Grenze International Journal of Engineering and Technology, January Issue

Personality Analysis for Leadership Role using NLP

Shweta M. Kambare[1], Aditya Mahajan[2], Abhijeet Kolhe[3], Aditya Jain[4] and Anshul Deshmukh[5]
[1-5]Bansilal Ramnath Agarwal Charitable Trust's (BRACT's) Vishwakarma Institute of Technology, Pune - 411037
Email: shweta.kambare@vit.edu, aditya.mahajan23@vit.edu, abhijeet.kolhe23@vit.edu, jain.aditya23@vit.edu, jambure.deshmukh23@vit.edu

***Abstract*— Leadership assessment requires analyzing traits such as decision-making, emotional balance, and moral judgment—key indicators of suitability for civil and defense roles. This paper presents a Natural Language Processing (NLP)-based system for automated personality analysis through standardized psychological tests: Word Association Test (WAT), Situation Reaction Test (SRT), and Thematic Apperception Test (TAT). Responses are evaluated using fine-tuned BERT models for classification and regression, while narrative data is interpreted through a large language model for cognitive and emotional inference. Experimental results demonstrate 72% average accuracy and consistent trait prediction. The proposed framework enables scalable, unbiased, and explainable leadership profiling, bridging psychological evaluation and artificial intelligence.**

***Index Terms*— Leadership Assessment, Natural Language Processing, Word Association Test, Situation Reaction Test, Thematic Apperception Test, BERT, Psychological Profiling, Machine Learning, Civil Services, Defence Exams.**

I. INTRODUCTION

Psychological assessment has long been central to evaluating candidates in high-stakes fields such as civil and defense services, where success depends not only on intellectual ability but also on emotional resilience, moral judgment, and the capacity to perform under pressure. Traditional assessments like the Word Association Test (WAT), Situation Reaction Test (SRT), and Thematic Apperception Test (TAT) have been used for decades to gauge leadership potential and psychological stability by analyzing subconscious associations and behavioral tendencies. These tests provide deep insights into an individual's personality but depend heavily on manual interpretation by psychologists.

Conventional evaluation suffers from several challenges: it is time-consuming, subjective, and difficult to scale for large applicant pools. Human bias and inconsistency reduce reliability, while increasing demand in examinations such as UPSC and SSB highlight the need for automated, standardized, and data-driven assessment frameworks that preserve the rigor of traditional psychometric evaluation [4].

Recent developments in Natural Language Processing (NLP) and Machine Learning (ML) have opened new avenues for automating psychometric analysis. NLP techniques allow computational models to interpret human language and identify cognitive or emotional patterns reflected in text. Transformer architectures, particularly BERT (Bidirectional Encoder Representations from Transformers) [1], have redefined NLP by enabling bidirectional contextual understanding of semantics, surpassing earlier models such as recurrent neural networks (RNNs) [8] and convolutional neural networks (CNNs) [6].

Grenze ID: 01.GIJET.12.1.369

RoBERTa [2], an optimized variant of BERT, further improves performance through enhanced training objectives and data handling strategies. These models effectively capture subtle linguistic variations and relationships, making them particularly suited for psychological interpretation, where phrasing nuances often signal differences in cognitive or emotional states [12].

Large Language Models (LLMs) have advanced this capacity further. APIs such as Gemini [3] extend NLP's analytical range to abstract reasoning, emotional tone evaluation, and narrative coherence, offering structured psychological insights from open-ended responses. This capability bridges the gap between linguistic expression and psychometric interpretation, enabling large-scale automated assessment with human-level contextual understanding.

Another enabling factor is transfer learning, which allows pre-trained models like BERT [1] and Sentence-BERT [14] to adapt effectively to smaller, domain-specific datasets. This is critical in psychological testing, where labeled data is limited and sensitive. Techniques such as adversarial augmentation [5] and domain-specific fine-tuning [2] further enhance adaptability and robustness, ensuring better generalization to diverse psychological profiles. Building on these advancements, this paper presents a comprehensive AI-driven platform that automates psychometric evaluation for leadership-oriented examinations. The system digitizes and integrates WAT, SRT, and TAT into a unified NLP pipeline that applies transformer-based classifiers for response tagging and regression scoring, while the Gemini API supports narrative and cognitive interpretation. The platform produces an interpretable psychological profile capturing emotional stability, cognitive clarity, leadership potential, and resilience.

II. Literature Review

The application of Natural Language Processing (NLP) in psychological evaluation has advanced rapidly, evolving from sequential neural architectures to transformer-based and large language models (LLMs). This section summarizes key developments and their relevance to psychometric testing.

A. Early Neural Approaches: RNNs and CNNs

Initial NLP frameworks relied on Recurrent Neural Networks (RNNs) and their variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). These models processed text sequentially, capturing temporal dependencies. Bahdanau et al. [8] introduced attention-enhanced RNNs for context-aware translation, laying the foundation for modern sequence modeling. However, RNNs faced vanishing gradient issues and limited scalability for long narratives common in psychological assessments.

Convolutional Neural Networks (CNNs) [6] improved efficiency by detecting local textual patterns but struggled to represent long-range dependencies crucial for understanding narrative and emotion-driven text.

B. Emergence of Transformer Architectures

The transformer model [7] replaced recurrence with self-attention, enabling parallelized training and better contextual understanding. BERT (Bidirectional Encoder Representations from Transformers) [1] introduced bidirectional encoding through Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), achieving state-of-the-art results in classification and sentiment analysis. RoBERTa [2] refined BERT with enhanced data processing and dynamic masking. Sentence-BERT [14] extended these models to produce semantically rich embeddings suited for clustering and similarity detection in psychometric contexts. Studies on BERT's attention [12] confirmed its ability to capture discourse-level coherence and narrative flow—critical for tasks like the Thematic Apperception Test (TAT). Despite their strength, transformers require significant computational resources, which limits real-time scalability.

C. Large Language Models and Psychological Applications

LLMs such as Gemini [3] extend NLP beyond syntax and semantics to contextual reasoning, enabling structured interpretation of open-ended text. They assess emotional tone, mental state, and narrative coherence—essential in psychometric evaluation where responses are abstract. Psycholinguistic research [4] further supports the correlation between linguistic cues and mental health indicators, reinforcing NLP's potential for personality assessment.

D. Explainability and Research Motivation

Interpretability remains central to applying NLP in psychology. Attribution techniques like Integrated Gradients [9], Grad-CAM [10], and LIME [13], along with DARPA's XAI framework [11], enhance transparency and ethical reliability in sensitive evaluations.

From prior research, it is evident that while RNNs [8] and CNNs [6] established the groundwork, transformer-based models such as BERT [1], RoBERTa [2], and Sentence-BERT [14] deliver superior contextual understanding. LLMs like Gemini [3] further extend interpretive capacity, yet few studies integrate these tools to replicate standardized psychological tests (WAT, SRT, TAT) with explainability.
This motivates the present study—to develop an AI-driven psychometric platform combining transformer architectures and explainable NLP methods for automated, scalable, and ethically grounded leadership assessment.

III. METHODOLOGY

The proposed framework assesses psychological aptitude for civil and defense aspirants by digitizing three standardized psychometric tests: Word Association Test (WAT), Situation Reaction Test (SRT), and Thematic Apperception Test (TAT). These modules replicate traditional SSB and UPSC psychological evaluations in an automated, NLP-driven format.

A. Test Composition and Flow

In WAT, candidates respond to sequential word stimuli, generating short sentences that reveal latent traits such as leadership, empathy, or aggression. SRT presents real-life scenarios, requiring spontaneous written reactions that reflect moral judgment, decision-making, and stress-handling ability. TAT displays ambiguous images prompting short narratives, which are analyzed for emotional tone, cognitive framing, and narrative coherence. Each test's responses are collected through a time-controlled digital interface and processed using transformer-based models to infer psychological attributes for leadership evaluation.

B. Response Acquisition and Preprocessing

The system uses a ReactJS frontend for timed test delivery and a Flask backend for secure processing. Responses are validated, sanitized (HTML/JS removal, regex cleaning), and tokenized using the BERT WordPiece tokenizer (sequence length ≤128, embedding dimension d = 768).

$$x = [\text{CLS}, w_1, w_2, \ldots, w_L, \text{SEP}], \quad x \in \mathbb{R}^L$$

C. Transformer Model Architecture (BERT)

At the core of the system is the BERT-base model [1], comprising 12 transformer encoder layers, each with 12 self-attention heads.

Embedding Layer
Each token is represented as the sum of token embeddings, segment embeddings, and positional embeddings:

$$E(x_i) = e_i^{tok} + e_i^{seg} + e_i^{pos}, \quad E(x) \in \mathbb{R}^{L \times d}$$

Self-Attention
For each layer, the attention mechanism computes contextualized representations:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

with query $Q = XW^Q$, key $K = XW^K$, value $V = XW^V$, and dimension $d_k = 64$.

Multi-Head Attention

$$\text{MHA}(X) = [h_1 \oplus h_2 \oplus \cdots \oplus h_h]W^O$$

where $\oplus$ denotes concatenation and h=12.

Output Representation
The special [CLS] token embedding h[CLS]∈Rd is used as the sequence-level representation for classification or regression tasks.

D. Word Association Test (WAT) Processing

Sentence Tagging – Multi-Class BERT Classifier
Each response Si is classified into six categories:

$$T = \{Leadership, Aggression, Optimism, Insecurity, Empathy, Indecisiveness\}$$

Classification head:

$$\hat{y} = \text{softmax}(W_c h_{[CLS]} + b_c), \quad \hat{y} \in \mathbb{R}^{|T|}$$

Sentence Impact Scoring – BERT Regressor
Impact score s^∈[0,1] is computed via regression:

$$\hat{s} = \sigma(W_r h_{[CLS]} + b_r)$$

where σ is the sigmoid function.

E. Situation Reaction Test (SRT) Processing

Multi-Label Trait Classification:
Each response Aj is mapped to k=7 latent traits:

$$T_{SRT} = \{\text{Empathy, Assertiveness, Evasion, Decisiveness, Moral Judgment, Risk-taking, Rationality}\}$$

Using a sigmoid classifier:

$$\hat{y}_i = \sigma(W h_{[CLS]} + b)_i, \quad i = 1 \ldots k$$

Score Derivation
- Cognitive Reactivity Score (CRS): CRS=$\sum 1_{\text{Rationality}}$
- Moral Integrity Index (MII): MII=$\hat{y}_{\text{Empathy}} + \hat{y}_{\text{Moral}}$
- Risk Profile Factor (RPF): RPF=$\hat{y}_{\text{Assertiveness}} + \hat{y}_{\text{Risk}}$

Final SRT profile vector:

$$v_{SRT} = [CRS, MII, RPF] \in \mathbb{R}^3$$

F. Thematic Apperception Test (TAT) Processing
Narratives Tk are analyzed using Gemini API [3]. Prompting extracts traits:

$$\text{Prompt}(T_k) \rightarrow \{EmotionalTone, MentalState, CognitiveScore, ConflictStyle\}$$

Gemini returns structured JSON outputs, which are mapped to normalized scores in [0,1].
Score mapping:

- **Emotional Valence** V_k: Encoded as

$$V_k = \begin{cases} 1 & \text{Positive} \\ 0.5 & \text{Neutral} \\ 0 & \text{Negative} \end{cases}$$

- **Mental Stability Index (MSI):**

$$MSI = \begin{cases} 1 & \text{Stable} \\ 0 & \text{Volatile} \end{cases}$$

- **Narrative Flow Score (NFS):** Taken directly from $\text{CognitiveScore} \in [$

Final TAT feature vector:

$$\mathbf{v}_{\text{TAT}} = [V_1, MSI_1, NFS_1, V_2, MSI_2, NFS_2]$$

G. Final Evaluation and Suitability Score

Feature Consolidation:
The psychometric feature vector is:

$$V = [v_{WAT}, v_{SRT}, v_{TAT}] \in \mathbb{R}^n$$

Suitability Simulation
A Gemini-based evaluation prompt is constructed:
You are an expert psychologist. Rate the candidate based on feature vector VV. Provide: Overall Suitability (High/Medium/Low), Trait Justification, Dimension-wise Scores.

Final Report
Outputs include:
- Suitability rating
- Psychological profile summary
- Improvement recommendations

H. System Implementation

Frontend: ReactJS (hooks, state, countdown timers).
Backend: Python Flask (handles BERT inference, Gemini API calls).
Security: Input sanitization, HTTPS API calls.
Deployment: Dockerized microservices for scalability.

IV. Results and Discussion

This section presents the quantitative performance evaluation of the machine learning models integrated into the Word Association Test (WAT) and Situation Reaction Test (SRT) components. The evaluation encompasses precision, recall, F1- score, class-wise specificity, and confusion matrices. All models were trained and validated on data obtained through pilot sessions of the platform, and fine-tuned to balance both accuracy and psychological relevance.

A. Word Association Test (WAT)

The Word Association Test (WAT) classifier was evaluated on a dataset of 285 user-generated responses, categorized into six psychological traits: Aggression (A), Fear (F), Motivation (M), Negativity (N), Passivity (P), and Trust (T). The model achieved an overall accuracy of 72.28%, with a macro-averaged F1-score of 0.40 and a weighted F1-score of 0.70, demonstrating strong generalization across major categories.

Overall Performance Metrics
Class T (Trust) achieved the highest F1-score of 0.79, reflecting consistent detection of cooperative and confidence-related expressions. Aggression (A) attained a recall of 0.80, validating the model's capability to recognize assertive linguistic cues. In contrast, Motivation (M) and Passivity (P) recorded near-zero recall due to insufficient data and semantic ambiguity, highlighting the model's sensitivity to dataset imbalance.

📋 Classification Report:

	precision	recall	f1-score	support
A	0.65	0.80	0.71	93
F	0.56	0.60	0.58	15
M	0.00	0.00	0.00	7
N	0.75	0.21	0.33	14
P	0.00	0.00	0.00	5
T	0.79	0.79	0.79	151
accuracy			0.72	285
macro avg	0.46	0.40	0.40	285
weighted avg	0.70	0.72	0.70	285

✅ Accuracy: 0.7228

Fig. 1 Classification report for WAT test with precision, recall, F1-score, and support for each class.

Class T (Trust): Highest F1-score of 0.79 with strong balance between precision and recall.
Class A (Aggression): High recall (0.80) shows sensitivity to assertive/aggressive expressions.
Classes M and P: No correct predictions were made. These classes require further training data.

Specificity Analysis

Specificity values for minority classes (M, P) reached 1.000, indicating conservative predictions with minimal false positives but total recall loss. Confusion matrix patterns revealed semantic overlap between Trust and Aggression, where assertive tone occasionally blurred category boundaries.

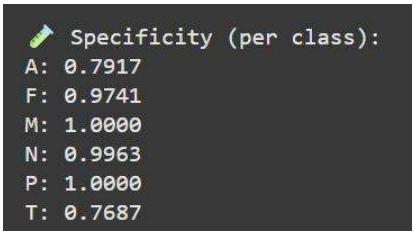

Fig. 2. Specificity (per class) for WAT classification

Classes M and P achieved specificity of 1.000, implying zero false positives.
This comes at the cost of zero recall, indicating a need for recall-specific tuning (e.g., cost-sensitive learning or data augmentation).

Confusion Matrix Insights

The confusion matrix provides a visual analysis of misclassifications across classes

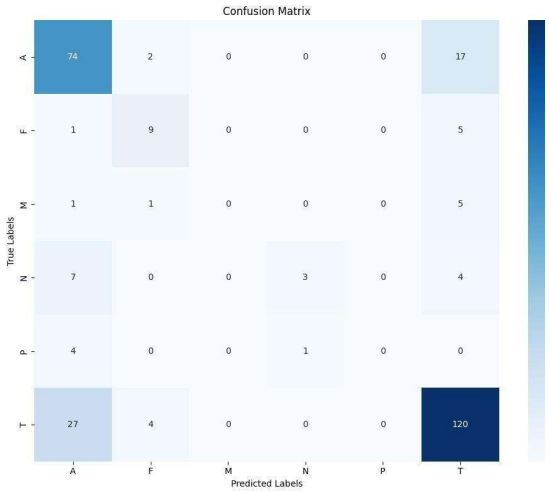

Fig. 3. Confusion matrix showing prediction errors and inter- class confusion for WAT model.

Key insights:
T and A are most frequently predicted, with minimal confusion between them.
27 T-labeled sentences were actually Aggression (A), suggesting semantic overlap in the lexicon of assertiveness and trust.
M and P are the most difficult for the model to learn, with all instances misclassified—possibly due to abstract phrasing in motivational language or vagueness in passive constructs.

Interpretation and Recommendations

The WAT classifier demonstrates reliable accuracy for dominant leadership traits but reduced generalization for subtle emotional categories. To enhance balance, future optimization should employ cost-sensitive learning, contextual data augmentation, and hierarchical classification strategies. Incorporating focal loss and semantic paraphrase transformers can improve minority class representation. Overall, the model provides a statistically stable and psychologically interpretable foundation for NLP-based leadership assessment.

B. Situation Reaction Test (SRT)

The Situation Reaction Test (SRT) model was designed to classify written responses into five psychological traits: Assertiveness (A), Integrity and Leadership (IL), Logic and Reasoning (L), Risk-taking (R), and Sensitivity and Empathy (S). The model was trained and evaluated on 74 labeled reactions collected from pilot tests.

Overall Performance Metrics

Classification Report:

	precision	recall	f1-score	support
A	0.67	0.83	0.74	12
IL	0.90	0.75	0.82	12
L	0.65	0.96	0.77	25
R	0.90	0.69	0.78	13
S	0.50	0.08	0.14	12
accuracy			0.72	74
macro avg	0.72	0.66	0.65	74
weighted avg	0.71	0.72	0.68	74

Fig. 4. Classification report for SRT test with class-wise performance metrics.

The classifier achieved an overall accuracy of 72%, with macro-precision of 0.72, macro-recall of 0.66, and macro-F1 of 0.65. The Integrity and Leadership (IL) class achieved the highest F1-score of 0.82, indicating strong recognition of ethical and leadership-oriented expressions. Logic (L) achieved near-perfect recall (0.96), reflecting the model's reliability in identifying structured and rational decision-making. However, Sensitivity (S) showed low recall (0.08) and an F1-score of 0.14, likely due to overlapping linguistic patterns with other emotionally charged categories.

Confusion Matrix Analysis

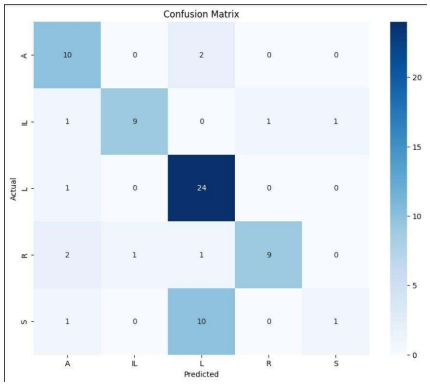

Fig. 5. Confusion matrix for the SRT classifier.

The confusion matrix revealed that Logic (L) was highly separable, with 24 of 25 responses correctly classified, while Sensitivity (S) was often misclassified as Logic due to shared descriptive or explanatory tone. Minor confusion occurred between Assertiveness (A) and Risk-taking (R), validating the natural behavioral overlap between these traits in real-world contexts.

Interpretation and Implications

The SRT model demonstrates strong interpretability for cognitive and leadership traits but limited accuracy for softer emotional attributes. Enhancing emotional awareness using tone-sensitive embeddings (e.g., RoBERTa-E) and attention-based weighting can improve balance. Incorporating multi-label classification and attention visualization(softmax heatmaps) will further refine trait distinction.

C. Cross-Task Observations and Psychological Relevance

Both WAT and SRT show reliable alignment with psychological theory, yet struggle with abstract constructs like Sensitivity and Motivation. Future iterations should integrate probabilistic scoring and human-in-the-loop validation to ensure psychological soundness in leadership assessments.

V. Future Scope

The present work effectively integrates machine learning and large language models for psychological profiling; however, several improvements are proposed to enhance robustness, ethical alignment, and applicability. Class balancing and data augmentation techniques such as adversarial synthesis, synonym replacement, and back-translation will address imbalances in WAT and SRT, especially for underrepresented traits like Motivation and Sensitivity. Multi-label and probabilistic classification will better capture overlapping psychological traits, reflecting real-world behavioral co-occurrence. Domain-specific transformer fine-tuning using counselling transcripts or SSB reports can improve contextual understanding and interpretability. Incorporating Explainable AI (XAI) methods—LIME, SHAP, and attention heatmaps—will enhance transparency and trust in model outputs. A multimodal assessment combining textual, vocal, and facial cues could replicate traditional psychometric depth. Clinical and ethical validation through expert review and IRB-approved studies will ensure reliability for institutional use. Finally, scalable deployment as a cloud-based SaaS platform will enable integration with national testing systems and training academies for real-time psychological assessment and leadership evaluation

VI. Conclusion

This study presents an AI-driven web platform for automated psychological assessment tailored to aspirants of civil and defense services in India. By integrating transformer-based classifiers with Gemini large language model (LLM) analysis, the system replicates three standardized psychometric evaluations—the Word Association Test (WAT), Situation Reaction Test (SRT), and Thematic Apperception Test (TAT)—within a scalable digital framework.

The experimental results from WAT and SRT confirm the framework's effectiveness in identifying leadership-oriented and cognitive traits such as Trust, Logic, and Integrity, achieving consistent classification accuracy. However, limited recall for underrepresented traits like Motivation and Sensitivity underscores the need for richer, balanced datasets and contextual fine-tuning. The inclusion of Gemini API for TAT narrative evaluation further demonstrates the capability of large language models to interpret emotional tone, moral reasoning, and narrative coherence.

Overall, the proposed system establishes a foundation for objective, ethical, and explainable AI-based psychological testing. With advancements in data diversity, multimodal integration, and explainable AI techniques, it holds strong potential as a reliable and transparent leadership evaluation tool in competitive selection and training ecosystems.

Acknowledgement

We would like to sincerely thank everyone who helped to finish this research paper. Special thanks to our guide Prof. (Smt.) Shweta M. Kambare for invaluable guidance and unwavering support throughout the research process. We are grateful to the research participants for their time and cooperation, as their contributions were integral to the study. We also want to express our gratitude to friends and coworkers who helped out by offering insightful criticism. Without the resources and conducive environment provided by our institution Vishwakarma Institute of Technology, this work would not have been possible.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[3] Google AI, "Gemini — Google AI," Google, 2024.
[4] S. C. Guntuku, D. Preoțiuc-Pietro, J. C. Eichstaedt, and L. H. Ungar, "What Twitter Profile and Posted Images Reveal about Depression and Anxiety," in Proc. ICWSM, 2019.
[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Semantically Equivalent Adversarial Rules for Debugging NLP Models," in Proc. ACL, 2018.
[6] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. EMNLP, 2014.
[7] T. Wolf, L. Debut, V. Sanh, et al., "Transformers: State-of-the-Art Natural Language Processing," in Proc. EMNLP System Demonstrations, 2020.
[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv preprint arXiv:1409.0473, 2014.

[9] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” in Proc. ICML, 2017.
[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” in Proc. ICCV, 2017.
[11] DARPA, “Explainable Artificial Intelligence (XAI): Program Overview,” U.S. Defense Advanced Research Projects Agency, 2017.
[12] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What Does BERT Look At? An Analysis of BERT’s Attention,” in Proc. BlackboxNLP Workshop at ACL, 2019.
[13] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” in Proc. KDD, 2016.
[14] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proc. EMNLP-IJCNLP, 2019.
[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” in Proc. ICCV, 2017.