Mining Arguments in Presidential Campaign Debates

PM Mining Opinions & Arguments WiSe 21/22, Universität Potsdam

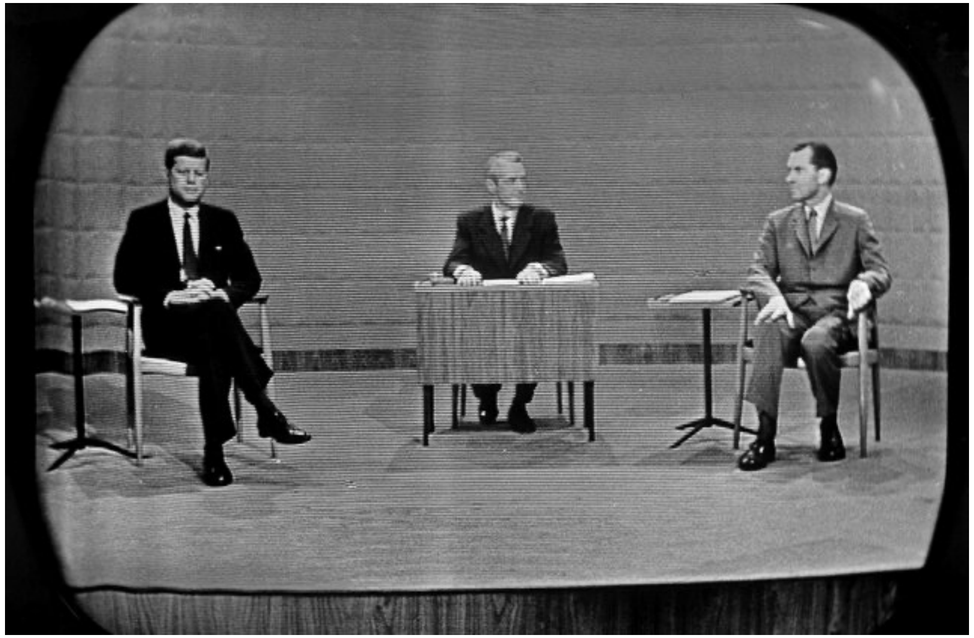
Prof. Dr. Manfred Stede

Team Chen Peng, Mariia Poiaganova, Milena Voskanyan

11. February 2022

Plan

- Motivation
- Tasks
- Corpus and Data
- Experiments
- Replication summary
- Further steps and extension
- Conclusions so far



John F. Kennedy and Richard Nixon debates, 1960

Project Motivation

Paper *"Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates", 2019*

Shohreh Haddadan, Elena Cabrio, Serena Villata

- Replication of the two main classification tasks in the paper
- To match and if possible, exceed the classification results reported by the authors
- Two tasks
  - Task 1: Binary classification of **all** sentences, based on whether they contain an argument component or not
  - Task 2: Binary classification of sentences which **contain an argument component,** based on whether they contain a claim **or** a premise (evidence)

Corpus

- Argument mining in past debates from U.S. presidential campaigns
  - 42 debate transcripts from 1960 - 2016, *USElecDeb60To16 v.01 dataset*

- Two main argument components: claims and premises
  - 6 expert annotators involved in the definition and annotation process

Claims

- *"can be <u>a policy advocated</u> by a party or a candidate to be undertaken which needs to be justified in order to be accepted by the audience"*
- *"provide <u>judgments about the other candidate or parties</u>"*
- *"taking a <u>stance towards a controversial subject</u>, or an opinion towards a specific issue"*

  - **BUSH:** Over 60 nations involved with disrupting the trans-shipment of information and/or weapons of mass destruction materials. And **<u>[we've been effective]</u>**. *[We busted the A.Q. Khan network. This was a proliferator out of Pakistan that was selling secrets to places like North Korea and Libya]. [We convinced Libya to disarm].*

  - Bush is defending the decisions taken by his administration by claiming that his policy has been effective

Premises

- *"Premises are <u>assertions</u> made by the debaters for supporting their claims (i.e., reasons or justifications). A type of premise commonly used by candidates is referring to <u>past experience</u>."*
- *"Statistics are very commonly used as evidence to justify the claims"*

    - **CARTER:** *[Well among my other experiences in the past, I've - I've been a nuclear engineer, and did graduate work in this field].* **<u>[I think I know the - the uh capabilities and limitations of atomic power].</u>**

Dataset

- USElecDeb60To16 v.01
  - https://github.com/ElecDeb60To16/Dataset

Dataset statistics

Total, sent	Arg, sent	Non-arg, sent	Claims, sent	Premises, sent
29.621	22.280	7.252	11.964	10.316

	Train, sent	Test, sent	Validation, sent
Task 1	14.044	8.455	7.033
Task 2	10.464	6.575	5.241

Experimenal settings

For each of the two tasks, the authors used:

- Tf-idf with unigrams + linear SVM. Experiment 1
- Engineered Features + rbf SVM. Experiment 2
- FastText embeddings + LSTM. Experiment 3
- Engineered Features + FFNN. Experiment 4

- Engineered Features are: tf-idf for each unigram, n-grams (bi- and tri-grams), NER, POS for adverbs and adjectives, tenses for verbs, syntactic features, discource connectives, sentiment of a sentence

Experiment 1 : Arg vs Non-arg

- Tf-idf features, unigrams, full vocab (9.833), SVM with **linear kernel**, C=10
- Kept the linear kernel, tuned **gamma** and **C** parameters on validation set

	[Haddadan et al., 2019] F1-measure	Exact replication, F1-measure	Replication after tuning, F1-measure
Baseline	0.551		
Argument	0.855	**0.896**	0.894
Non-argument	**0.486**	0.441	0.457
Both classes	0.737	0.795	**0.797**

Gamma=1, C=1

Experiment 1: Claims vs Premises

- Tf-idf features, unigrams, full vocab, SVM with **linear kernel**, C=10
- Kept the linear kernel, tuned **gamma** and **C** parameters on validation set

	[Haddadan et al., 2019] F1-measure	Exact replication, F1-measure	Replication after tuning, F1-measure
Baseline	0.35		
Claim	**0.685**	0.641	0.670
Premise	0.599	0.579	**0.610**
Both classes	**0.643**	0.610	**0.641**

Gamma=1, C=1

Experiment 2 : Arg vs Non-arg

- Tf-idf features: uni-, bi- and tri-grams (vocab=10.000), NER-features, POS-features, discourse connectives and sentiment of a sentence
- SVM with **rbf kernel**, C=10
- Tuned **gamma** and **C** parameters, tried both **rbf and linear** on validation set

	[Haddadan et al., 2019] F1-measure	Exact replication, F1-measure	Replication after tuning, F1-measure
Baseline	0.551		
Argument	**0.916**	0.879	**0.895**
Non-argument	0.433	**0.509**	**0.491**
Both classes	**0.823**	0.797	**0.805**

Gamma=1, C=1, kernel=linear

Experiment 2 : Claims vs Premises

- Tf-idf features: uni-, bi- and tri-grams (vocab=10.000), NER-features, POS-features, discourse connectives and sentiment of a sentence
- SVM with **rbf kernel**, C=10
- Tuned **gamma** and **C** parameters, tried both **rbf and linear** on validation set

	[Haddadan et al., 2019] F1-measure	Exact replication, F1-measure	Replication after tuning, F1-measure
Baseline	0.35		
Claim	**0.717**	0.65	0.674
Premise	0.581	0.59	**0.606**
Both classes	**0.651**	0.62	**0.641**

Gamma=1, C=1, kernel=linear

Experiment 3 : Arg vs Non-arg

- **FastText** word embeddings, Neural Network with two bidirectional **LSTM** layers

	[Haddadan et al., 2019] F1-measure	Exact replication, F1-measure
Baseline	0.551	
Argument	**0.913**	**0.895**
Non-argument	**0.547**	0.460
Both classes	**0.843**	0.798

Experiment 3 :Claims vs Premises

- **FastText** word embeddings, Neural Network with two bidirectional **LSTM** layers

	[Haddadan et al., 2019] F1-measure	Exact replication, F1-measure
Baseline	0.35	
Claim	**0.819**	0.661
Premise	**0.710**	0.634
Both classes	**0.673**	**0.648**

Experiment 4 : Arg vs Non-arg

- Engineered features, **Feed Forward Neural Network**, two hidden layers with 64 and 32 neurons for the 1st and 2nd hidden layer

	[Haddadan et al., 2019] F1-measure	Exact replication, F1-measure
Baseline	0.551	
Argument	**0.872**	**0.858**
Non-argument	**0.498**	**0.483**
Both classes	**0.800**	**0.775**

Experiment 4 :Claims vs Premises

- Engineered features, **Feed Forward Neural Network**, two hidden layers with 64 and 32 neurons for the 1st and 2nd hidden layer

	[Haddadan et al., 2019] F1-measure	Exact replication, F1-measure
Baseline	0.35	
Claim	**0.667**	0.629
Premise	**0.611**	0.596
Both classes	**0.640**	0.613

Replication results summary

…so far

- Replicated most of the engineered features and all models
- Exceeded the given results on 1 out 8 experiments
- Got comparable results on 4(6) out of 8 experiments

Next and final steps

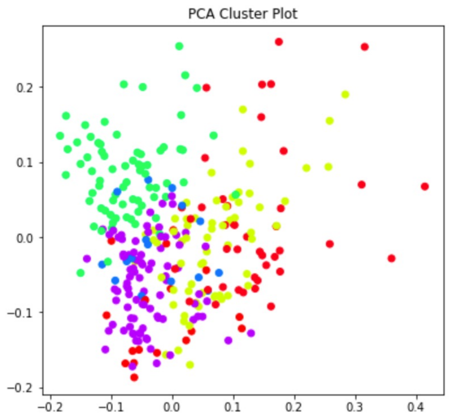
- Improve performance on experiments with LSTM and possibly FFNN
- Add syntactic features
- Perform the linguistic analysis of errors on our best model

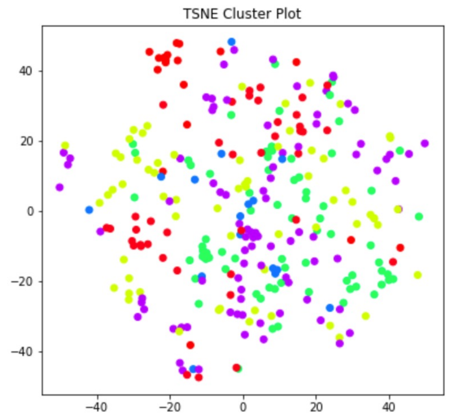
- Extension ideas

Extension directions (and trials...)

- Three experimental directions

1. **clustering**
- 6 types of evidences (Al Khatib, 2016)
- what if clustering algorithms could capture them?
...unclear borderline from type to type





2. **tried BERT** (pretrained, ‘bert-base-cased’)
avg accuracy 0.82 on task 1 (arg vs non-arg)

3. **adding other features**: signular first person pronouns (Stab and Gurevych, 2014), argument words as unigrams (Nguyen and Litman, 2015), comparative and superlative forms of adj and adv (Nguyen and Litman, 2016)

Conclusions so far

- **Reproducibility crisis in NLP/AI**
- **70%** of scientists reporting **failure to reproduce** someone else's results, and more than half reporting failure to reproduce their own (Baker, 2016)
- Only **15%** of AI studies share their code (state of AI report, 2021)
- The **challenges** we faced: no source code, very brief techincal side explanations, no mention of tools…
- The **good part**: the original train-test data, really positive experience contacting the authors

Conclusions so far

- Our changes to the original research/contributions

- preprocessing (lowercasing and punctuation removal)
- tuned hyperparams
- used other tools (our SpaCy, vader vs authors' CoreNLP; genism for FastText implementation)

Related Materials

Literature

- Baker, M. (2016). Reproducibility crisis. Nature, 533(26):353–66.
- Haddadan, S., Cabrio, E., & Villata, S. (2018). Annotation Guideline for Argumentation Structure in Political Debates Dataset, https://github.com/ElecDeb60To16/Dataset/blob/master/ElectDeb60To16_Guidelines.pdf
- Haddadan, S., Cabrio, E., & Villata, S. (2019). Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates. *ACL*.
- Stab, C., Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland, pages 1501–1510. ACL.
- Nguyen, Huy & Litman, Diane. (2016). Context-aware Argumentative Relation Mining. 1127-1137. 10.18653/v1/P16-1107.
- State of AI Report, 2021. https://www.stateof.ai/

Tools

- Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
- Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.