

IntentifierBench: An Open-Source Web Tool for Benchmarking Authorship Identification using Illocutionary Force Large Language Models Based Classifiers

Godfrey Inyama

Independent Researcher

United Kingdom

godfreyinyama@outlook.com

Abstract

This work introduces *IntentifierBench*, an open-source web-based benchmarking tool for authorship identification grounded in illocutionary force modelling. The framework leverages large language model (LLM)-based classifiers to infer sentence-level communicative intent and inter-sentential argumentative relations, enabling the construction of discourse-aware author representations. IntentifierBench integrates two complementary components: (i) an illocutionary force classifier that captures sentence-level speech acts, and (ii) a sentence-pair classifier that models local discourse dynamics. The system is evaluated on the PAN2019 cross-domain authorship attribution dataset using a standardised preprocessing pipeline that constructs unified author profiles and controlled evaluation subsets. Experimental results demonstrate that illocutionary force distributions provide a meaningful and interpretable signal for authorship identification, outperforming pairwise discourse modelling in both accuracy and computational efficiency. Furthermore, the framework offers an explainable interface that visualises communicative patterns across authors, bridging the gap between black-box neural models and interpretable stylometric analysis. By shifting the focus from surface-level features to communicative intent, IntentifierBench establishes a novel direction for discourse-driven authorship identification and provides a reproducible benchmark for future research at the intersection of natural language processing, pragmatics, and explainable AI.

1 Introduction

Authorship identification has traditionally relied on lexical, syntactic, and stylistic features, often overlooking the underlying communicative intent that shapes how authors construct discourse, such as Illocutionary force. Illocutionary force classification can be attributed as a subtask of illocutionary

analysis (Wanderer, 2023). Illocutionary analysis aims to reveal the communicative functions behind spoken and written utterances. At the core of illocutionary analysis are illocutionary acts, which are actions performed through language, and illocutionary forces, which capture the speaker’s communicative intent. Holdcroft (Holdcroft, 1978) framed illocutionary force as the constituent of an utterance to perform a specific communicative action. Leveraging this fundamental theory, I aim to classify illocutionary forces to better understand communicative intentions in dialogue for Authorship Identification using a newly developed open-source tool. Computationally, thus far, dialogue acts have been used for intent recognition at the dialogue level. However, explicit treatment of illocutionary force classification as a standalone task for intent recognition is still underexplored, especially for authorship identification. IntentifierBench gives users the ability to identify the author of an unknown document using fine-tuned illocutionary force models. It also aims to support transparency and explainability by displaying each candidate author’s and unknown author’s profile across the classifier labels to indicate how it made its decision.

To show the initial use case of IntentifierBench, an authorship identification task was set up using the PAN 2019 dataset. Only one problem set and 5 unknown authored documents were experimented with. The contributions of this preliminary work are:

- **Discourse-Driven Authorship Representation:** I introduced a novel authorship identification framework that models writing style through *illocutionary forces* and *argumentative discourse relations*, moving beyond traditional lexical and syntactic features toward pragmatically grounded representations.
- **Embedding Pipeline for Stylometry:** I propose a unified pipeline that integrates

transformer-based sentence classification and sentence-pair modelling to construct document-level embeddings, enabling similarity-based authorship identification using discourse-informed signals.

- **IntentifierBench - A Benchmark for Discourse-Aware Authorship Identification:** I present IntentifierBench ¹, a proof-of-concept open-source benchmark built on the PAN2019 dataset with a novel preprocessing strategy that aggregates author profiles and standardises evaluation, facilitating reproducible research in discourse-aware stylometry.

2 Background

Illocutionary force analysis, which is the identification and classification of the communicative intent encoded within speech acts, have been foundational to pragmatic theory since Austin and Searle first formalised the distinction between locutionary act, illocutionary act, and illocutionary force (Austin, 1962)(Searle, 1969). Both Pragmatic and linguistic studies have continued to refine these theoretical distinctions, stressing that illocutionary force can often be encoded at both the lexical and syntactic levels and is not restricted to root clauses (Woods, 2016). Despite these advances and improvements, computational approaches often confound illocutionary force with related concepts such as dialogue act or explicit intent classification, which fail to capture the full range of communicative function present in natural discourse (Mann, 1980) (Reed, 2010).

Recent research acknowledges that the current Natural Language Processing (NLP) models, such as the transformer-based models like BERT and notable open-source LLMs are good at semantic and syntactic tasks but cannot operationalise and generalise pragmatic features such as illocutionary force in dialogue (Inyama, 2025) (Woods, 2016). For example, experiments with both the Argument Interchange Format (AIF) and custom dialogue structures reveal that classification performance for illocutionary force lags behind other established utterance-level classification tasks, owing in part to class imbalance and the contextual subtlety of illocutionary distinctions (Inyama, 2025). Parsing and modeling illocutionary structure at a moderate level

remains relatively scarce in NLP literature, with only isolated successes in argument mining and dialogue structure parsing (Reed, 2010) (Inyama, 2025).

Therefore, a significant gap persists whereby most dialogue-oriented NLP systems and tools do not systematically model or classify illocutionary force at the utterance level, limiting the depth and responsiveness of conversational agents and automated argument analysers. By directly focusing on illocutionary force classification, leveraging novel model architectures and datasets, current research will advance both theory and practice in computational pragmatics, aiming to improve both analytical precision and downstream conversational agent utility.

3 IntentifierBench

IntentifierBench is an open-source benchmark for authorship identification that leverages *illocutionary forces* and *argumentative discourse structures* as stylistic signals. Unlike the traditional approaches that rely on lexical or syntactic features, IntentifierBench models an author’s communicative intent through sentence-level and inter-sentential classification using transformer-based fine-tuned models on the Argument Interchange Format datasets (Lawrence et al., 2012). The framework operationalises authorship attribution as a similarity matching problem over discourse-informed embeddings derived from illocutionary force classification and argumentative relation prediction between coherent sentences. These representations provide a higher-level abstraction of writing style grounded in pragmatic and discourse theory. Figure 1 below shows the architecture of IntentifierBench.

3.1 Key Components

IntentifierBench is a Python-Flask web application that consists of the following core components:

- **Illocutionary Force Classifier (Model 1):** This is a transformer-based sequence classification model that assigns each sentence an illocutionary label, e.g. assertion, directive, commissive ². The output logits are aggregated to form a document-level embedding.

¹<https://github.com/Godfrey2712/IntentifierBench>

²https://huggingface.co/Godfrey2712/amf_illoc_force_intent_recognition

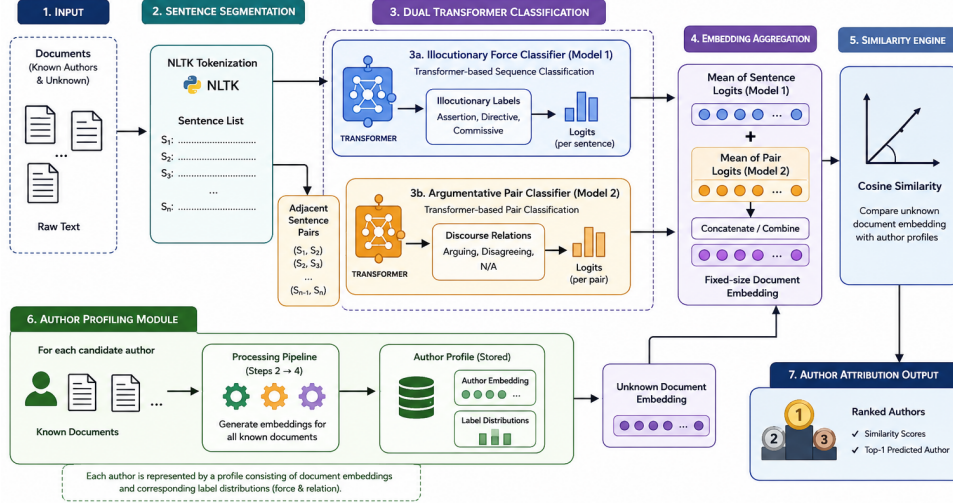


Figure 1: IdentifierBench Architecture

- **Argumentative Pair Classifier (Model 2):** A second transformer model that processes pairs of coherent sentences to capture discourse relations, e.g, Arguing, Disagreeing, N/A³. This captures structural writing patterns beyond isolated sentences.
- **Sentence Segmentation Module:** Texts are segmented into sentences using NLTK tokenisation, forming the basis for both classification pipelines.
- **Embedding Aggregation Mechanism:** Sentence-level and sentence-pair logits are averaged to produce fixed-size document embeddings representing discourse behaviour.
- **Similarity Engine:** Cosine similarity is used to compare the embeddings of unknown documents with known author profiles.
- **Author Profiling Module:** Each candidate author is represented by embeddings derived from their known documents, along with corresponding label distributions.

3.2 Operational Mechanism

IdentifierBench follows a multi-stage pipeline that transforms raw text into discourse-informed representations for similarity-based authorship identification.

³https://huggingface.co/Godfrey2712/arg_mining_us2016_locutions

Step 1: Sentence Tokenisation - Given a document D , it is segmented into a sequence of sentences:

$$D = \{s_1, s_2, \dots, s_n\} \quad (1)$$

Step 2: Illocutionary Force Encoding - Each sentence s_i is passed through the illocutionary classifier:

$$\mathbf{z}_i = f_{\text{illoc}}(s_i) \quad (2)$$

where $\mathbf{z}_i \in \mathbb{R}^k$ represents the logits vector.

The document-level embedding is later obtained via mean pooling:

$$\mathbf{z}_D^{(1)} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \quad (3)$$

Step 3: Argumentative Pair Encoding - Adjacent sentence pairs (s_i, s_{i+1}) are processed:

$$\mathbf{p}_i = f_{\text{pair}}(s_i, s_{i+1}) \quad (4)$$

The pairwise embedding is later aggregated as:

$$\mathbf{z}_D^{(2)} = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbf{p}_i \quad (5)$$

Step 4: Joint Representation - When both models are used, embeddings are concatenated:

$$\mathbf{z}_D = [\mathbf{z}_D^{(1)} \parallel \mathbf{z}_D^{(2)}] \quad (6)$$

Step 5: Similarity-Based Attribution - Given an unknown document embedding $\mathbf{z}_u$ and a set of candidate author embeddings $\{\mathbf{z}_a\}$, cosine similarity is computed:

$$\text{sim}(\mathbf{z}_u, \mathbf{z}_a) = \frac{\mathbf{z}_u \cdot \mathbf{z}_a}{\|\mathbf{z}_u\| \|\mathbf{z}_a\|} \quad (7)$$

The predicted author $\hat{a}$ is:

$$\hat{a} = \arg \max_a \text{sim}(\mathbf{z}_u, \mathbf{z}_a) \quad (8)$$

The Authorship Identification attribution pipeline is described in the algorithm below.

Algorithm 1 IntentifierBench Attribution Pipeline

Require: Unknown document D_u , set of known documents $\mathcal{D}_A$ for each author A

Ensure: Predicted author $\hat{A}$

- 1: Tokenize D_u into sentences $\{s_1, \dots, s_n\}$
- 2: Compute $\mathbf{z}_u^{(1)}$ using illocutionary classifier
- 3: Compute $\mathbf{z}_u^{(2)}$ using sentence-pair classifier
- 4: $\mathbf{z}_u \leftarrow$ concatenate embeddings (if both models used)
- 5: **for** each author A **do**
- 6: Aggregate known documents into $\mathbf{z}_A$
- 7: Compute similarity score:

$$\text{sim}_A = \cos(\mathbf{z}_u, \mathbf{z}_A)$$

- 8: **end for**
 - 9: **return** $\hat{A} = \arg \max_A \text{sim}_A$
-

3.2.1 Usage of IntentifierBench

To use IntentifierBench in practice, these steps are followed:

1. Prepare a dataset of known author documents and unknown documents and upload to the open-source tool from the backend. By default, the PAN 2019 dataset is used. Each file’s content can be updated, and the “DOCUMENTS_PATH” can be changed.
2. Select a model configuration: After running the application, you can select the model to use from the options below. The model options are seen below.
 - model1: Illocutionary force only
 - model2: Argumentative pairs only
 - both: Combined representation
3. Run the pipeline to extract embeddings for all documents. Once the unknown author documents are in the correct directory in the backend, they are reflected on the front-end, and you can select the unknown file you want to identify based on the candidate authors’ known files.

4. Computes cosine similarity between unknown and known embeddings.
5. Assign authorship based on maximum similarity and displays the likely author and an author profile graph for transparency and explanation.

This framework is particularly suited for research in discourse-aware stylometry, explainable authorship attribution, and cognitively grounded language modelling.

3.3 Dataset and Model

The dataset use for experimentation is the PAN-2019 Authorship Identification Dataset (Kestemont et al., 2019). The dataset consists of 20 problems across 5 languages. For the task in these foundational studies, only the English corpus was used and one problem set. Each problem set consists of various candidate known author files, some unknown author files, and a ground truth file. Each problem instance is organised as follows:

- A set of candidate author directories (candidateXXXXX), each containing multiple text files written by the same author.
- A directory of unknown texts (unknown) containing documents whose authorship must be determined.

The two models used are a fine-tuned illocutionary force classification transformer-based LLM:

- Model 1: The first model is a RoBERTa-large based classifier fine-tuned to identify the illocutionary force—i.e., the communicative intent—of individual utterances in dialogue. It predicts one of ten speech-act categories, including asserting, agreeing, disagreeing, challenging, and several forms of questioning. Trained on multiple argumentation-focused corpora from the AIFdb (MM2012⁴, MM123 (Schad et al., 2024), QT30 (Hautli-Janisz et al., 2022), US2016 (Visser et al., 2019)), the model achieves an overall accuracy of 0.67 and a macro-F1 of 0.55, with strong performance on classes such as Asserting and Pure Questioning, and weaker performance on underrepresented or ambiguous categories. Its purpose is to support fine-grained analysis of dialogical structure by automatically labeling the pragmatic function of utterances.

⁴<https://corpora.aifdb.org/mm2012>

- **Model 2:** The second model is a DeBERTa-v2 based text classification system designed to identify the illocutionary relationship between pairs of coherent locutions in dialogue. Given two linked statements, it classifies their interaction as Arguing, Disagreeing, or N/A, where the latter captures cases with no clear argumentative force. Trained on the US2016 corpus from AIFdb, the model achieves an overall accuracy of 0.69, with F1-scores of 0.69 for Arguing, 0.52 for Disagreeing, and 0.72 for N/A, reflecting stronger performance on well-represented classes. This model supports computational argumentation research by enabling automatic detection of argumentative relations within dialogical data.

The fine-tuning process of these models can be found in Inyama et al.'s work (Inyama, 2025).

3.4 Preprocessing of the PAN2019 Authorship Identification Corpus

The PAN2019 Cross-Domain Authorship Attribution dataset is employed as the primary benchmark for evaluating IdentifierBench. The dataset consists of multiple attribution problems, each containing a set of candidate authors with known documents and a collection of anonymous (unknown) texts to be attributed. The preprocessing pipeline is designed to:

1. Construct unified author profiles by aggregating all documents belonging to a candidate.
2. Standardise document representation across all problems.
3. Select a controlled subset of unknown documents for evaluation.

For each candidate author, all associated text files are concatenated into a single document. Formally, let a candidate author A_i have a set of documents:

$$\mathcal{D}_i = \{d_1, d_2, \dots, d_m\} \quad (9)$$

The aggregated author profile is defined as:

$$D_i = \bigoplus_{j=1}^m d_j \quad (10)$$

where $\oplus$ denotes concatenation with newline separation. This transformation ensures that each author is represented by a single consolidated document

capturing their stylistic and discourse patterns. To ensure consistency and computational tractability, a fixed subset of unknown documents is selected per problem. Specifically, the first 5 unknown files are retained:

$$\mathcal{U} = \{u_1, u_2, \dots, u_5\} \quad (11)$$

Each selected unknown document is copied into the processed dataset without modification. For each problem instance, a corresponding output directory is created, containing:

- One aggregated document per candidate author:

candidateXXXXX.txt

- A subset of unknown documents:

unknownXXXXX.txt

This results in a flattened and uniform structure that facilitates efficient downstream processing.

The Authorship Identification preprocessing pipeline for the PAN2019 dataset is described in the algorithm below.

Algorithm 2 PAN2019 Preprocessing Pipeline

Require: Root dataset directory $\mathcal{P}$, set of problems $\{p_1, \dots, p_n\}$

Ensure: Preprocessed dataset $\mathcal{P}'$

```

1: for each problem  $p \in \mathcal{P}$  do
2:   Create output directory  $p'$ 
3:   for each candidate author  $A$  in  $p$  do
4:     Initialise empty text buffer  $D_A$ 
5:     for each document  $d$  in  $A$  do
6:       Read  $d$  and append to  $D_A$ 
7:     end for
8:     Write  $D_A$  to candidateXXXXX.txt
9:   end for
10:  Select first 20 unknown documents  $\mathcal{U}$ 
11:  for each  $u \in \mathcal{U}$  do
12:    Copy  $u$  to output directory  $p'$ 
13:  end for
14: end for
15: return  $\mathcal{P}'$ 
```

4 Experiment

All experiments were carried out on a 8GB RAM Nvidia-GPU computer. In the PAN 2019 dataset, only problem00001 was used. All the known author text files were concatenated into one large candidate author individual files. Only Unknown00001 to Unknown00005 of problem00001 were set out for the authorship identification task. After each unknown author document file was identified, the candidate author prediction was recorded across the three models used. Also, the time taken to complete each identification for each unknown document across each model was recorded. To cross-check the accuracy of the classification and attribution, the ground truth (true candidate author of an unknown document) was used for verification. To identify the likely author of an unknown document loaded, the Identify Author button beside the unknown document is clicked while ensuring the model to use has been selected.

4.1 Results

After the experiments across all models for the 5 unknown author documents, I record the results in Table 1 below, where GT = Ground Truth, Illoc = Illocutionary Force Classifier, Pair = Argumentative Pair Classifier, and Both = Combination of Illoc and Pair.

To offer transparency and explanation on the authorship identification decision for the correctly attributed documents, the visualisations are shown below.

Model 1 (Illoc) - Unknown00001

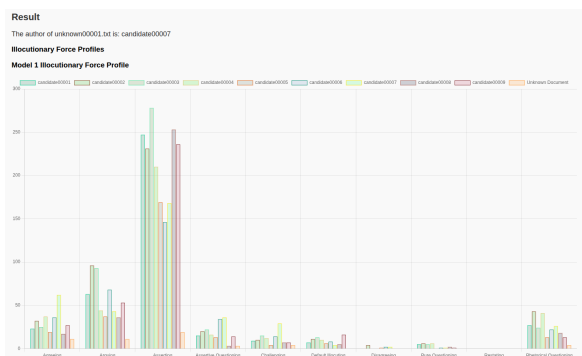


Figure 2: Model 1 (Illoc) - Unknown00001 Profile of Accurate Prediction on IntentifierBench

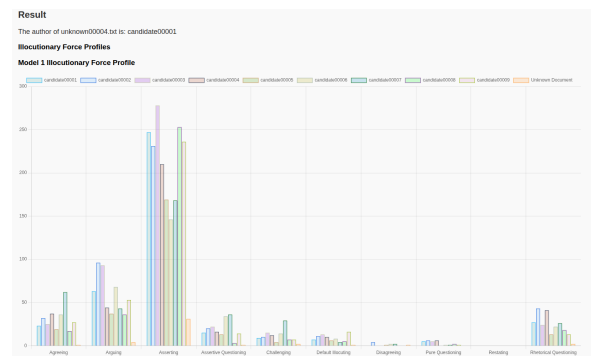
Model 1 (Illoc) - Unknown00007

Figure 3: Model 1 (Illoc) - Unknown00007 Profile of Accurate Prediction on IntenifierBench

Model 3 (Both) - Unknown00001 and **Model 3 (Both) - Unknown00007** are found in the Appendix below.

4.2 Analysis

Authors 00001 to 00009 from the PAN dataset are represented as 1 to 9 in figure 4 below.

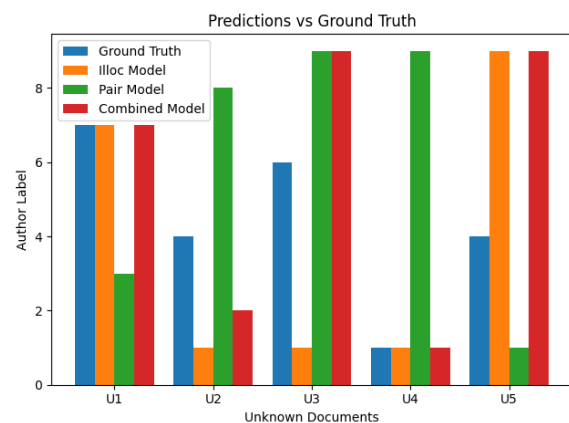


Figure 4: Comparison of model predictions against ground truth labels for each unknown document: The horizontal axis shows the unknown document IDs. The vertical axis shows the Candidate Authors.

In Figure 4 above, it is seen that the ground truth demarcated as blue only aligns with a model’s prediction at Unknown00001 (U1) and Unknown00004 (U4). The next section gives an overview and implications of this analysis.

5 Discussion

Table 1 shows the authorship identification performance across three model configurations: (i) illocutionary force classification, (ii) argumenta-

Unknown	GT	Illoc	Time (s)	Pair	Time (s)	Both	Time (s)
00001	00007	00007	630.96	00003	2455.35	00007	3035.21
00002	00004	00001	504.16	00008	2611.21	00002	3130.49
00003	00006	00001	537.72	00009	2610.82	00009	2905.13
00004	00001	00001	529.52	00009	2594.50	00001	3138.18
00005	00004	00009	505.02	00001	2366.38	00009	2903.58

Table 1: Authorship attribution results using illocutionary force, argumentative pair, and combined models.

tive sentence-pair modelling, and (iii) a combined representation.

The illocutionary force model achieves an accuracy of 40%, correctly identifying 2 out of 5 unknown documents. In contrast, the sentence-pair model fails to correctly classify any instance, yielding 0% accuracy. The combined model matches the performance of the illocutionary model with 40% accuracy, indicating that the inclusion of pairwise discourse features does not improve predictive performance in this setting for the data sample size used. This insight suggests that sentence-level communicative intent (illocutionary force) provides a more stable and discriminative signal for authorship identification rather than local inter-sentential relations alone. The poor performance of the pair model may be attributed to noise accumulation across sentence transitions or insufficient robustness of pairwise discourse classification, as seen in only utilising 3 labels of Arguing, Disagreeing, or N/A.

For the computational cost, a significant disparity is observed in computational efficiency. The illocutionary model requires approximately 541 seconds per document on average, whereas the pair model incurs a substantially higher cost of approximately 2528 seconds. The combined model exhibits an additive computational cost, reaching approximately 3023 seconds per document. This difference arises from the quadratic nature of sentence-pair processing, where each document of length n yields $n - 1$ inference steps, compared to n independent sentence evaluations in the illocutionary force model.

In contrast to expectations, combining both models does not yield performance gains. This suggests that the two representations may not be complementary in their current form, or that the concatenation strategy does not effectively integrate the information. There is also a possibility that the pairwise model introduces noise that degrades the overall embedding quality.

5.1 Implications

These results highlight the importance of discourse-level abstraction in authorship identification while also exposing the limitations of naive sentence-pair modelling. The findings suggest that:

- Illocutionary force distributions capture meaningful stylistic signatures.
- Argumentative pair modelling requires further refinement or alternative aggregation strategies.
- Computational efficiency is a critical consideration for scalable authorship identification for illocutionary force classifiers.

In general, the proposed framework demonstrates the feasibility of discourse-driven authorship attribution, while motivating future work on more robust and efficient integration of discourse structures.

6 Conclusion

IntentifierBench, a proof of concept, offers the ability to carry out authorship identification based on Illocutionary Forces LLM classifiers. The experiment carried out further demonstrates its impact on practical authorship identification tasks. While recorded successes, there are some limitations of IntentifierBench. Future studies should critically explore how the Illocutionary Forces LLM-classifiers perform against other gold linguistic features used in authorship identification, and also evaluate at an ablation level where these linguistic features are integrated with the illocutionary force features. In this foundational work, only a small sample of the dataset from the PAN 2019 dataset is used. Offering better insight, future work should include more data samples and reevaluate performance for a more generalised report for the PAN 2019 dataset, and consider non-cross-domain datasets. The tool can also be improved by including the ability to select from a variety of features other than the illocutionary force features, and a selection capability

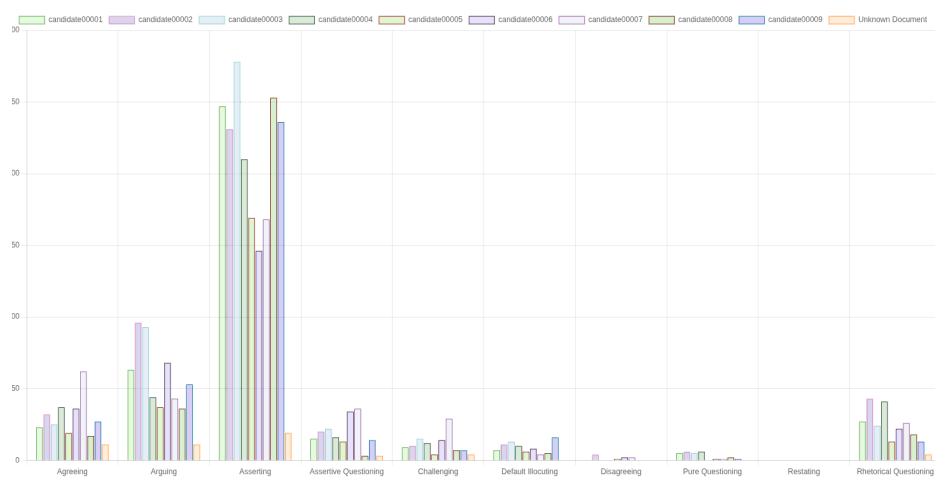
of more than one feature/model choice. Finally, while I leverage a gold dataset benchmark for the studies, a more practical approach would consider using real-time participants in dialogue and developing mechanisms to build their profiles and cross-examine them over unknown sample documents written by the candidate participants.

References

- J.L. Austin. 1962. *How to Do Things with Words*. Oxford University Press.
- Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. [QT30: A corpus of argument and conflict in broadcast debate](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.
- David Holdcroft. 1978. *Words and Deeds: Problems in the Theory of Speech Acts*. Oxford University Press.
- Godfrey Inyama. 2025. Exploiting illocutionary forces in dialogue structures for enhancing authorship identification. Master’s thesis, University of Dundee.
- Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. [Pan19 authorship analysis: Cross-domain authorship attribution](#).
- John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. 2012. [Aifdb: infrastructure for the argument web](#). In *Computational Models of Argument*, Frontiers in artificial intelligence and applications, pages 515–516, Netherlands. IOS Press. Fourth International Conference on Computational Models of Argument, COMMA 2012; Conference date: 10-09-2012 Through 12-09-2012.
- W.C. Mann. 1980. Toward a speech act theory for natural language processing. Technical report, USC/Information Sciences Institute.
- C. Reed. 2010. The role of illocutionary force in computational models of argument. In *COMMA*, pages 57–68.
- Ella Schad, Jacky Visser, and Chris Reed. 2024. [The RIP corpus of collaborative hypothesis-making](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16047–16057, Torino, Italia. ELRA and ICCL.
- J.R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cam: Verschiedene Aufl. Cambridge University Press.
- Jacobus Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2019. [Argumentation in the 2016 us presidential elections: Annotated corpora of television debates and social media reaction](#). *Language Resources and Evaluation*, 54:123–154.
- Jeremy Wanderer. 2023. [Clarifying illocutionary force](#). *Inquiry*, 0(0):1–28.
- Rebecca Louise Woods. 2016. *Embedding illocutionary force*. Ph.D. thesis, University of York.

APPENDIX

A Model 3 (Both) - Unknown00001



Model 2 Illocutionary Force Profile

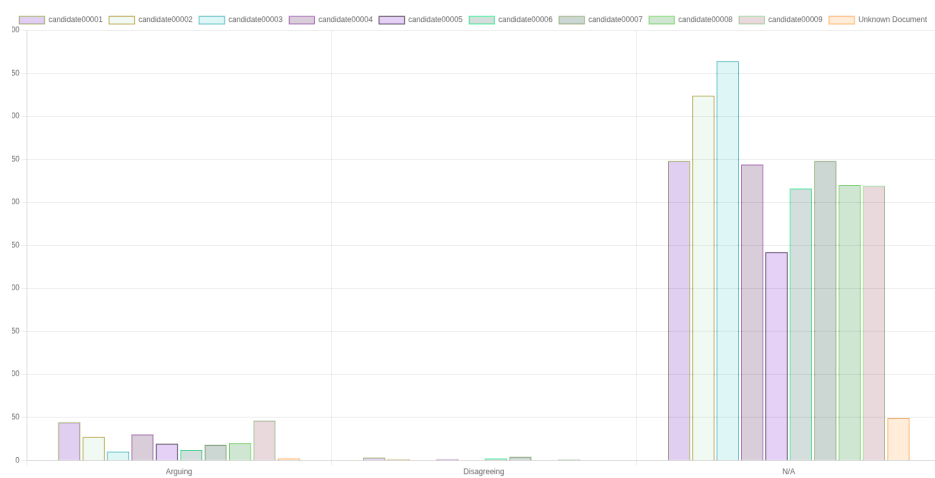


Figure 5: Model 3 (Both) - Unknown00001 Profile of Accurate Prediction on IntentifierBench

B Model 3 (Both) - Unknown00007

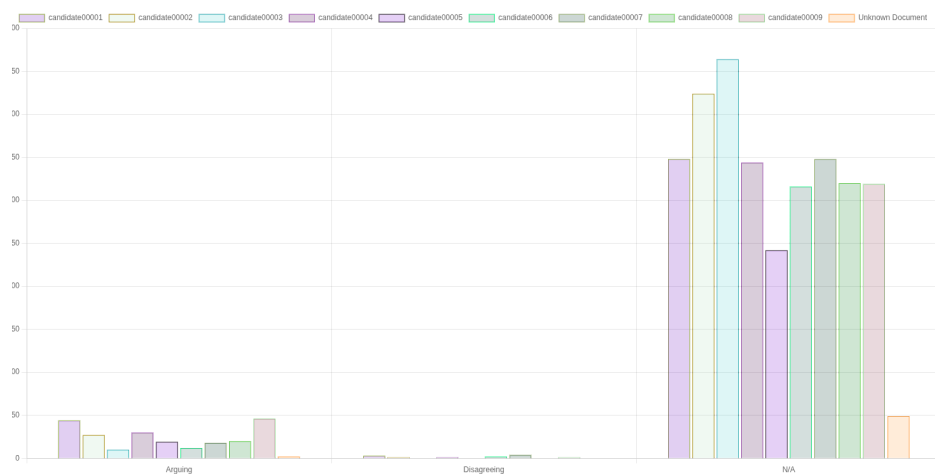
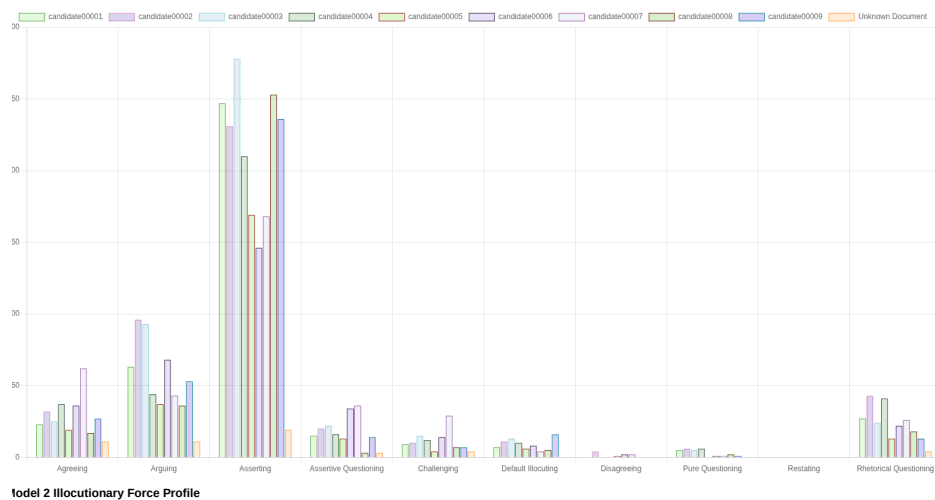


Figure 6: Model 3 (Both) - Unknown00007 Profile of Accurate Prediction on IdentifierBench