NLP results discussion

NLP results discussion

A number on a white background

AI-generated content may be incorrect.

Macro à Calculates F1 score **for each class independently**, then takes the **unweighted average**

* Sensitive to poor performance on **underrepresented or difficult** classes

Micro à Aggregates the contributions of all classes to compute a **global** F1

* Skewed by **majority classes**, so weaker performance in rare classes (e.g., ORG) doesn’t heavily affect this score
* calculated by **summing up all the true positives, false positives, and false negatives across all classes**, and then computing **precision**, **recall**, and the **F1 score** from those totals

Weighted à Like Macro F1, but **weighs each class by its support (frequency),** takes a weighted average

* This high value means the model performs especially well on **frequent classes**, which dominates the dataset

Interpretation:

* PER and LOC:High F1 scores indicate strong alignment between silver and gold labels,
* These entities are easier to detect due to surface form clues (e.g., capitalization, context)
* More consistently labeled in WikiAnn, as names and places are generally well-defined in Wikipedia
* ORG: Lower F1 score indicates that silver labels often **mislabel** or **miss** organizations compared to human annotations
* Common causes: ambiguity (e.g., “Amazon” – company vs. river), may omit less notable organisations

When low performance on the ORG can be acceptable:

1.General Domain Models Used for Downstream Fine-Tuning

* Pretraining models, with the assumption that fine-tuning will be done later
* initial misclassifications in less frequent classes (like ORG) are acceptable, as the final task-specific model will adapt using cleaner data

2. . Use Cases Focused on People and Places

* Social media analytics for public sentiment or influencer tracking, where individuals (PER) and event venues (LOC) are more relevant than companies

3. Exploratory or Prototype-Stage Projects

* Early-stage research and experiments often prioritize speed and coverage over precision

Where it can cause problems:

1. News and media mining à In journalism, politics, or businesses analytics organizations are central players
2. Financial and Legal NLP

**3. What Can We Say About the Results?**

**Overall Impression**

* Silver data **is not perfect**, but **reasonably good** — the **Macro F1 of 73%** suggests a **moderate level of quality** across all classes.
* **Micro and Weighted F1 are higher (~80-82%)**, which suggests that **for frequent labels**, silver data is quite accurate.

**Important:**   
Macro F1 is lower because **some classes are much worse than others** (see low per-class F1 for some categories).

**Classes and Their F1 Scores**

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Label** | **F1 Score** | **Interpretation** |
| 0 | O (Other / Outside of any entity) | 0.8792 | Very strong — model reliably identifies non-entity tokens. |
| 1 | PER1 (Person - type 1) | 0.8981 | Excellent — person entities are recognized very well. |
| 2 | PER2 (Person - type 2) | 0.8352 | Very good — slight drop but still strong recognition. |
| 3 | ORG3 (Organization - type 3) | 0.5639 | Weak — organization entities are not detected reliably. |
| 4 | ORG4 (Organization - type 4) | 0.5231 | Weak — severe struggles on organization labeling. |
| 5 | LOC5 (Location - type 5) | 0.7524 | Good — locations are generally recognized well. |
| 6 | LOC6 (Location - type 6) | 0.6608 | Moderate — some room for improvement on locations. |

**Analysis in One Paragraph**

Our analysis shows that silver-standard labels achieve high F1 scores for non-entity tokens (O) and person entities (PER1 and PER2), with scores between 0.83 and 0.89. Location entities (LOC5 and LOC6) are moderately well recognized, achieving F1 scores between 0.66 and 0.75. However, silver data exhibits significant weaknesses in recognizing organization entities (ORG3 and ORG4), where F1 scores fall below 0.57. This suggests that while silver data is generally reliable for frequent and less ambiguous classes, it struggles considerably with more complex entity types like organizations, which may require further correction or human supervision. 

In our evaluation, we observed an overall accuracy of 75.48% when comparing silver-standard labels to our manually created gold-standard annotations. This suggests that while silver data captures the majority of labels correctly, approximately one in four tokens is mislabeled or disagreed upon. This result aligns with the Macro F1 score (73%), indicating that silver data offers reasonable coverage but introduces systematic errors that could impact model performance, especially for more complex entity types.   
   
**Compresed Notes from above into one text:**

In our project, we evaluated the reliability of silver-standard labels (WikiAnn) against manually annotated gold-standard data for Named Entity Recognition (NER) tasks. We focused on three evaluation metrics: Macro F1, Micro F1, and Weighted F1, each highlighting different aspects of model performance.

**Macro F1**, which calculates the F1 score independently for each class and then averages them equally, was **0.730**. This score indicates a **moderate quality** across all classes and highlights that the model struggles more with underrepresented or difficult classes.

**Micro F1**, aggregating contributions across all classes, reached **0.801**, suggesting that **frequent classes dominate the overall performance**.

**Weighted F1** was slightly higher at **0.822**, reflecting that the model performs especially well on frequent classes, which make up the majority of the dataset.

Breaking down the performance by entity type, we observed **high F1 scores for person entities** (PER1 and PER2) and **non-entity tokens** (O), with scores ranging between **0.83 and 0.89**. These strong results are likely due to surface clues such as capitalization and consistent labeling of names and locations in Wikipedia data.

**Location entities** (LOC5 and LOC6) were **moderately well recognized**, with F1 scores of **0.7524** and **0.6608**, respectively.

However, silver data **struggled significantly with organization entities** (ORG3 and ORG4), achieving **low F1 scores below 0.57**.

This weakness in recognizing organizations can be explained by several factors:

* **Ambiguity**: Words like "Amazon" can refer to a company or a river, depending on context.
* **Incomplete labeling**: Silver annotations may omit less notable organizations.
* **Contextual errors**: Organization identification often depends heavily on broader sentence context, which silver data models may fail to capture.

In some use cases, **lower performance on organization entities may be acceptable**. For instance:

* In **general-domain models** used for downstream fine-tuning, initial errors in rare classes like ORG can be corrected later during task-specific training.
* For **applications focused on people and places**, such as **social media analysis** or **public sentiment tracking**, detecting organizations may not be critical.
* During **exploratory or prototype-stage projects**, speed and coverage often take precedence over entity-level precision.

However, **low ORG performance would pose significant problems** in domains like **news and media mining**, **business analytics**, or **financial and legal NLP**, where organizations are central and errors could lead to serious misinformation.

In addition to F1 scores, we calculated **overall accuracy** by comparing silver and gold labels token-by-token. The model achieved an **accuracy of 75.48%**, meaning that approximately **one in four tokens was mislabeled or disagreed upon**. This finding is consistent with the Macro F1 score and reinforces that while silver data captures the majority of labels correctly, it introduces systematic errors — particularly in more complex entity types — that could affect downstream model performance.

**In conclusion**, our results show that silver-standard labels offer **reasonable coverage for frequent and simpler classes** (such as persons and locations), but **significant weaknesses remain for complex or ambiguous entities** like organizations. Depending on the target application, silver data may either serve as a cost-effective and scalable starting point or require additional correction efforts to ensure robust performance, especially in domains where precise entity recognition is critical.