# Political Discussion at Scale

## *Argument Mining with Deep Learning*

**Isaac D. Mehlhaff**

Department of Political Science
The University of North Carolina at Chapel Hill
imehlhaff.net
mehlhaff@live.unc.edu

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

## Objectives

- **Understand** use of argumentation techniques and strategies in political discussion.
- **Push past** argument mining's primary focus on structural tasks (e.g. extracting premises and conclusions) to include more holistic characteristics of interest to social scientists.
- **Assess** the potential for deep learning to facilitate the study of political argumentation, particularly in online discourse.

## Data and Classifiers

- **Internet Argument Corpus** (Abbott, Ecker, et al. 2016), **Argument Extraction Corpus** (Swanson, Ecker, and Walker 2015)
  - Discussions/debates from online forums
  - Each document assigned scalar value on nine characteristics by 5-7 annotators
  - Dichotomized to create classification task
- 80%-10%-10% train-validate-test split
- **Feature extraction**:
  - Unigrams for lexical baseline
  - Bidirectional encoding representations from transformers (BERT) for deep learning models (Devlin et al. 2019)
- **Classifiers**:
  - **Baselines**: Naïve (random) and unigram (support vector machine)
  - Logistic regression **(LR)**
  - Support vector machine with stochastic gradient descent **(SVM w/ SGD)**
  - Random forest **(RF)**
  - Extremely randomized trees **(ERT)**
  - Random forest with extreme gradient boosting **(RF w/ XGB)**
  - Fine-tuned BERT neural network

| Task | N | Range | Mean | SD | Class Balance |
|---|---|---|---|---|---|
| **Disagree / Agree** | 28,171 | [-5, 5] | -0.916 | 1.689 | 0.802 / 0.198 |
| **Emotion / Fact** | 23,057 | [-5, 5] | 0.065 | 1.536 | 0.468 / 0.532 |
| **Attacking / Respectful** | 24,819 | [-5, 5] | 0.628 | 1.485 | 0.242 / 0.758 |
| **Nasty / Nice** | 26,612 | [-5, 5] | 0.895 | 1.439 | 0.159 / 0.841 |
| **Individual / Audience** | 5,997 | [-5, 5] | -1.271 | 2.09 | 0.816 / 0.184 |
| **Defeater / Undercutter** | 5,603 | [-5, 5] | -0.671 | 2.245 | 0.673 / 0.327 |
| **Counterargue / Attack** | 5,831 | [-5, 5] | -0.479 | 2.293 | 0.625 / 0.375 |
| **Question / Assert** | 6,216 | [-5, 5] | 0.717 | 2.368 | 0.319 / 0.681 |
| **Argument Quality** | 3,895 | [0, 1] | 0.54 | 0.277 | 0.595 / 0.405 |

**Table 1:** Descriptive statistics of document annotations.

## Deep Learning Pipeline

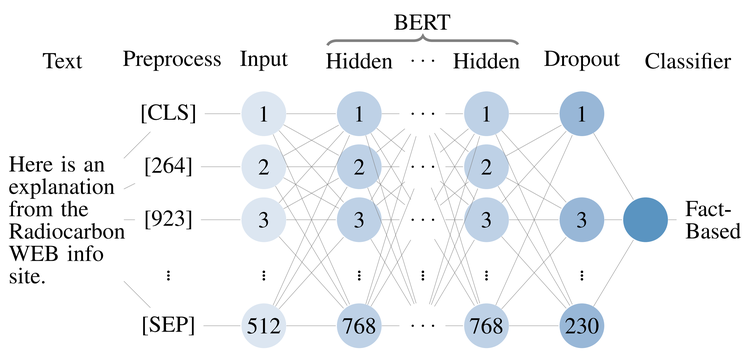


**Figure 1:** Visual Representation of Deep Learning Pipeline. Hidden layers are transformer blocks. Fine-tuned BERT models have two hidden layers with 128 neurons and two attention heads. Other classifiers using BERT feature extraction have twelve hidden layers with 768 neurons and twelve attention heads. Dropout and output layers are fully connected in fine-tuned BERT model.

## Performance Metrics

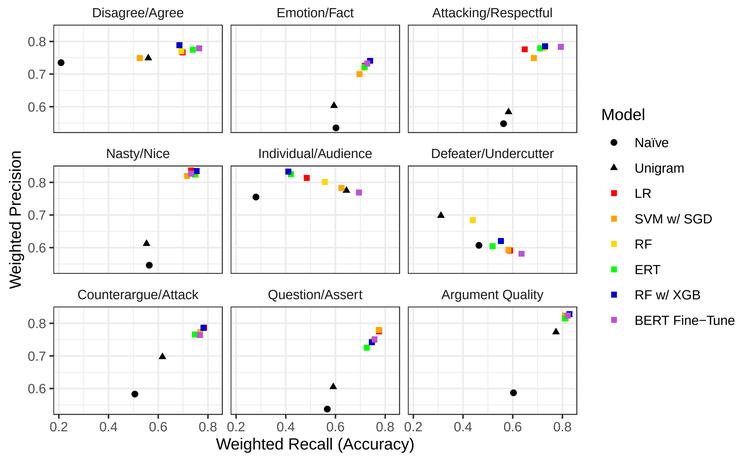


**Figure 2:** Weighted Precision and Recall. Weighted recall is mathematically equivalent to accuracy. Black points represent baselines, colored points represent deep learning models.

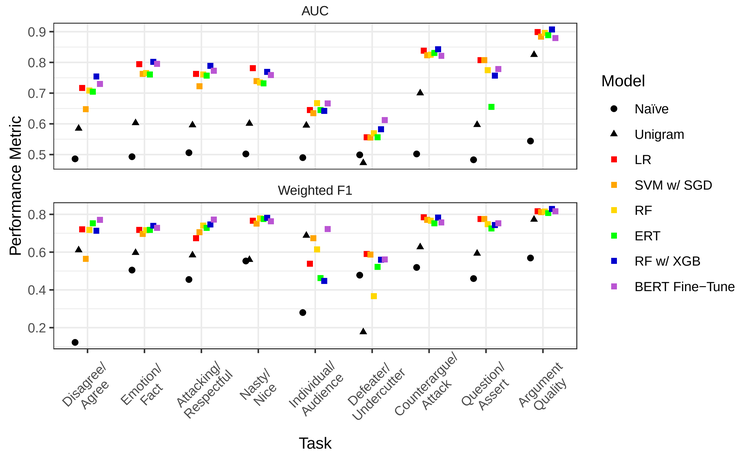


**Figure 3:** Area Under the ROC Curve and Weighted F1 Score. Black points represent baselines, colored points represent deep learning models.

| Task | Best F1 | Improvement over Naïve Baseline: Absolute | Improvement over Naïve Baseline: Relative | Improvement over Lexical Baseline: Absolute | Improvement over Lexical Baseline: Relative |
|---|---|---|---|---|---|
| **Disagree / Agree** | 77.1% | 64.9% | 531.967% | 16% | 26.187% |
| **Emotion / Fact** | 73.9% | 23.4% | 46.336% | 14.2% | 23.786% |
| **Attacking / Respectful** | 78.7% | 33.2% | 72.967% | 20.3% | 34.76% |
| **Nasty / Nice** | 78.1% | 22.8% | 41.23% | 22.1% | 39.464% |
| **Individual / Audience** | 72.3% | 44.3% | 158.214% | 3.5% | 5.087% |
| **Defeater / Undercutter** | 59% | 11.2% | 23.431% | 41.3% | 233.333% |
| **Counterargue / Attack** | 78.5% | 26.6% | 51.252% | 15.8% | 25.199% |
| **Question / Assert** | 77.6% | 31.6% | 68.696% | 18.3% | 30.86% |
| **Argument Quality** | 82.8% | 25.9% | 45.518% | 5.5% | 7.115% |

**Table 2:** Absolute and Relative Improvement Over Baselines

| Task | Citation | Metric | Previous | New | Absolute Gain | Relative Gain |
|---|---|---|---|---|---|---|
| **Disagree / Agree** | Wang and Cardie (2014) | F1 | 63.57% | 77.1% | 13.53% | 21.28% |
| **Disagree / Agree** | Abbott, Walker, et al. (2011) | Acc. | 68.2% | 76.5% | 8.3% | 12.17% |
| **Emotion / Fact** | Oraby et al. (2015) | F1 | 46.2% | 73.9% | 27.7% | 59.96% |
| **Nasty / Nice** | Lukin and Walker (2013) | F1 | 69% | 78.1% | 9.1% | 13.19% |

**Table 3:** Absolute and Relative Improvement Over Previous State-of-the-Art Metrics

## Example Classifications

| Document | Disagree / Agree: D | Disagree / Agree: A | Emotion / Fact: E | Emotion / Fact: F | Counterargue / Attack: C | Counterargue / Attack: A |
|---|---|---|---|---|---|---|
| So why do the elderly marry? Why do those who know they can never conceive children marry? Marriage is not simply a way of providing a stable home life for children at all. Marriage today is for companionship and love. | ✓ | | ✓ | | ✓ | |
| Cost would certainly be an issue. If the adoptive parents were willing to take on that cost, that would be best. I wonder, though, how much demand there really is out there for adopted babies. | | ✓ | ✓ | | ✓ | |
| Another critical error. Evolution does not say life is chaotic. Mutations are random, not natural selection or evolution. It is a heavily guided process. We have already shown that the primary building blocks of life can be simulated and created. They can even explain how the human eye developed—a recent discovery. | ✓ | | | ✓ | ✓ | |
| I know what you mean. I also wonder how they think their children or grandchildren will be paying for their medical bills. Note, by way of comparison, this Pittsburgh local TV clip of the September March for Jobs. See how clearly and intelligently the protestors articulate their objectives. | | ✓ | | ✓ | | ✓ |
| What about the pain and death inflicted upon the innocent women and girls forced to give birth? Have you no regard for them? | ✓ | | ✓ | | ✓ | |

**Table 4:** Predicted classifications of exemplar documents from test set. For purposes of demonstration, examples were selected for variation on classes. All classifications produced using random forest classifier with XGBoost.

## Highlights

- Argument mining on online discourse is a **challenging task—**messages are often brief and do not conform to standard language structures (e.g. sentence fragments), and aspects of argumentation can be highly subjective.
- Lexical models and other **popular methods of feature extraction may not be appropriate** for use on these data. Deep learning may present a more reliable option for understanding political discussion and argumentation.
- As usual, **no free lunch**: Different models perform well on different tasks, though fine-tuned BERT and random forests with XGBoost tend to perform best across all tasks.
- **Deep learning architectures improve** upon baseline F1 scores by 5.1%-233.3% and improve upon state-of-the-art F1 scores from lexical models by 12.2%-60%.

## References

Abbott, Rob, Brian Ecker, et al. (May 2016). "Internet Argument Corpus 2.0: An SQL Schema for Dialogic Social Media and the Corpora to Go With It". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 4445–4452.

Abbott, Rob, Marilyn Walker, et al. (June 2011). "How Can You Say Such Things?!? Recognizing Disagreement in Informal Political Argument". In: *Proceedings of the Workshop on Language in Social Media*. Portland, OR: Association for Computational Linguistics, pp. 2–11.

Devlin, Jacob et al. (May 2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". Pre-Print. Google AI Language. arXiv: 1810.04805.

Lukin, Stephanie and Marilyn Walker (June 2013). "Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue". In: *Proceedings of the Workshop on Language in Social Media*. Atlanta: Association for Computational Linguistics, pp. 3–40.

Oraby, Shereen et al. (May 2015). "And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue". In: *Proceedings of the 2nd Workshop on Argumentation Mining*. Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies. Denver.

Swanson, Reid, Brian Ecker, and Marilyn Walker (Sept. 2015). "Argument Mining: Extracting Arguments from Online Dialogue". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague: Association for Computational Linguistics, pp. 217–226.

Wang, Lu and Claire Cardie (June 2014). "Improving Agreement and Disagreement Identification in Online Discussions with A Socially-Tuned Sentiment Lexicon". In: *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Baltimore: Association for Computational Linguistics, pp. 97–106.