**Research Report: Stance Detection Using Transformer Models**

**Date**: April 13, 2025  
**Prepared by**: [Your Name or Team Name, if applicable]  
**Institution**: [Your Institution, if applicable]  
**Report Type**: Faculty of Information Analysis (FIA)

**Abstract**

This study evaluates a stance detection pipeline using transformer models to classify social media texts on topics. The pipeline employs a T5 model for initial training and a fine-tuned BERTweet model for testing. The pre-trained T5 model showed poor performance (F1: 0.0), necessitating fine-tuning. Testing with BERTweet on three datasets yielded improved but moderate results (overall F1: 0.3884). Challenges included noisy data and dataset imbalance. Recommendations focus on enhancing data quality and model training.

**1. Introduction**

Stance detection is vital for understanding public sentiment on social media, particularly for divisive issues like gay rights, creationism, climate change, and public health policies. Transformer models, such as T5 and BERTweet, offer robust frameworks for such tasks due to their contextual understanding capabilities.

This report analyzes two Jupyter Notebooks:

* **Training**: In this phase, we implements a T5-based pipeline with pre-training evaluation and initial fine-tuning.
* **Testing**: Stance\_Testing.ipynb, which tests a fine-tuned BERTweet model on three datasets.

The objectives are to assess model performance, identify limitations, and propose improvements for stance detection accuracy.

**2. Methodology**

**2.1 Environment and Dependencies**

Both notebooks were executed in a Python 3.10 environment with GPU support (CUDA). Key libraries included:

**Table 1: Core Dependencies**

|  |  |  |
| --- | --- | --- |
| **Library** | **Version** | **Purpose** |
| transformers | 4.50.3 | T5 and BERTweet models, tokenizers. |
| torch | 2.6.0+cu126 | GPU-accelerated deep learning. |
| pandas | 2.2.3 | Data handling. |
| scikit-learn | 1.6.1 | Evaluation metrics (F1, precision). |
| nltk | 3.9.1 | NLP tools (METEOR score). |
| bert-score | 0.3.13 | Semantic similarity metric. |

**2.2 Datasets**

* **Training Dataset** (from Stance\_Updated\_2\_(1).ipynb):
  + ~1000 social media texts.
  + Labeled with four stances: gay rights, creationism, climate change is a concern, lockdown in New York state.
  + Distribution (based on 100 samples):

**Table 2: Training Dataset Stance Distribution**

| **Stance** | **Frequency** | **Percentage (%)** |
| --- | --- | --- |
| Gay rights | 34 | 34.0 |
| Creationism | 34 | 34.0 |
| Climate change is a concern | 28 | 28.0 |
| Lockdown in New York state | 4 | 4.0 |

* **Testing Datasets**:
  + Three CSV files: tse\_explicit.csv, tse\_implicit.csv, vast\_filtered\_im.csv.
  + Total: 4980 samples.
  + Labels: FAVOR, AGAINST, NONE.
  + Topics overlap with training (e.g., creationism, gay rights) but include broader issues (e.g., marijuana, immigration).

**2.3 Models and Training**

* **Training Phase** (Stance\_Updated\_2\_(1).ipynb):
  + **Model**: Pre-trained T5-base (892M parameters).
  + **Preprocessing**: Texts mapped to stance labels.
  + **Evaluation**: Pre-trained model tested with metrics like F1, ROUGE, BERTScore.
  + **Fine-tuning**: Initiated using transformers Trainer API with Weights & Biases logging (incomplete in output).
* **Testing Phase** (Stance\_Testing.ipynb):
  + **Model**: BERTweet-base, fine-tuned on test\_predictions\_vast\_filtered\_ex.csv.
  + **Fine-tuning**: 3 epochs, batch size 16, training loss decreased from 1.1412 to 0.5305.
  + **Testing**: Predictions generated using BERTweet and Ollama (LLaMA3) on three datasets.
  + **Input Format**: Post and keyphrase combined (e.g., post [SEP] keyphrase).
  + **Labels**: FAVOR, AGAINST, NONE.

**2.4 Evaluation Metrics**

Metrics used across both phases:

* **F1 Score**: Balances precision and recall.
* **Precision**: Proportion of correct predictions.
* Additional training metrics: ROUGE, METEOR, BERTScore, Word Mover's Distance (WMD).

**3. Results**

**3.1 Training Phase Results**

The pre-trained T5 model was evaluated on 100 samples:

**Table 3: Pre-training Evaluation Metrics**

| **Metric** | **Value** | **Interpretation** |
| --- | --- | --- |
| F1 | 0.0 | No correct predictions. |
| ROUGE-1 | 0.0045 | Minimal unigram overlap. |
| METEOR | 0.0094 | Poor semantic alignment. |
| BERTScore | 0.8134 | Moderate semantic similarity. |

**Sample Predictions**: **Table 4: Pre-training Sample Predictions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Text (Truncated)** | **Predicted Stance** | **Ground Truth Stance** | **Correct?** |
| "the god always remind us to live happily" | ",,,, the god always..." | gay rights | No |
| "evolution is the only of explaining" | "evolution is the only..." | creationism | No |
| "so hot and humid!" | "so hot and humid! ..." | climate change is a concern | No |

Fine-tuning was incomplete, with no post-fine-tuning metrics available.

**3.2 Testing Phase Results**

The fine-tuned BERTweet model was tested on three datasets, with results aggregated:

**Table 5: Testing Metrics (Fine-tuned BERTweet)**

| **Dataset** | **Precision (macro)** | **F1 Score (macro)** | **Accuracy** |
| --- | --- | --- | --- |
| tse\_explicit.csv | 0.3204 | 0.3455 | 0.48 |
| tse\_implicit.csv | 0.2994 | 0.3134 | 0.45 |
| vast\_filtered\_im.csv | 0.4622 | 0.4639 | 0.70 |
| **Overall (4980 samples)** | 0.3730 | 0.3884 | 0.56 |

**Table 6: Overall Classification Report**

| **Stance** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| FAVOR | 0.57 | 0.51 | 0.54 | 2167 |
| AGAINST | 0.55 | 0.72 | 0.63 | 2306 |
| NONE | 0.00 | 0.00 | 0.00 | 507 |

**Ollama Results** (for comparison):

* Overall F1: 0.4258, slightly better than BERTweet, but inconsistent across datasets.

**4. Discussion**

**4.1 Performance Analysis**

* **Training Phase**:
  + The pre-trained T5 model's zero F1 score indicates it was not adapted for stance detection, producing irrelevant outputs (e.g., repeating input text).
  + High BERTScore (0.8134) suggests semantic overlap due to shared vocabulary, not accurate stance prediction.
  + Incomplete fine-tuning limited improvements.
* **Testing Phase**:
  + The fine-tuned BERTweet model improved performance, with vast\_filtered\_im.csv achieving the highest F1 (0.4639) and accuracy (0.70).
  + Poor performance on NONE labels (F1: 0.0) indicates difficulty detecting neutral stances.
  + Ollama (LLaMA3) outperformed BERTweet slightly, suggesting potential in larger models.

**4.2 Challenges**

* **Data Quality**:
  + Training dataset: Noisy (hashtags, mentions) and imbalanced (lockdown: 4%).
  + Testing datasets: Implicit stances (tse\_implicit.csv) were harder to classify (F1: 0.3134).
* **Model Limitations**:
  + T5 struggled without fine-tuning.
  + BERTweet's uninitialized weights initially reduced effectiveness; fine-tuning helped but was limited by dataset size.
* **Technical Issues**:
  + Training: WMD score of 0.0, MoverScore approximation errors.
  + Testing: Wandb callback errors, RoBERTa weight initialization warnings.

**4.3 Recommendations**

1. **Enhance Dataset**:
   * Clean noise (e.g., remove hashtags).
   * Balance stances, especially underrepresented ones.
   * Increase training data size (>1000 samples).
2. **Improve Training**:
   * Complete T5 fine-tuning with optimized hyperparameters.
   * Extend BERTweet fine-tuning epochs or use larger models (e.g., RoBERTa).
3. **Address Technical Issues**:
   * Fix WMD and MoverScore implementations.
   * Set environment variables for Wandb and Hugging Face to avoid authentication issues.
4. **Explore Models**:
   * Test larger models like LLaMA3 (as used in Ollama) or RoBERTa for better contextual understanding.

**5. Conclusion**

This study implemented a stance detection pipeline using T5 for training and BERTweet for testing. The pre-trained T5 model failed (F1: 0.0), underscoring the need for fine-tuning. The fine-tuned BERTweet model achieved moderate success (overall F1: 0.3884), with best performance on vast\_filtered\_im.csv (F1: 0.4639). Data noise, imbalance, and technical issues limited results. Future work should prioritize dataset enhancement, extended fine-tuning, and robust model exploration to improve stance detection accuracy.

**6. References**

* Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140).
* Hugging Face. (2023). Transformers Documentation. <https://huggingface.co/docs/transformers/>
* Nguyen, D. Q., et al. (2020). BERTweet: A Pre-trained Language Model for English Tweets. *EMNLP 2020*.

**7. Appendices**

**Appendix A: Technical Notes**

* **Training Warnings**: T5 tokenizer legacy behavior, RoBERTa uninitialized weights, Wandb authentication issues.
* **Testing Warnings**: BERTweet uninitialized classifier weights, Ollama empty candidate sentence errors.
* **Environment**: User-specific library installations, suggesting permission constraints.

**Appendix B: Dataset Notes**

* **Training**: ~1000 samples, social media texts, noisy and imbalanced.
* **Testing**: 4980 samples across three datasets, varying explicitness (explicit vs. implicit).