introduce our

check for updates

Academic Editor: Ioannis Tsoulos

Received: 18 December 2024
Revised: 21 January 2025
Accepted: 28 January 2025
Published: 1 February 2025

Citation: Jha, A.; Gloor, P.A. Do What You Say—Computing Personal Values Associated with Professions Based on the Words They Use. *Algorithms* **2025**, *18*, 72. https://doi.org/10.3390/a18020072

novel frameworks and tools for personality analysis and profession classification. This will provide readers with information about both the methodology and the broader implications of profession prediction from text.

2. Background

2.1. Personality Traits and Occupational Roles

The relationship between personality traits and occupational roles has been a long-standing area of research. Recent advances in natural language processing (NLP) and machine learning (ML) have enabled the analysis of large-scale textual data from social media and professional networks, providing new insights into this dynamic [3]. Studies show that personality traits are crucial for professional success and organizational outcomes. Gloor et al. [4] emphasized the role of ethical values in fostering teamwork and improving efficiency.

Woods and Hampson [5] demonstrated the long-term influence of childhood personality traits, particularly openness/intellect and conscientiousness, on adult occupational environments. Their study further revealed that gender moderates these relationships, especially in strongly sex-typed vocations. Similarly, Floricia et al. [6] explored the role of personality in shaping professional choices, emphasizing how social, psychological, and economic factors interact to influence career paths in a rapidly changing society.

Eakman and Eklund [7] extended this understanding by showing that personality traits significantly impact perceptions of meaningful occupation and occupational value, which are strong predictors of life satisfaction and meaning in life. Building on this, Csikszentmihalyi et al. [8] highlighted the importance of finding purpose and flow in work, noting that while human consciousness offers freedom and flexibility in occupational choices, it also necessitates the creation of meaning to sustain effort and satisfaction.

Collectively, these studies demonstrate that personality traits not only influence occupational choices and success but also shape perceptions of meaningful work, well-being, and broader societal outcomes.

2.2. Advancements in Personality Prediction

Deep learning has revolutionized personality prediction by enabling the analysis of complex textual data. Sun et al. [9] demonstrated the effectiveness of models like RNN and LSTM in uncovering personality traits, while frameworks using transformer architectures such as BERT and RoBERTa have improved contextual understanding and prediction accuracy [10]. Jain et al. [11] introduced "Personality BERT", a fine-tuned transformer model for personality detection based on the Myers–Briggs Type Indicator (MBTI) framework, which effectively classifies personality types by analyzing writing styles. These advancements have facilitated the extraction of meaningful insights from diverse data sources.

The intersection of personality traits and behavioral vulnerabilities has also been a focus of psychological and computational research. For instance, recent work by Smith et al. [12] explores the correlation between the Big Five personality traits and susceptibility to phishing attacks, a prevalent social engineering technique. Their study employs a conditional generative adversarial network (C-GAN) to address the challenges of data scarcity and bias, generating realistic synthetic data for analysis.

2.3. Social Media as a Data Source

Social media platforms have emerged as rich data sources for personality analysis. Platforms like Twitter and Facebook allow researchers to analyze virtual communities, uncovering shared values and professional traits. Gloor et al. [2] demonstrated how

these virtual tribes provide valuable insights into personality and professional alignment, enabling researchers to bypass traditional survey-based approaches. Pradhan et al. [3] further emphasized the role of NLP and deep learning in enhancing personality analysis accuracy, showcasing their applicability in various fields such as marketing and user experience design.

Social media activity also reveals aspects of personality through shared content and interaction patterns. Recent work highlights the potential of NLP methods in analyzing such activity. For instance, a linguostylistic personality traits assessment (LPTA) system estimates Twitter users' personality traits using the Myers–Briggs-type indicator (MBTI) and big-five personality scales. The system employs an innovative input representation mechanism that converts tweets into real-valued vectors using frequency, co-occurrence, and context (FCC) measures, outperforming state-of-the-art systems with just 50 tweets per user [13].

2.4. Profession Prediction Models

In the context of profession prediction, RNN-LSTM-based systems have proven effective in linking linguistic features to occupational categories. These models not only predict current professions but also suggest suitable career paths based on personality profiles [14]. Personality BERT further advances this field by leveraging transformer models to capture nuanced textual features, which significantly improve the classification of professions based on personality traits [11]. Such systems provide valuable tools for both job seekers and hiring managers, offering a data-driven approach to career planning.

2.5. Personality Traits and Self-Employment

Research has also explored the connection between personality traits and self-employment. The Big Five personality traits have been shown to influence an individual's decision to become self-employed, particularly in professions categorized as part of the "creative class". A study based on the German Socio-Economic Panel (SOEP) found significant but varying associations between personality traits and self-employment propensity across different professions [15]. These findings highlight the complex interplay between personality, professional roles, and entrepreneurial tendencies.

Frameworks like Happimetrics[2] categorize professions into value personas, offering deeper insights into how personal values influence professional choices. However, existing models often struggle with context-specific language and the dynamic nature of personality traits. Addressing these challenges requires integrating multi-modal data, such as combining textual, visual, and auditory cues, to enrich personality assessments and improve prediction accuracy.

3. Our Approach: Professional Values Personas

This paper introduces a novel approach leveraging machine learning to predict personality characteristics of different professions. We go beyond a simple classification by exploring how individual traits, values, and worldviews influence professional identity. This section details the novel concepts introduced in our research:

3.1. Alternative Realities

In earlier work, we identified two groups of personas that combine related clusters of personality attributes commonly associated with different professions. *Alternative realities* refers to four distinct categories of individuals based on their personality traits and professional orientations. These categories, which serve as a framework for understanding societal dynamics, are referred to as *spiritualists*, *nerds*, *fatherlanders*, and *treehuggers* [16].

Each of these groups embodies a unique worldview and approach to life, providing a way for individuals to reflect on their own values and roles in society.

- Spiritualists provide spiritual guidance, whether as priests, meditation coaches, or yoga teachers. Their aim is to elevate moral values and ethical behavior.
- Nerds represent technocrats and scientists, using their knowledge to develop and apply technology in fields such as engineering, computer science, and innovation.
- Fatherlanders defend their nations and organizations, embodying leadership roles such as politicians, soldiers, or corporate executives.
- Treehuggers are environmental activists who resist the overuse of technology and advocate for sustainability, often feeling disempowered in the face of modern industrial practices.

These categories closely correspond to the ancient Indian *Varna* caste system, with *brahmins* (spiritual leaders), *ksathriyas* (warriors and rulers), *vaishyas* (merchants and artisans), and *shudras* (laborers and servants) serving as historical analogs [17]. The spiritualists align with the brahmins, nerds with the vaishyas, fatherlanders with the ksathriyas, and treehuggers with the shudras (as the "treehugger" extinction rebellion members are striving to save the world from a position of weakness by gluing themselves to the highway). Each modern category can be seen as a reflection of these traditional societal roles, providing insight into the historical and contemporary structures of society.

Ultimately, the concept of alternative realities offers a lens through which we can analyze the motivations, ethical behavior, and influence of individuals across these categories. By doing so, we can better understand how leadership, trust, and power operate within different *realities* of human interaction.

3.2. Behavioral Categories: Beeflow, Antflow, and Leechflow

The second group of personas identifies clusters of personality characteristics describing team player attributes. Inspired by the behaviors of bees, ants, and leeches, we classify human actions in terms of creativity, competition, and exploitation. These categories help explain not only professional roles but also how individuals engage in their work and interact with others [18].

3.2.1. Beeflow

"Beeflow" activity refers to individuals who, like creators and innovators, engage in collaborative activities that benefit both themselves and society. They derive satisfaction from creating something new or improving upon the existing structures. Their focus is on adding value to the world, whether through art, technology, or services. People in this category often experience a state of *flow* [8], where they become fully immersed in their work, losing track of time as they work towards their goals.

3.2.2. Antflow

"Antflow" individuals are competitive, disciplined, and success-driven. They focus on achieving personal goals, often through hard work and by outcompeting others. The competitive environment fuels their actions, and they thrive on recognition and victory. While ants can contribute to societal growth through their determination and efficiency, their behavior can sometimes lead to intense rivalry and stress.

3.2.3. Leechflow

"Leechflow" individuals are driven by a desire for wealth, power, and personal gain. They often engage in exploitative behaviors, taking from others without giving back in equal measure. Unlike bees and ants, leeches operate primarily for their own

benefit, often disregarding the well-being of others. While there are professions that may inherently involve such exploitative tendencies, individuals in these roles can still choose to act ethically.

3.3. Tribefinder: Personality Feature Extraction

Griffin, our advanced personality extraction tool, plays a key role in linking textual data to profession predictions [19]. It breaks down text into seven categories—adding Ideology, Recreation, Personality, and Emotions to Groupflow and Alternative Realities—providing a deeper understanding of how personality traits influence career decisions (see Table 1). It also uses social network analysis [20] that can capture trends through a person's emails, although, for this research, we rely on its NLP capabilities to identify personas.

Table 1. Classifications of tribes and their characteristics.

Dimension	Tribe	Characteristics
Alternative Reality	Fatherlander	God, country, and tradition.
	Nerd	Technology, science, social inclusion, and globalization.
	Spiritualist	Contemplation and search for meaning.
	Treehugger	Protection of nature and sustainable growth.
Ideology	Liberalism	Focus on individual freedom.
	Capitalism	Minimal government intervention.
	Socialism	Greater government influence.
	Complainers	Constantly complain about everything.
Personality	Stock-Trader	Emphasis on short-term profit at the expense of long-term investment.
	Politician	Complex and evasive language rather than plain speaking.
	Journalist	Descriptive and generally more honest language.
	Risk-Takers	Language reflects daring decisions and behavior.
Recreation	Art	Art forms stimulate appreciation for beauty and passion.
	Fashion	Focus on popular trends and latest styles.
	Sport	Watching, attending, and playing sports.
	Travel	Experiencing different cultures and environments.
Groupflow	Beeflow	Collaborative creators who add value through innovation, art, or services.
	Antflow	Competitive and success-driven individuals focused on personal achievement.
	Leechflow	Exploitative individuals driven by personal gain, often at the expense of others.

3.4. Understanding the Interplay Between Profession and Personality

In our research, we recognize and understand how a person's profession is tied to their personality and values. By aligning career choices with individual traits, we can offer insights into job satisfaction, career planning, and professional development [21]. We have based our study on collecting data pertaining to groups of professions:

- Lawyer (0): Lawyers often exhibit a strong need for autonomy, logical thinking, and empathy, allowing them to excel in client relationships and complex cases [22]
- Medicine and Academics (1): In medicine and academia, personality traits such as conscientiousness, empathy, and a passion for knowledge are key. These traits drive individuals to be diligent, ethical, and focused on continuous learning and research [23].
- Sports (2): Athletes strive for recognition, with personality factors like emotional stability and extroversion strongly linked to high performance and success in competitive settings [24].
- Engineer (3): Engineers are characterized by a focus on practical, material outcomes and a preference for orderliness and objectivity [25]
- Creative (4): Associated with artistic expression, originality, and a passion for innovation and beauty. Creatives are significantly influenced by personality traits such as openness to experience, allowing for greater innovation and exploration [26]
- MBA (5): MBA performance is strongly influenced by personality traits such as extraversion, competitiveness, conscientiousness, and openness to experience. These traits, combined with a deep approach to learning, correlate significantly with higher academic achievement [27].
- Accountant (6): Value precision, security, and tradition, often driven by a desire for financial stability and order. They exhibit the ESTJ personality type, characterized by extraversion, sensing, thinking, and judging traits [28].

4. Dataset

To develop a robust model for profession prediction, we constructed a diverse and comprehensive dataset by scraping audio content from YouTube videos and podcasts across seven distinct professional categories: Lawyer, Medicine, Sports, Engineer, Creative, MBA, and Accountant. The audio content was transcribed into text using **Whisper AI** v3.0, an advanced automatic speech recognition (ASR) tool [29]. This transcription process yielded a total dataset of approximately **70,000 words**.

4.1. Data Collection Details

The data collection involved a variety of recordings across different professions. For the **Lawyer** category, 110 videos and podcasts were collected, covering legal arguments, courtroom practices, and career paths in law. The **Medicine** category consisted of 95 recordings, which included medical lectures, interviews with doctors, and discussions of healthcare trends. In the **Sports** category, 100 videos were gathered, featuring athlete interviews, sports commentary, and motivational content. The **Engineer** category included 85 recordings, which focused on engineering innovations, career advice, and technical discussions. For the **Creative** field, 75 podcasts were collected, discussing creative professions such as writing, music, and design. The **MBA** category had 90 discussions on business strategies, case studies, and management careers. Lastly, 60 videos related to the **Accountant** field were included, covering financial practices, accounting principles, and career paths in finance.

The transcriptions of these recordings varied in length, with each averaging between 500 to 700 words, depending on the content's duration. In total, the transcriptions amounted to over **70,000 words** of structured text.

The sources for these recordings included major platforms like YouTube, Spotify, Apple Podcasts, and other public podcast platforms. The topics discussed in each category were specific to the respective field: for Lawyers, they covered legal case studies, law school experiences, and courtroom strategies; for Medicine, they included advances in medicine, patient care, and medical education; Sports recordings focused on personal experiences of

athletes, training routines, and mental preparation; Engineers discussed industry trends, project management, and technical innovations; Creatives explored inspirations, processes, and challenges in creative fields; MBA discussions delved into business operations, market strategies, and leadership skills; and Accountants covered financial planning, tax strategies, and accounting tools.

4.2. Data Structuring

The collected text data were organized into **seven separate datasets**, each corresponding to one of the predefined professional categories. These datasets contained chunks of text paired with their respective labels, effectively capturing the nuances and specificities associated with each profession.

4.3. Thematic Headers and Extended Analysis

We utilized a fine-tuned architecture combining a pre-trained BERT model with a Bidirectional LSTM layer to enhance the contextual understanding of text data. The training process employed the Adam optimizer with a learning rate of 2×10^{-5}. The Adam optimizer was selected due to its robust performance in optimizing deep learning models, particularly in NLP tasks, as it adapts the learning rate for each parameter, combining the advantages of both RMSProp and momentum optimizers.

After training the model and generating predictions, we extended our analysis by calculating probabilities of belonging to a certain tribe using **Griffin 1.0** [20]. The resulting secondary dataset provided a broad spectrum of features categorized into thematic dimensions such as **Ideology**, **Recreation**, **Personality**, **Emotions**, **Groupflow**, and **Alternative Realities**. These thematic dimensions allowed us to explore deeper patterns and correlations between text-based personality features and professional classifications (see Table 1).

- **IDEOLOGY**: Capitalism, Complainers, Liberalism, Socialism;
- **RECREATION**: Arts, Fashion, Sport, Travel;
- **PERSONALITY**: Journalist, Politician, Risk-Taker, Stock-Trader;
- **EMOTIONS**: Anger, Fear, Happy, Sad;
- **GROUPFLOW**: Antflow, Beeflow, Leechflow;
- **ALTERNATIVE REALITIES**: Fatherlander, Nerd, Spiritualist , Treehugger.

This comprehensive dataset, combined with the subsequent analysis, forms the foundation of our research, providing the necessary data to train and evaluate our model for profession prediction from textual content.

5. Machine Learning Methodology

This section outlines the various steps involved in the process of text preprocessing, model training, and evaluation. We also provide a detailed comparison of different models employed, highlighting their performance based on several key metrics.

5.1. Text Preprocessing

The text data collected for this study underwent a thorough preprocessing pipeline to ensure consistency and quality. These processes aimed to clean and normalize the text, reducing noise and retaining meaningful features for model training.

Initially, the text was tokenized using the `nltk` library, splitting it into individual words or tokens. This was followed by the removal of common stopwords, such as “the” and “and”, which helped eliminate uninformative words that could introduce noise into the dataset. Then, lemmatization was applied to reduce words to their root forms, ensuring uniformity between variations of the same word. Finally, all punctuation marks were

removed to prevent them from influencing the model's understanding of the text. This systematic preprocessing ensured that the data were clean, standardized, and ready for further analysis.

5.2. Data Augmentation: Merging Sentences

To enhance the dataset and provide richer contextual information within training examples, multiple short sentences were merged into longer ones. This process was conducted by grouping sentences in batches of $n = 6$, chosen empirically to ensure a balance between adding context and maintaining sequence length compatible with the model's input constraints. The merging was performed separately within the dataset of each category to maintain contextual coherence and relevance, ensuring that the concatenated text was derived exclusively from sentences sharing the same category.

Each group of n sentences was concatenated into a single sequence, separated by spaces, and the resulting category was assigned based on the first sentence in the group, assuming uniformity within the batch.

5.3. Model Training and Evaluation

To predict professions based on textual data, we experimented with a range of models, progressing from classical machine learning approaches to advanced deep learning architectures. Initially, a Random Forest model was employed as a baseline due to its simplicity, interpretability, and effectiveness in handling structured data. However, its performance was limited when applied to the nuanced patterns in textual data, motivating the transition to deep learning approaches.

Next, DistilBERT was utilized for its efficiency and reduced computational overhead compared to BERT while retaining strong language representation capabilities. Although DistilBERT provided improved performance over Random Forest, the need for capturing sequential dependencies in text led to the adoption of a hybrid BERT + BiLSTM architecture. This model combines the contextual understanding of BERT with the sequential learning capabilities of BiLSTM, allowing it to better model the complex relationships in text relevant to profession classification.

5.3.1. Random Forest Classifier

The Random Forest model was applied as a baseline. Before training, we used the TF-IDF vectorizer to convert the text into numerical representations. The Random Forest was trained using categorical cross-entropy as the loss function, and feature importance was derived based on Gini impurity. The model's output for each class was calculated based on the following:

$$L = -\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{1}$$

where y_i represents the true label and $\hat{y}_i$ the predicted probability for class i.

5.3.2. DistilBERT Model

DistilBERT was used to capture the deep semantic relationships within the text. Each token in a sentence was embedded and the final output of the DistilBERT model was a sequence of hidden states H. These hidden states were then passed through a Softmax layer to predict the profession [30].

$$H = \text{DistilBERT}(T) \tag{2}$$

where T represents the tokenized text and H is the set of hidden states.

5.3.3. BERT with BiLSTM

The BERT model was combined with a BiLSTM layer to capture sequential dependencies alongside contextual information. The final embeddings from BERT were passed to a bidirectional LSTM, which produced an enhanced understanding of the sequence before the dense layer for classification.

$$H' = \text{BiLSTM}(\text{BERT}(T)) \tag{3}$$

where T represents the tokenized text and H is the output of the BiLSTM layer.

5.4. Analysis and Feature Importance

5.4.1. Correlation Analysis and Clustering

We performed correlation analysis to uncover relationships between features and professional categories. The correlation matrix C was computed using Pearson's correlation coefficient, which measures the linear relationship between variables. This method was chosen for its simplicity and effectiveness in identifying linear dependencies, making it suitable for the initial exploration of feature relationships.

An empirical assessment of the dataset led to the selection of a significant criterion of $\rho > 0.3$. It was found that the threshold struck a balance between identifying significant correlations and avoiding adding too much noise from weaker associations. To further facilitate data interpretation, hierarchical clustering algorithms were used to group features with similar patterns in the correlation matrix.

To group professions based on feature similarities, K-means clustering was employed. After scaling the data, the K-means algorithm partitioned the dataset into seven clusters, corresponding to the seven professional categories. This clustering approach provided insights into the natural groupings of professions, complementing the correlation analysis in identifying key features relevant to each cluster.

5.4.2. SHAP Values for Model Interpretability

After the formation of the dummy network and passing it through Griffin, we obtain values of different features for each entry. To better understand the predictions made by our models, we used SHAP (SHapley Additive exPlanations) values. The SHAP value ϕ_i for feature i is computed as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{4}$$

where N is the set of all features, and $f(S)$ is the model's prediction with feature subset S.

SHAP values ensure fairness and transparency in feature contribution by explaining each prediction at the granular level, in contrast to Random Forest's feature importance, which mainly represents the average effect of each feature across all predictions. This method works especially well for deciphering complicated models like BERT + BiLSTM, where feature interactions can have a big impact on predictions.

5.4.3. Feature Importance Using Random Forests

Feature importance was also derived from the Random Forest model, where the Gini importance score indicated the significance of each feature in predicting the profession. Top features were visualized using bar plots, providing insights into the driving factors for each profession [31].

$$\text{Importance}(X_i) = \sum_{t \in T} \Delta G_t \quad \text{for feature } X_i \tag{5}$$

where ΔG_t represents the reduction in Gini impurity at tree node t that splits on feature X_i.

5.4.4. Radar Charts for Feature Insights

To visualize the top features for each profession, radar charts were created. These charts highlighted the top 10 most important features for each profession, based on the Random Forest feature importance scores.

5.5. Evaluation Metrics

To evaluate the performance of the models, several metrics were employed. Accuracy was used to measure the proportion of correct predictions out of all predictions made. It is calculated as Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$, where TP represents true positives, TN true negatives, FP false positives, and FN false negatives.

Precision, defined as Precision $= \frac{TP}{TP+FP}$, quantifies the number of correct positive predictions among all positive predictions made by the model. Recall (or sensitivity), calculated as Recall $= \frac{TP}{TP+FN}$, measures the number of correct positive predictions out of all actual positive instances.

Lastly, the F1 score, which provides a balance between precision and recall, is computed as the harmonic mean of the two using the formula $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

Since the dataset contains seven classes, these metrics were adapted for multi-class classification using a one-vs.-rest (OvR) approach. For each class, the model treats it as the positive class while considering all other classes as negative. Precision, recall, and F1 score were then computed for each class independently, and the final metric values were obtained by averaging across all classes. Macro averaging was employed to give equal weight to each class, ensuring that smaller classes were not overshadowed by larger ones. This adaptation ensures a comprehensive evaluation of the model's predictive performance across all classes.

6. Model Architecture

The model architecture used in this study is designed to effectively capture both the semantic and sequential nature of textual data. To achieve this, we employed a hybrid architecture consisting of the **BERT** (Bidirectional Encoder Representations from Transformers) model and a **Bidirectional Long Short-Term Memory** (BiLSTM) layer.

Adding a BiLSTM layer enables enhanced modeling of temporal patterns and relationships within the contextual embeddings generated by BERT. This hybrid approach leverages the strengths of both transformer-based contextualization and recurrent sequential processing, as supported by recent studies [32].

In this study, we utilized a pre-trained BERT model, specifically the `bert-base-uncased` variant, for feature representation. Pre-trained BERT models are widely recognized for their ability to provide high-quality contextual embeddings, which are derived from large-scale unsupervised pretraining on extensive corpora.

6.1. BERT: Contextual Embeddings

The BERT model is the core component of our architecture, responsible for generating rich, contextual embeddings of the input text. BERT, or Bidirectional Encoder Representations from Transformers, processes input sentences as sequences of tokens and leverages a transformer-based architecture to capture the relationships between these tokens. One of the key advantages of BERT is its ability to model bidirectional context, meaning it considers both the preceding and succeeding tokens when encoding each word, allowing for more nuanced representations of the language [11].

BERT uses a subword tokenization method, such as WordPiece, which splits words into smaller units and assigns each unit a unique identifier. For example, a sentence

$s = \{x_1, x_2, \ldots, x_n\}$, where each x_i is a token, is tokenized into subword units that are then embedded into vectors.

Once the sentence is tokenized, BERT processes the input and produces a sequence of hidden states $H = \{h_1, h_2, \ldots, h_n\}$, where each h_i represents the context-aware embedding of token x_i. These hidden states encapsulate the meaning of each token in relation to the entire sentence and are computed as follows:

$$H = \text{BERT}(T) \tag{6}$$

where T is the tokenized input sequence, and H is the corresponding set of embeddings. The embeddings serve as the foundational representations that are further refined by the subsequent layers of the model.

6.2. BiLSTM: Sequential Modeling

While BERT captures context on a global scale, we introduced a Bidirectional LSTM (BiLSTM) layer to model the sequential dependencies in the token embeddings. The BiLSTM processes the sequence of embeddings in both forward and backward directions, allowing the model to capture information from both past and future tokens [33].

The hidden states from BERT $H = \{h_1, h_2, \ldots, h_n\}$ are passed into the BiLSTM, which generates the transformed hidden states $H' = \{h'_1, h'_2, \ldots, h'_n\}$. These transformed hidden states encapsulate the temporal relationships between tokens, providing a more robust representation of the input.

$$H' = \text{BiLSTM}(H) \tag{7}$$

6.3. Classification Layer

The output from the BiLSTM is passed into a fully connected classification layer with a Softmax activation function. This layer produces the final probabilities for each profession class. Given the transformed hidden states H', the classification layer computes the probability distribution over the profession classes:

$$P(y|H') = \text{softmax}(W \cdot H' + b) \tag{8}$$

where W is the weight matrix, b is the bias term, and y is the predicted profession class.

6.4. Number of Classification Labels

The model is trained to classify text into one of seven distinct professional categories, each corresponding to a specific profession. These categories include Lawyer (Class 0), Medicine and Academics (Class 1), Sports (Class 2), Engineer (Class 3), Creative (Class 4), MBA (Class 5), and Accountant (Class 6). The classification task is implemented as a seven-way classification problem, where the Softmax layer outputs the probability distribution across these classes for a given input. The highest probability value determines the predicted profession, effectively leveraging the model's understanding of linguistic and contextual features to categorize the text.

6.5. Model Training

The model was trained using the categorical cross-entropy loss function, defined as follows:

$$L = -\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{9}$$

where y_i is the true label for class i and $\hat{y}_i$ is the predicted probability for class i. The model was optimized using the Adam optimizer, with a learning rate of 2×10^{-5}, and trained for three epochs with a batch size of 16.

6.6. Summary of the Architecture

The final architecture of the model comprises several layers, each designed to perform a specific role in processing and classifying the input text. The first component is the **Input Layer**, which takes tokenized input sequences prepared during preprocessing. These tokenized inputs are passed to the **BERT Layer**, which generates rich, contextual embeddings for each token by leveraging bidirectional context from the entire input sequence. The contextual embeddings are then processed by the **BiLSTM Layer**, which captures sequential dependencies and relationships between tokens, enhancing the model's understanding of the temporal structure within the text. Finally, a **Dense Layer with Softmax** computes the probability distribution over the seven profession classes, allowing the model to assign the most appropriate profession to the input based on its learned features. The overall architecture is illustrated in Figure 1.The model's predictions are evaluated based on confidence scores assigned to each profession class. Table 2 presents example sentences with the top two predicted profession categories and their corresponding probabilities.

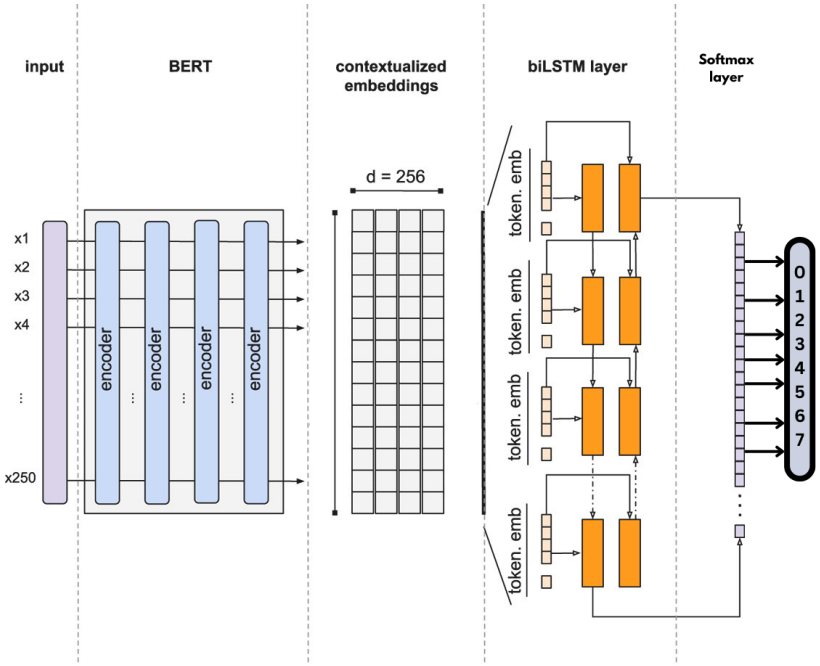

Figure 1. Final model architecture.

Table 2. Top 2 predicted probabilities for example sentences.

Example Sentence (Shortened)	Top Class (Probability)	Second Class (Probability)
"Singing is my passion. I love drawing and painting".	Creative (0.6758)	Medicine (0.1564)
"Court overturned lower court citing procedural errors".	Lawyer (0.9407)	Accountant (0.0191)
"Team showed resilience to secure victory in the final moments".	Sports (0.6649)	Engineer (0.0993)
"I excel in budgeting, auditing, and financial management".	MBA (0.7305)	Accountant (0.1756)

7. Results

This section presents a detailed analysis of how various features contribute to professional classifications, based on correlations, hierarchical clustering, feature importance, and model interpretability through SHAP values.

7.1. Unsupervised Trend Detection of Dataset

In our study, we explored the application of unsupervised learning methods to detect trends across different professions. The core idea was to utilize sentence embeddings generated by the Sentence Transformer model and apply clustering techniques to uncover hidden patterns in the text data [34]. This approach allowed us to group similar professions based on the textual content and identify the most frequent topics or trends in each cluster.

7.2. Data Preparation and Embedding Generation

The text data were sourced from multiple professions, including Lawyers, Medicine and Academics, Sports, Engineers, Creatives, MBA, and Accountants. After preprocessing the text, which included tokenization, stopword removal, and lemmatization, we used the Sentence Transformer model, specifically "paraphrase-MiniLM-L6-v2", to generate embeddings for each sentence.

$$E_i = \text{SentenceTransformer}(T_i) \tag{10}$$

where T_i is the preprocessed text and E_i is the corresponding sentence embedding. These embeddings serve as a dense representation of the text, capturing both semantic meaning and contextual information.

7.3. Clustering Approach

We applied K-means clustering on the embeddings to detect underlying trends within the text data. The choice of the number of clusters K was determined empirically, with $K = 7$ representing the seven major professional categories. The K-means algorithm grouped similar sentences into distinct clusters, with each cluster representing a collection of professions or topics with high similarity in text content. The resulting clusters are visualized in Figure 2.

$$C = \arg\min_{S} \sum_{i=1}^{n} ||E_i - \mu_S||^2 \tag{11}$$

where C represents the cluster, E_i is the embedding, and μ_S is the cluster centroid.

Figure 2. Clusters using K-means.

7.4. Model Comparison

The performance of the models is summarized in Table 3. Each model was evaluated based on its accuracy, precision, recall, and F1 score on the test set.

Table 3. Model performance comparison.

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.78	0.76	0.75	0.75
DistilBERT	0.85	0.85	0.80	0.77
BERT + BiLSTM	0.88	0.87	0.83	0.82

Evaluating the different machine learning and deep learning models, we found that advanced models such as BERT with BiLSTM layers significantly outperformed traditional models such as Random Forest in the task of profession prediction. The best-performing model, BERT + BiLSTM, achieved an F1 score of 0.82, highlighting its ability to capture both contextual and sequential dependencies in text data.

7.5. Trend Detection in Clusters

After clustering the text, we analyzed the most frequent words in each cluster to uncover key topics and trends. The law-based cluster prominently featured terms such as **"court"**, **"law"**, and **"justice"**, reflecting discussions centered around legal concepts and practices. Similarly, the MBA and accountant clusters were dominated by words like **"stock market"**, **"investment"**, and **"profit"**, highlighting themes of finance and business. These clusters effectively captured the professional focus and vocabulary relevant to these fields.

The sports-based group emphasized terms such as **"competition"**, **"game"**, and **"training"**, highlighting physical activity and performance-related themes. On the other hand, the creatives-based cluster highlighted words like **"design"**, **"art"**, and **"innovation"**, aligning with discussions on artistic and creative pursuits. This cluster analysis provided a comprehensive understanding of the dominant topics in different professions, offering insights into how language reflects professional contexts.

7.6. Visualization and Trend Analysis of Final Dataset

To visualize the clusters, we reduced the dimensionality of the sentence embeddings using Principal Component Analysis (PCA). The two-dimensional representation of the clusters helped in understanding the relationships between different professions. Additionally, the cosine similarity between cluster centroids and predefined profession embeddings helped further refine our understanding of which clusters aligned with which professions.

The top trends for each cluster were analyzed based on the cosine similarity between cluster centroids and professional embeddings, showing a strong correlation between text features and profession classification.

7.6.1. Correlation Insights

Our correlation analysis revealed meaningful relationships between personal traits and professional identities. We focus on highly correlated characteristics (threshold $\rho > 0.3$) for each profession, as summarized in Table 4:

These correlations highlight strong links between individual traits such as recreation, ideology, and groupflow with specific professions.

Table 4. Correlation insights across professional identities.

Profession	Correlated Feature	ρ	Interpretation
Lawyer	Alternative Realities Fatherlander	+0.37	Structured, traditional ideologies
	Professional Complexity	−0.40	Nuanced professional identity
Medicine & Academic	Recreation Travel	+0.31	Intellectual curiosity, global perspective
Sports	Recreation Sport	+0.61	Strong link to physical activity
	Groupflow Antflow	+0.60	Collaborative behavioral patterns
	Personality Risk-Taker	+0.35	Propensity for risk and challenge
Creative	Ideology Complainers	+0.34	Critical and analytical mindset
MBA & accountant	Groupflow Leechflow	+0.35	Competitive professional dynamics
	Personality Stock-Trader	+0.35	Profit-driven professional approach
	Professional Competitiveness	+0.36	Strategic career orientation

7.6.2. Hierarchical Clustering

We applied hierarchical clustering to understand how different professional categories share common feature profiles [35]. The resulting dendrogram revealed several meaningful clusters. **Cluster 1** includes **Medicine**, **MBA**, and **Accountant**, which are professions that share traits related to financial and managerial skills. **Cluster 4** groups **Lawyer** with traits such as *Politician Persona* and *Anger Emotions*, suggesting that lawyers often manage emotionally charged situations with structured, political thinking. In **Cluster 6**, **Engineers** and **Creatives** are clustered together, reflecting a shared tendency for ideological dissent, described as *Complainers Ideology*. Finally, **Cluster 9** focuses on **Sports Professionals**, who are grouped based on their engagement in physical activity (*Recreation Sports*) and their collaborative approach (*Antflow*).

The clusters reveal shared personal traits between seemingly different professions, driven by common ideologies and emotional preferences.

7.6.3. Feature Importance

We analyzed feature importance using a Random Forest model to rank the most influential features for each profession [36]. This analysis provided insights into the primary characteristics that define each professional category. For the **Lawyer** profession, the top feature was *Alternative Realities Fatherlander* (importance: 0.17), indicating a connection to traditional, structured thinking. Other important features included *Ideology Liberalism*, *Recreation Arts*, and *Emotion Anger*, highlighting both creative and emotional preferences within the legal profession (see Figure 3).

For **Medicine and Academics**, the most significant feature was *Recreation Travel* (importance: 0.15), emphasizing the intellectual curiosity and global mindset common in academia. Additional features like *Groupflow Beeflow*, *Ideology Socialism*, and *Personality Stock-Trader* suggested exploratory, collaborative and spontaneous behaviors prevalent in academic and medical professionals (see Figure 4).

In the **Sports** profession, *Recreation Sport* dominated the feature importance (importance: 0.52), followed by *Groupflow Antflow* and *Personality Risk-Taker*, which emphasized the physical and competitive nature of athletes (see Figure 5).

For **Engineers**, the most significant feature was *Personality Stock-Trader* (importance: 0.08), reflecting their analytical and strategic approach to problem-solving. Features like *Groupflow Beeflow*, *Ideology Capitalism*, and *Alternative Realities Treehugger* highlight their collaborative, practical, and innovative traits, while *Groupflow Antflow* and *Emotions Fear*

emphasize adaptability and cautious decision-making common in the engineering field (see Figure 6).

For **Creative** professionals, the most significant feature was *Ideology Complainers* (importance: 0.16) highlighting the critical and introspective nature of creative professionals. *Groupflow Beeflow* and *Recreation Fashion* were the other important features obtained in the process (see Figure 7).

Finally, the **MBA** and **Accountant** professions were strongly influenced by features such as *Groupflow Leechflow* and *Personality Stock-Trader*, indicating competitive, financially driven behaviors that are characteristic of these fields (see Figures 8 and 9).

This feature importance analysis provided a clear understanding of how various traits, such as recreation and personality, contribute to the likelihood of belonging to a particular profession.

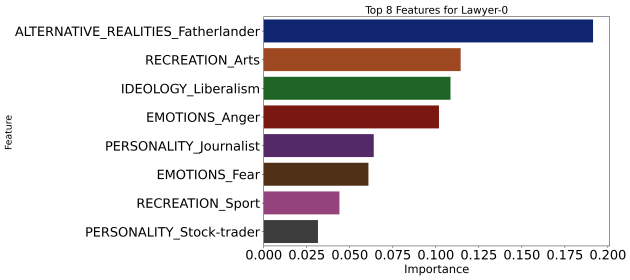

Figure 3. Feature importance bar plot for Lawyer.

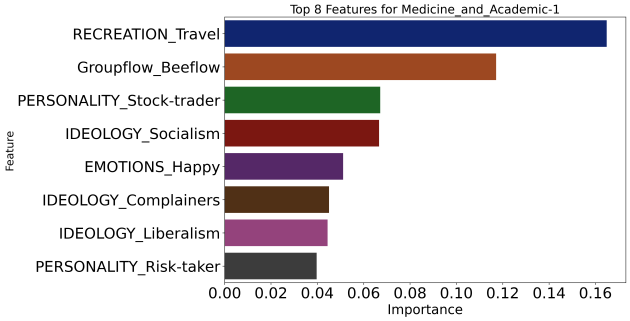

Figure 4. Feature importance bar plot for Medicine and Academics.

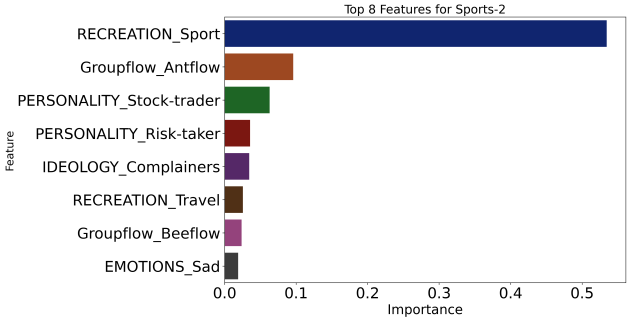

Figure 5. Feature importance bar plot for Sports.

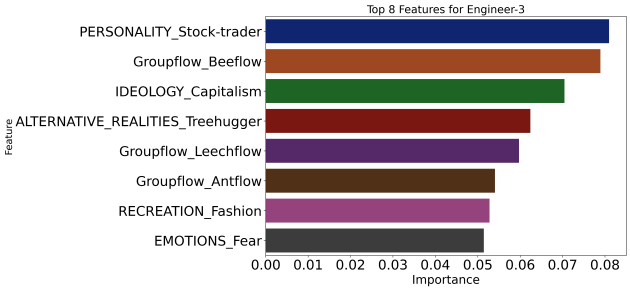

Figure 6. Feature importance bar plot for Engineer.

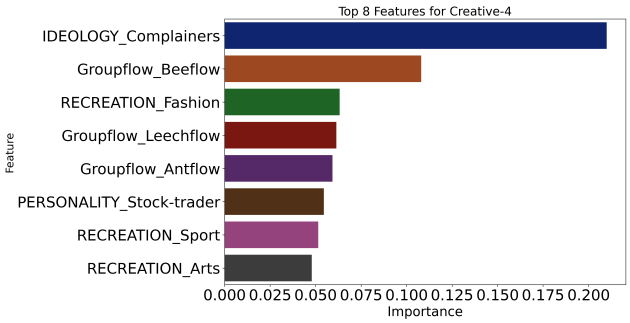

Figure 7. Feature importance bar plot for Creative.

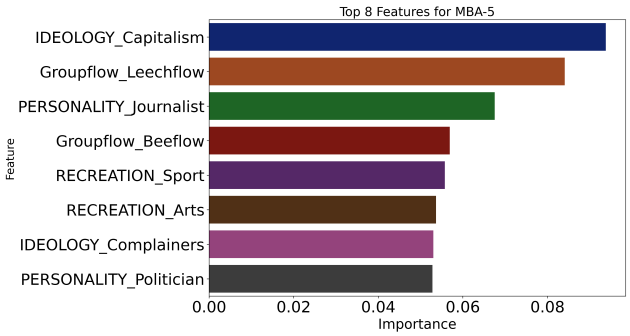

Figure 8. Feature importance bar plot for MBA.

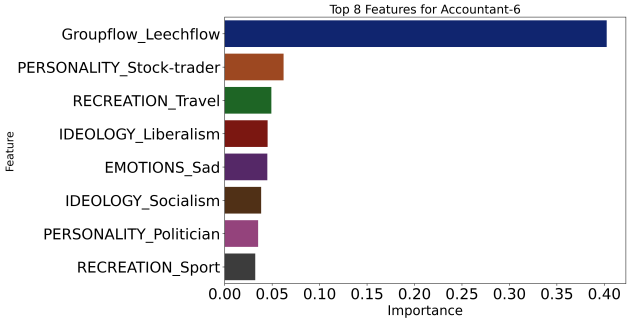

Figure 9. Feature importance bar plot for Accountant.

7.6.4. Radar Charts for Feature Visualization

To better visualize the distribution of features across professions, we generated radar charts for each professional category. These charts allowed for a comparative analysis of the most influential features in each profession, as shown in Figures 10–16. The insights derived from the radar charts include the following:

IDEOLOGY: A diverse range of ideological preferences was observed across professions. For instance, **Liberalism** dominated in **Lawyer-0** (Figure 10), while **Capitalism** was prominent in both **MBA-5** and **Engineer-3** (Figures 15 and 13), and **Socialism** was

more common in **Medicine and Academic-1** (Figure 11). The *Complainers ideology* was particularly evident in **Creative-4** (Figure 14) and also present in **MBA-5**, indicating a critical mindset. Interestingly, the *Fatherlander ideology* was uniquely strong in **Lawyer-0**, suggesting a patriotic inclination.

RECREATION: **Travel** emerged as a common recreational interest, particularly prominent in **Medicine and Academic-1** (Figure 11), but also present in several other profiles. **Sports** was overwhelmingly dominant in **Sports-2** (Figure 12), as expected, and moderately present in **MBA-5**. **Arts** was uniquely strong in **Lawyer-0**, suggesting a cultural interest within the legal profession.

PERSONALITY: The *Stock-Trader personality* was common across multiple professions, with a particularly strong presence in **Accountant-6** (Figure 16). The *Journalist trait* was notable in both **Lawyer-0** and **MBA-5**, highlighting strong communication skills. The *Risk-Taker personality* was uniquely present in **Sports-2**, aligning with the competitive nature of athletes.

EMOTIONS: Emotional traits were not consistently represented across all professions. *Anger* was particularly prominent in **Lawyer-0** (Figure 10), potentially reflecting the adversarial nature of the profession. The *Happy emotion* was uniquely strong in **Medicine and Academic-1** (Figure 11), suggesting job satisfaction in these fields. *Sadness* appeared in **Accountant-6**, albeit at a low level.

GROUPFLOW: *Leechflow* was notably dominant in **Accountant-6** and significant in **MBA-5** (Figures 16 and 15), possibly indicating a tendency to leverage others' work. *Beeflow* was prominent in **Medicine and Academic-1** and present in **Creative-4** (Figures 11 and 14), suggesting a tendency for collaboration. *Antflow* appeared in both **Creative-4** and **Sports-2**, which may point to individualistic or contrarian behaviors in these professions.

ALTERNATIVE REALITIES: *Treehugger* was strongly present in **Engineer-3** (Figure 13), uniquely combining with a capitalist ideology. The *Fatherlander* ideology was also significant in **Lawyer-0**, but not prominently represented in other professions. This category was not consistently represented across all professional profiles.

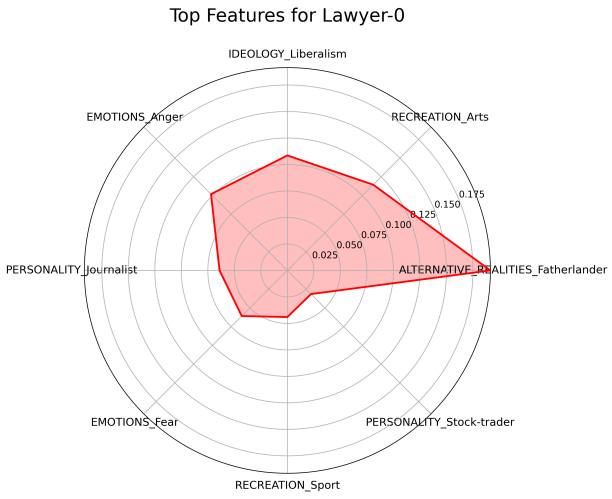

Figure 10. Radar chart for Lawyer.

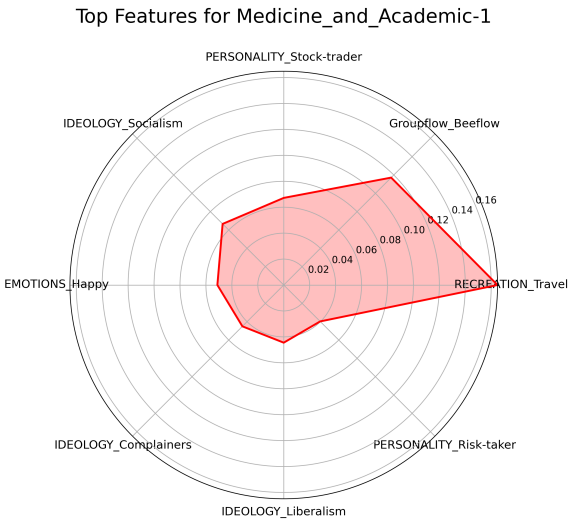

Figure 11. Radar chart for Medicine and Academics.

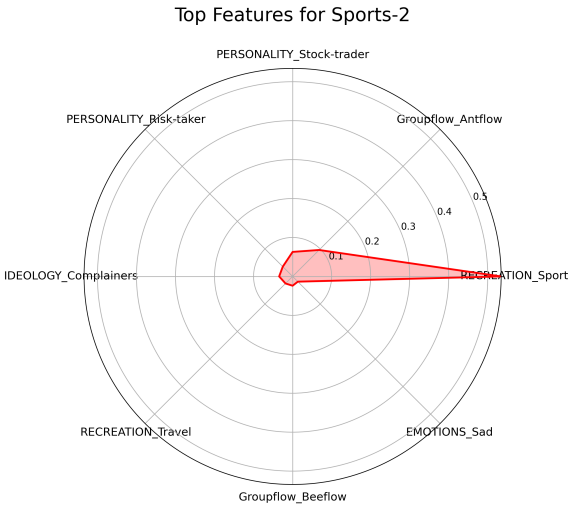

Figure 12. Radar chart for Sports.

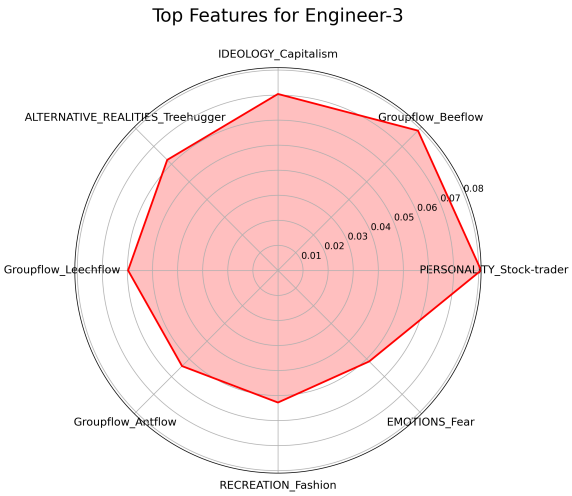

Figure 13. Radar chart for Engineer

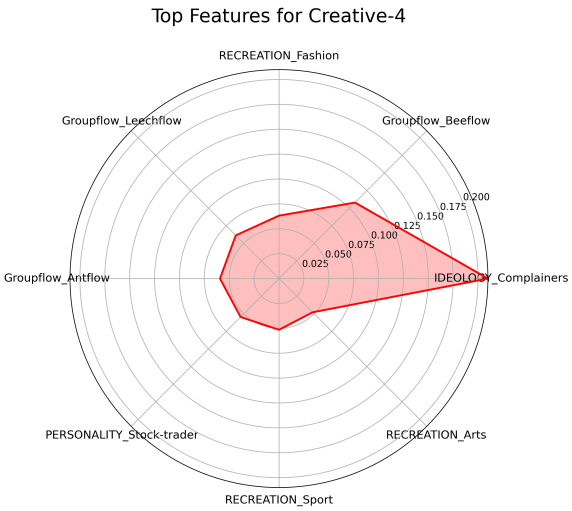

Figure 14. Radar chart for Creatives.

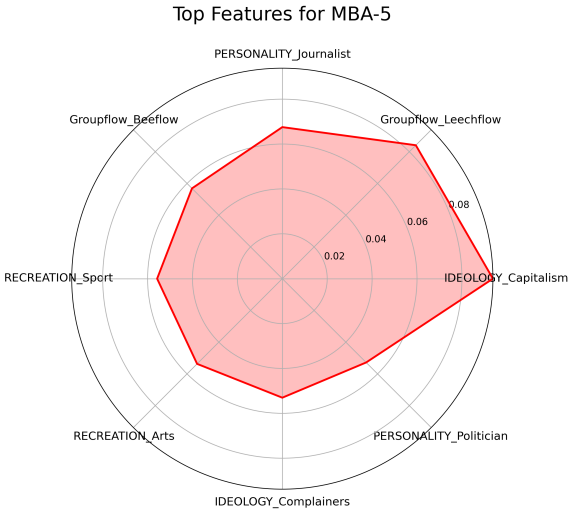

Figure 15. Radar chart for MBA.

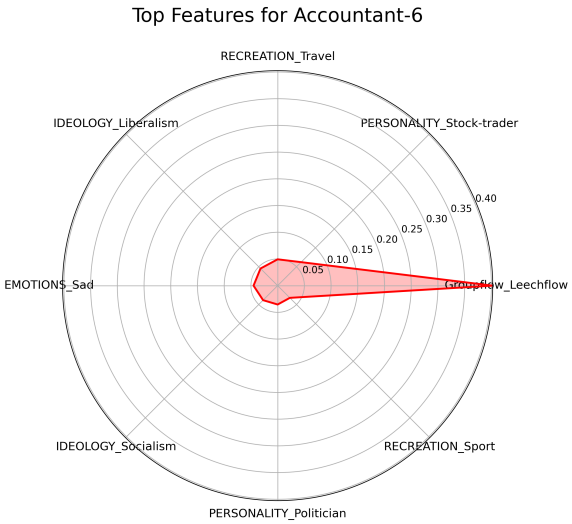

Figure 16. Radar chart for Accountant.

7.6.5. Unique Inferences

Several unique inferences were drawn from the analysis of the professional profiles. The **Lawyer-0** profile exhibits an interesting combination of **liberal ideology**, **artis-**

tic interests, and strong patriotic tendencies, particularly with a *Fatherlander* inclination. The **Engineer-3** profile is unique for blending a **capitalist ideology** with strong environmentalist traits, notably a *Treehugger* mentality. The **Creative-4** profile shows a higher tendency toward *complaining* and a *beeflow* groupflow, which may reflect the nature of creative work, which often involves long periods of focused effort. In the case of **Accountant-6**, an extremely high *Leechflow* score stands out, suggesting a strong tendency to rely on others' work or resources, much higher than any other trait across the profiles. The **Sports-2** profile is highly focused on sports-related traits, with little variation in other areas, indicating a highly specialized focus. Lastly, the **MBA-5** profile strikes a balance between **capitalist ideology**, a keen interest in **sports**, and a combination of traits such as *journalist* and *politician*, which likely reflects the diverse skill set required for success in business administration.

7.6.6. SHAP Value Interpretation

To improve interpretability, we employed SHAP (SHapley Additive exPlanations) values [37], which allowed us to break down the contributions of each feature to individual predictions. SHAP values revealed the specific traits that influenced the model's decision-making process for each profession.

Table 5 presents the top SHAP feature contributions for each profession. For **Sports**, SHAP values indicated that *Recreation Sport* and *Groupflow Antflow* had a positive contribution to the predictions, confirming that physical activity and collaboration are central to athletes. In the case of **Medicine and Academics**, *Recreation Travel* and *Personality Stock-Trader* were key contributors, suggesting a mix of exploratory and managerial traits that align with academia and healthcare professions. For **Lawyer**, SHAP values highlighted the importance of *Alternative Realities Fatherlander* and *Personality Politician*, aligning with the structured, politically oriented behavior often seen in legal professionals. By leveraging SHAP values, we provided a transparent view of how individual features influenced the model's predictions, enhancing interpretability and trust in the decision-making process.

Table 5. Top SHAP feature contributions for each profession.

Profession	Feature 1 (Value)	Feature 2 (Value)	Feature 3 (Value)
Lawyer	Arts (0.0299)	Fatherlander (0.0268)	Anger (0.0245)
Medicine & Academic	Travel (0.0895)	Beeflow (0.0507)	Stock-Trader (0.0195)
Sports	Sport (0.1501)	Antflow (0.0672)	Complainers (0.0368)
Engineer	Capitalism (0.0228)	Beeflow (0.0215)	Fashion (0.0177)
Creative	Complainers (0.0426)	Fashion (0.0271)	Travel (0.0241)
MBA	Politician (0.0342)	Leechflow (0.0323)	Complainers (0.0310)
Accountant	Leechflow (0.1320)	Stock-Trader (0.0263)	Sad (0.0165)

8. Discussion

This study examined how natural language processing (NLP) and machine learning (ML) models can predict professions by analyzing text, specifically focusing on how different personality traits and value systems, which we term **Alternative Realities**, are reflected in professional personas. By leveraging deep learning models such as BERT and BiLSTM, we were able to categorize professions like Lawyers, Engineers, and Sports professionals with a high degree of accuracy.

While there are many personality characteristics frameworks, most prominently the MBTI and FFI [6,15], they all are of limited usefulness in helping people find their ideal profession. The purpose of introducing an additional features layer with alternative realities, groupflow, personalities, and recreational features is to give the professional job seeker additional assistance in discovering the profession ideally suited to their personality

characteristics. This paper describes first steps for mapping the new features framework introduced here to prominent job categories. By being told these characteristics based on their language, job seekers obtain valuable additional input to make this life-course-influencing decision.

8.1. Alternative Realities: Personalities and Values

The concept of **Alternative Realities** plays a crucial role in understanding the intersection between personality traits and professional identity. We classified individuals into four broad categories, each representing different worldviews, values, and behaviors.

The **Fatherlander** category includes individuals who exhibit a deep belief in tradition, nation, and family. These individuals uphold the values of the "good old times" and view their fatherland as superior. Such personalities tend to align with professions that emphasize order, authority, and the preservation of cultural values, such as lawyers, military leaders, or politicians. Our models detected this strong traditionalism in legal professions, suggesting a correlation between conservative values and structured, rule-driven jobs.

The **Nerd** category is characterized by a belief in progress, science, and technology as forces for good. Nerds often aspire to transcend human limitations and are enthusiasts of global connectivity and advancements like space exploration. These individuals tend to thrive in professions related to engineering, technology, and scientific research. The BERT + BiLSTM model effectively captured these traits in Engineers and other tech-centric professions.

The **Spiritualist** group seeks meaning through subjective experiences of the sacred. Their behaviors are driven by a quest for spiritual fulfillment and contemplation. Professions that align with these values may include religious leaders, yoga instructors, or philosophers. While this group was more challenging to capture with traditional NLP models, future work could focus on refining models to detect the subtle language of spiritual guidance and contemplation.

The **Treehugger** category includes individuals who advocate for sustainability and environmental preservation. They challenge certain technological advancements like genetic manipulation while supporting others, such as alternative energy sources. Their value system often conflicts with industrial or corporate norms, leading them to professions in environmental activism, sustainability consulting, or conservation. Our model detected some alignment between these values and professions in academia or NGOs focused on sustainability.

8.2. Groupflow: Collaborative and Competitive Dynamics

In addition to **Alternative Realities**, we introduced the concept of **Groupflow**, which describes how individuals engage in teamwork and their behavioral dynamics in professional settings. These categories are crucial for understanding professional performance and interpersonal relationships within organizations.

Beeflow refers to individuals who are collaborative creators. They focus on creating value for both themselves and society. This profile aligns with professions in creative industries, research, and innovation. Beeflow members experience a state of flow while working, indicating high levels of engagement and satisfaction in collaborative environments.

Antflow describes competitive, disciplined individuals who are driven by personal goals and hard work. Professions such as athletes or business executives often display these traits, thriving in environments where success is measured by individual accomplishments and victories. Our models found strong correlations between Antflow behaviors and professions in sports and competitive business environments.

Leechflow individuals are characterized by exploitative tendencies, benefiting themselves often at the expense of others. They may be found in roles that allow for opportunism or manipulation of systems for personal gain, such as high-risk stock trading or certain management positions. This group was more difficult to define within traditional professional categories but could be inferred through language patterns associated with competitive and self-serving behaviors.

8.3. Recreation and Personality Traits: Enhancing Profession Classification

In addition to the **Alternative Realities** and **Groupflow** models, we explored how recreational interests and personality traits influence professional identities.

Recreation interests in **Art**, **Fashion**, **Sport**, and **Travel** reveal much about an individual's professional alignment. For instance, professionals in the creative industries were frequently linked with interests in art and fashion, while individuals in travel-related professions showed a strong interest in global exploration and cross-cultural experiences.

Personality Traits such as being a **Risk-Taker**, **Journalist**, or **Politician** also correlated with professional categories. For example, risk-takers were prevalent in high-risk, high-reward professions like sports and stock trading, while journalists and politicians exhibited traits aligned with professions involving communication and influence.

8.4. Model Performance and Limitations

The BERT + BiLSTM model showed robust performance in predicting professions based on text data, particularly when aligned with distinct personality categories. Professions like Sports, Engineering, and Law were consistently predicted with high accuracy, showcasing the model's ability to capture both contextual and sequential information. However, some professional categories, such as MBAs and Accountants, exhibited overlapping language patterns, which reduced the model's ability to distinguish between them. The model faced challenges in capturing subtle nuances associated with personality traits or alternative realities like Spiritualist behaviors, which are often implicit and context-dependent.

The dataset, while comprehensive, has potential biases stemming from its reliance on publicly available content from platforms like YouTube and podcasts, which may introduce selection bias. Additionally, the English-centric nature of the data limits their applicability to non-English contexts. Minor transcription errors from Whisper AI, particularly for domain-specific terms, could also have affected data quality. Addressing these biases through more diverse sources and broader demographic representation would improve the model's generalizability.

8.5. Future Directions

This study opens the door to further research on the relationship between personality and professional identity. Future work could explore the incorporation of more advanced personality models and extend the dataset to include a wider range of professions. Additionally, exploring more refined group behaviors within organizations through **Groupflow** could provide deeper insights into how personality traits influence team dynamics and professional success.

This work has wide practical applicability. For instance, based on the words people use in their everyday interactions, for example, in email and chat, an NLP analysis system could make recommendations for suitable jobs for a job seeker. Such a system could be put at the disposal of individuals to better realign their professional interests with their personalities by asking them to write job-related stories which are then used as input for a system like the one described in this paper to make referrals on available job opportunities inside and outside the company. Such a system could also be extended

to include recommendations derived from standardized job classifications such as the International Standard Classification of Occupations (ISCO) to give a prospective job seeker broad access to the many different professions available.

Alternatively, corporate HR managers could identify promising candidates based on their language in emails. Of course, all of these applications of AI in the HR context have to be carefully calibrated to respect privacy and adhere to the highest ethical standards.

Overall, our findings highlight the complex interplay between individual traits, societal values, and professional identities, offering a new lens through which to view career prediction and development.

9. Conclusions

Throughout this paper, we were able to uncover key insights about how specific personality traits and ideological preferences contribute to professional identity. Our findings underscore the complex interplay between individual characteristics and profession, with certain traits consistently aligning with particular professional categories. This multi-faceted analysis not only improves model accuracy but also provides a deeper understanding of how individual attributes influence career outcomes.

Author Contributions: Conceptualization, P.A.G.; methodology, P.A.G. and A.J.; software, A.J.; validation, A.J.; formal analysis, A.J. and P.A.G.; data curation, A.J.; writing—original draft preparation, A.J.; writing—review and editing, P.A.G.; visualization, A.J. and P.A.G.; supervision, P.A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available because of copyright restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, H.; Chatterjee, I.; Zhou, M.; Lu, X.S.; Abusorrah, A. Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 1358–1375.
2. Gloor, P.A. *Happimetrics: Leveraging AI to Untangle the Surprising Link Between Ethics, Happiness and Business Success*; Edward Elgar Publishing: Cheltenham, UK, 2022.
3. Pradhan, T.; Bhansali, R.; Chandnani, D.; Pangaonkar, A. Analysis of personality traits using natural language processing and deep learning. In Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 15–17 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 457–461.
4. Gloor, P.; Fronzetti Colladon, A.; Grippa, F. Measuring ethical behavior with AI and natural language processing to assess business success. *Sci. Rep.* **2022**, *12*, 10228.
5. Woods, S.A.; Hampson, S.E. Predicting adult occupational environments from gender and childhood personality traits. *J. Appl. Psychol.* **2010**, *95*, 1045.
6. Floricia, C.M.; Luminita, S.M.; Filotia, S. The influence of personality features in choosing the profession. *Tech. Soc. Sci. J.* **2021**, *25*, 447.
7. Eakman, A.M.; Eklund, M. The relative impact of personality traits, meaningful occupation and occupational value on meaning in life and life satisfaction. *J. Occup. Sci.* **2012**, *19*, 165–177.
8. Csikszentmihalyi, M.; Csikszentmihalyi, M.; Abuhamdeh, S.; Nakamura, J. Flow. *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*; Springer Science+Business Media: Dordrecht, The Netherlands, 2014; pp. 227–238.
9. Sun, X.; Liu, B.; Cao, J.; Luo, J.; Shen, X. Who am I? Personality detection based on deep learning for texts. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
10. Koroteev, M.V. BERT: A review of applications in natural language processing and understanding. *arXiv* **2021**, arXiv:2103.11943.
11. Jain, D.; Kumar, A.; Beniwal, R. Personality bert: A transformer-based model for personality detection from textual data. In Proceedings of the International Conference on Computing and Communication Networks: ICCCN 2021, Athens, Greece, 19–22 July 2021; Springer: Singapore, 2022; pp. 515–522.

12. Rahman, A.U.; Al-Obeidat, F.; Tubaishat, A.; Shah, B.; Anwar, S.; Halim, Z. Discovering the correlation between phishing susceptibility causing data biases and big five personality traits using C-GAN. *IEEE Trans. Comput. Soc. Syst.* **2022**, *11*, 4800–4808.
13. KN, P.K.; Gavrilova, M.L. Latent personality traits assessment from social network activity using contextual language embedding. *IEEE Trans. Comput. Soc. Syst.* **2021**, *9*, 638–649.
14. VVR, M.R.; Silpa, N.; Gadiraju, M.; Reddy, S.S.; Bonthu, S.; Kurada, R.R. A plausible RNN-LSTM based profession recommendation system by predicting human personality types on social media forums. In Proceedings of the 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 23–25 February 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 850–855.
15. Fritsch, M.; Rusakova, A. Personality Traits, Self-Employment, and Professions 2010. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1736576 (accessed on 17 December 2024).
16. Gloor, P.; Colladon, A.F.; de Oliveira, J.M.; Rovelli, P. Put your money where your mouth is: Using deep learning to identify consumer tribes from word usage. *Int. J. Inf. Manag.* **2020**, *51*, 101924.
17. Deshpande, M.S. History of the Indian Caste System and Its Impact on India Today. Bachelor's Thesis, California Polytechnic State University, San Luis Obispo, CA, USA, 2010.
18. Gloor, P. Chapter 6: Beeflow, antflow and leechflow. In *Happimetrics*; Edward Elgar Publishing: Cheltenham, UK, 2022; pp. 68–79. https://doi.org/10.4337/9781803924021.00012.
19. Gloor, P.A.; Fronzetti Colladon, A.; de Oliveira, J.M.; Rovelli, P.; Galbier, M.; Vogel, M. Identifying tribes on twitter through shared context. In *Collaborative Innovation Networks: Latest Insights from Social Innovation, Education, and Emerging Technologies Research*; Springer Nature: Cham, Switzerland, 2019; pp. 91–111.
20. Altuntas, E.; Gloor, P.A.; Budner, P. Measuring ethical values with AI for better teamwork. *Future Internet* **2022**, *14*, 133.
21. Baral, G. Delving into the Happiness of Professional Accountants: Examining the Interplay between Personality Traits, and Job and Life Satisfaction. *İktisadi İdari Ve Siyasal Araştırmalar Derg.* **2024**, *9*, 727–739.
22. Daicoff, S. Lawyer, know thyself: A review of empirical research on attorney attributes bearing on professionalism. *Am. UL Rev.* **1996**, *46*, 1337.
23. Sobowale, K.; Ham, S.A.; Curlin, F.A.; Yoon, J.D. Personality traits are associated with academic achievement in medical school: A nationally representative study. *Acad. Psychiatry* **2018**, *42*, 338–345.
24. Khan, B.; Ahmed, A.; Abid, G. Using the 'Big-Five'for assessing personality traits of the champions: An insinuation for the sports industry. *Pak. J. Commer. Soc. Sci.* **2016**, *10*, 175–191.
25. Beall, L.; Bordin, E.S. The development and personality of engineers. *Pers. Guid. J.* **1964**, *43*, 23–32.
26. Feist, G.J. 14 The Influence of Personality on Artistic and Scientific Creativity. In *Handbook of Creativity*; Cambridge University Press: Cambridge, UK, 1999; p. 273.
27. Taher, A.M.M.; Chen, J.; Yao, W. Key predictors of creative MBA students' performance: Personality type and learning approaches. *J. Technol. Manag. China* **2011**, *6*, 43–68.
28. Bealing, W.E., Jr.; Baker, R.L.; Russo, C.J. Personality: What it takes to be an accountant. *Account. Educ. J.* **2006**, *16*, 37.
29. Spiller, T.R.; Rabe, F.; Ben-Zion, Z.; Korem, N.; Burrer, A.; Homan, P.; Harpaz-Rotem, I.; Duek, O. Efficient and Accurate Transcription in Mental Health Research-A Tutorial on Using Whisper AI for Audio File Transcription. *OSF Preprints* **2023**.
30. Silva Barbon, R.; Akabane, A.T. Towards transfer learning techniques—bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: A case study. *Sensors* **2022**, *22*, 8184.
31. Menze, B.H.; Kelm, B.M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F.A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform.* **2009**, *10*, 1–16.
32. Huang, Z.; Xu, P.; Liang, D.; Mishra, A.; Xiang, B. TRANS-BLSTM: Transformer with Bidirectional LSTM for Language Understanding. *arXiv* **2020**, arXiv:2003.07000.
33. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. The performance of LSTM and BiLSTM in forecasting time series. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3285–3292.
34. Devika, R.; Vairavasundaram, S.; Mahenthar, C.S.J.; Varadarajan, V.; Kotecha, K. A deep learning model based on BERT and sentence transformer for semantic keyphrase extraction on big social data. *IEEE Access* **2021**, *9*, 165252–165261.
35. Bridges, C.C., Jr. Hierarchical cluster analysis. *Psychol. Rep.* **1966**, *18*, 851–854.
36. Hasan, M.A.M.; Nasser, M.; Ahmad, S.; Molla, K.I. Feature selection for intrusion detection using random forest. *J. Inf. Secur.* **2016**, *7*, 129–140.
37. Lee, Y.G.; Oh, J.Y.; Kim, D.; Kim, G. Shap value-based feature importance analysis for short-term load forecasting. *J. Electr. Eng. Technol.* **2023**, *18*, 579–588.