**Language Model/ Dataset Paper**

**Venue & Timeline:**

ACL 2021,

|  |  |
| --- | --- |
| Abstract deadline (*long & short papers*): | January 25, 2021 |
| Submission deadline (*long & short papers*): | February 1, 2021 |

Dec 1-11 – Find novelty

Dec 12 – 30 – Experiments

Jan – Write up experiments

Feb – Paper writing

**Paper to reference for style:** GoEmotions: A Dataset of Fine-Grained Emotions

**Data:** Current set of annotations where speaker tells the truth against (Reasoning/ Game Move/ Rapport/ Share Information)

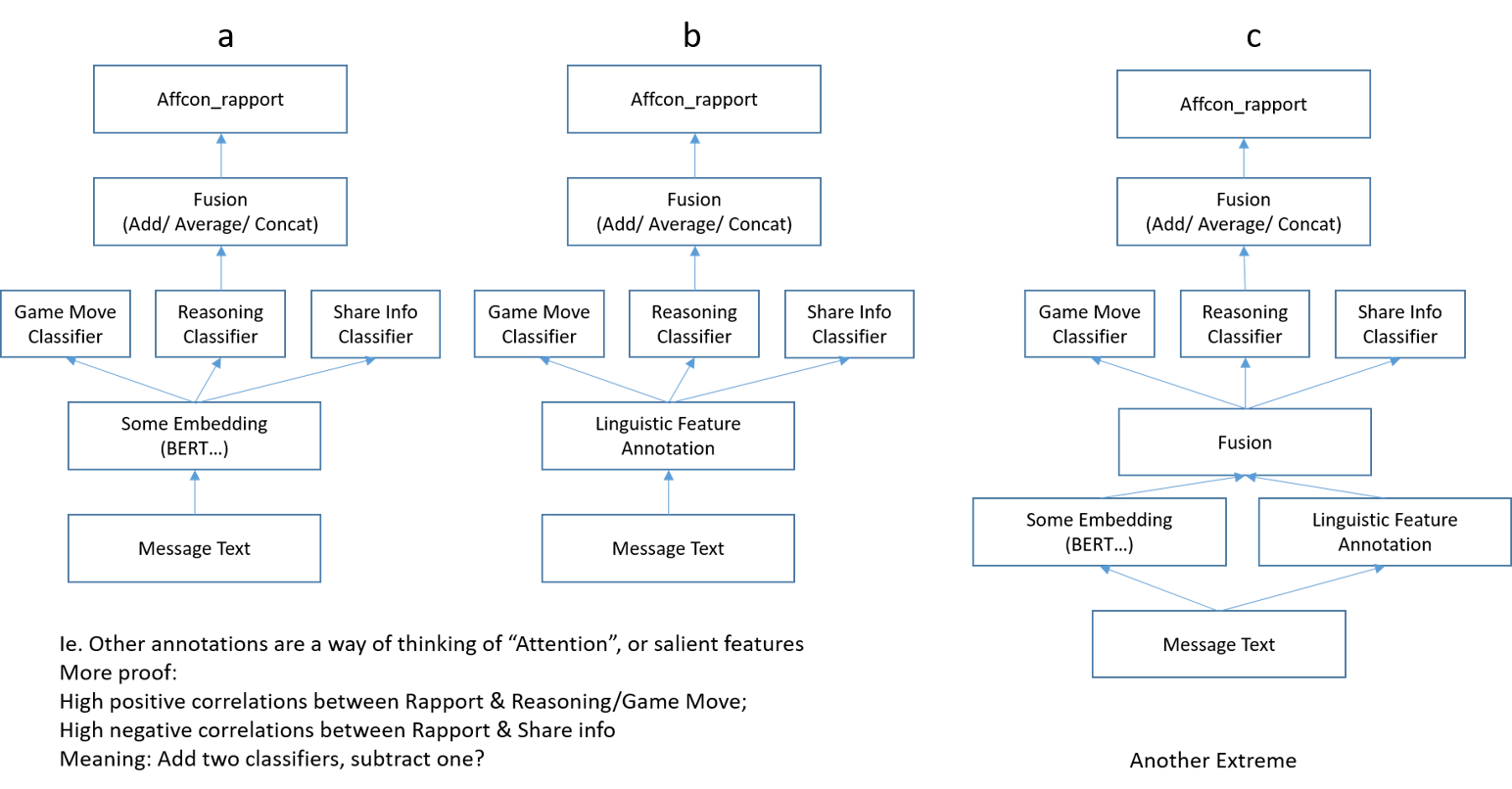
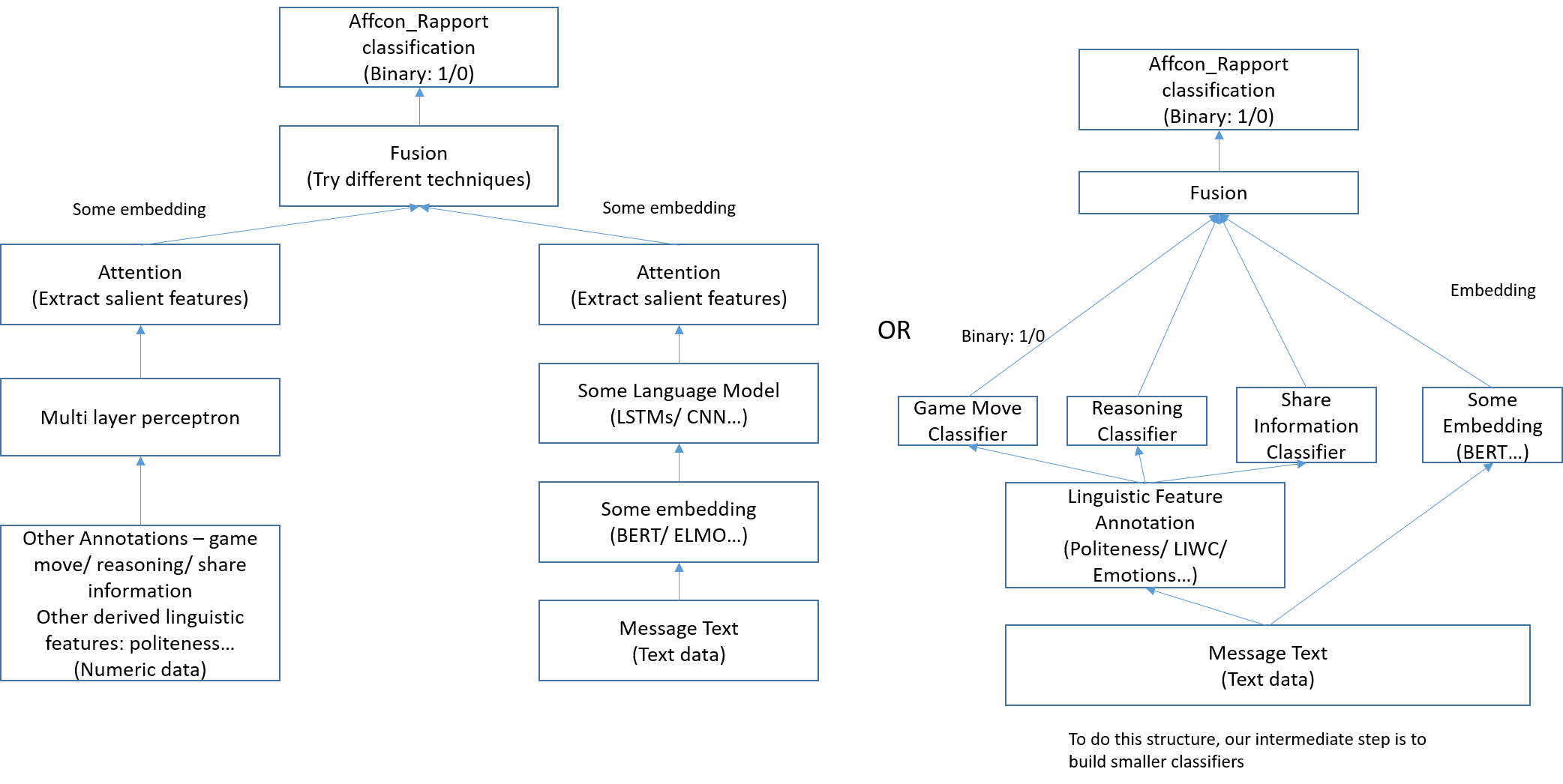
**Possible Direction**: What is the best way to enrich the data for the model? **[Novelty could be in enrichment construction or fusion construction]**

Enrich stuff in hierarchical models – identify different points in the NN pipeline to incorporate language features & examine which works out best

+ Figures with different training set size/ epochs

Enrich datasets with generative adversarial networks

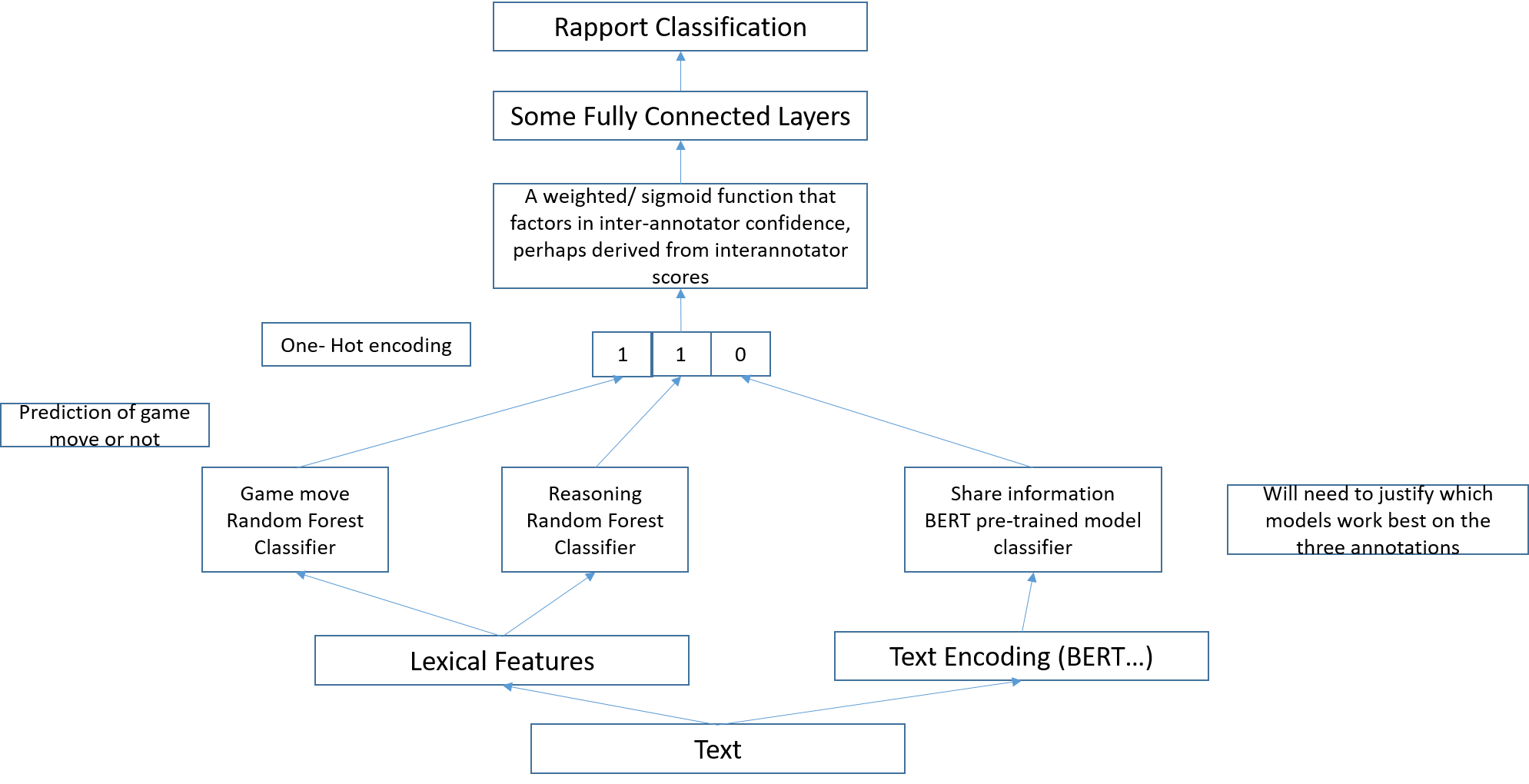
**Proposed Structure (Lynnette)**



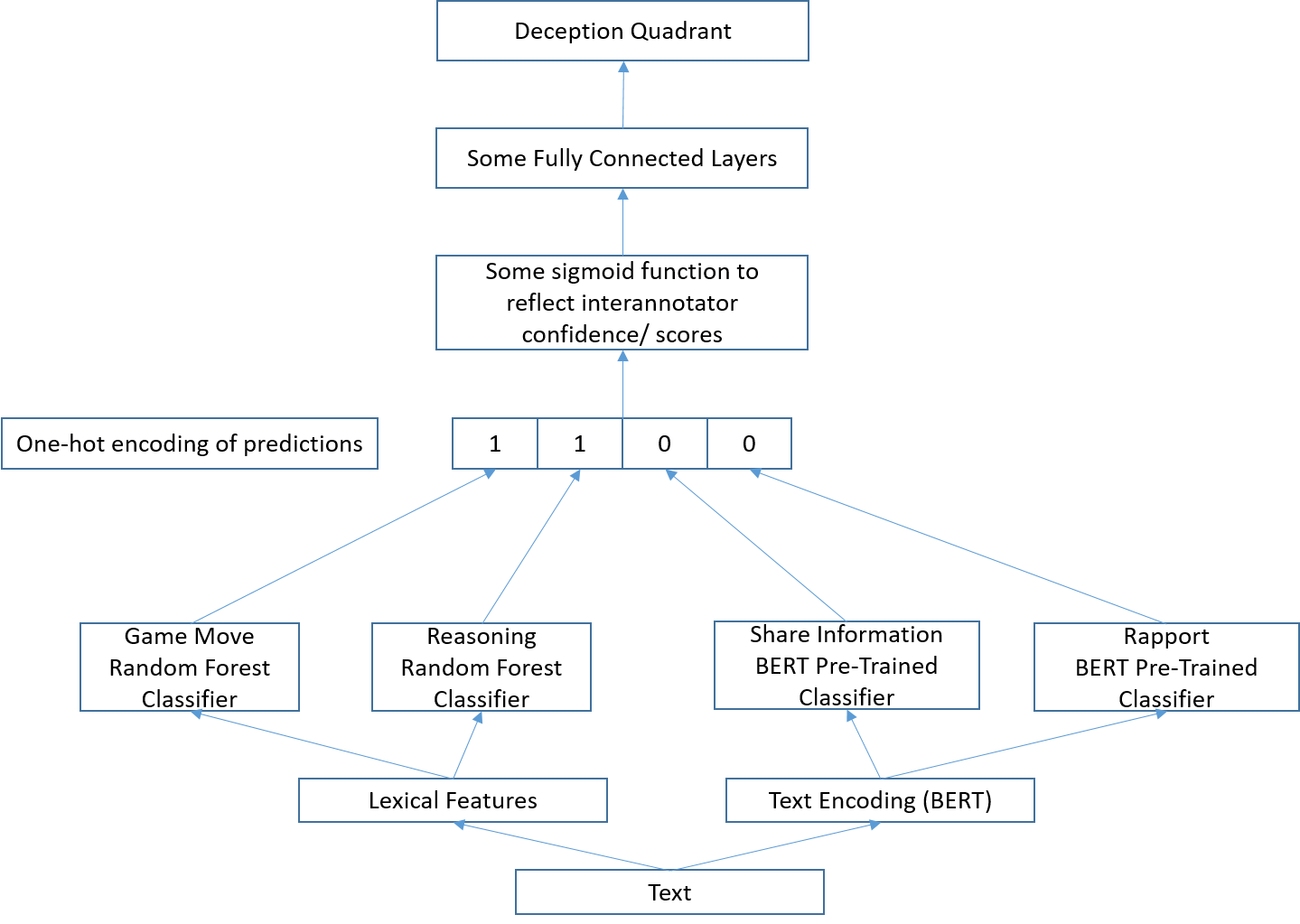
Structure a:

* Fusion of game move + reasoning + share info: 50%
* Fusion of game move + reasoning – share info: 60%

Structure b: 20%



Note: can remove linguistic feature annotation to just an embedding



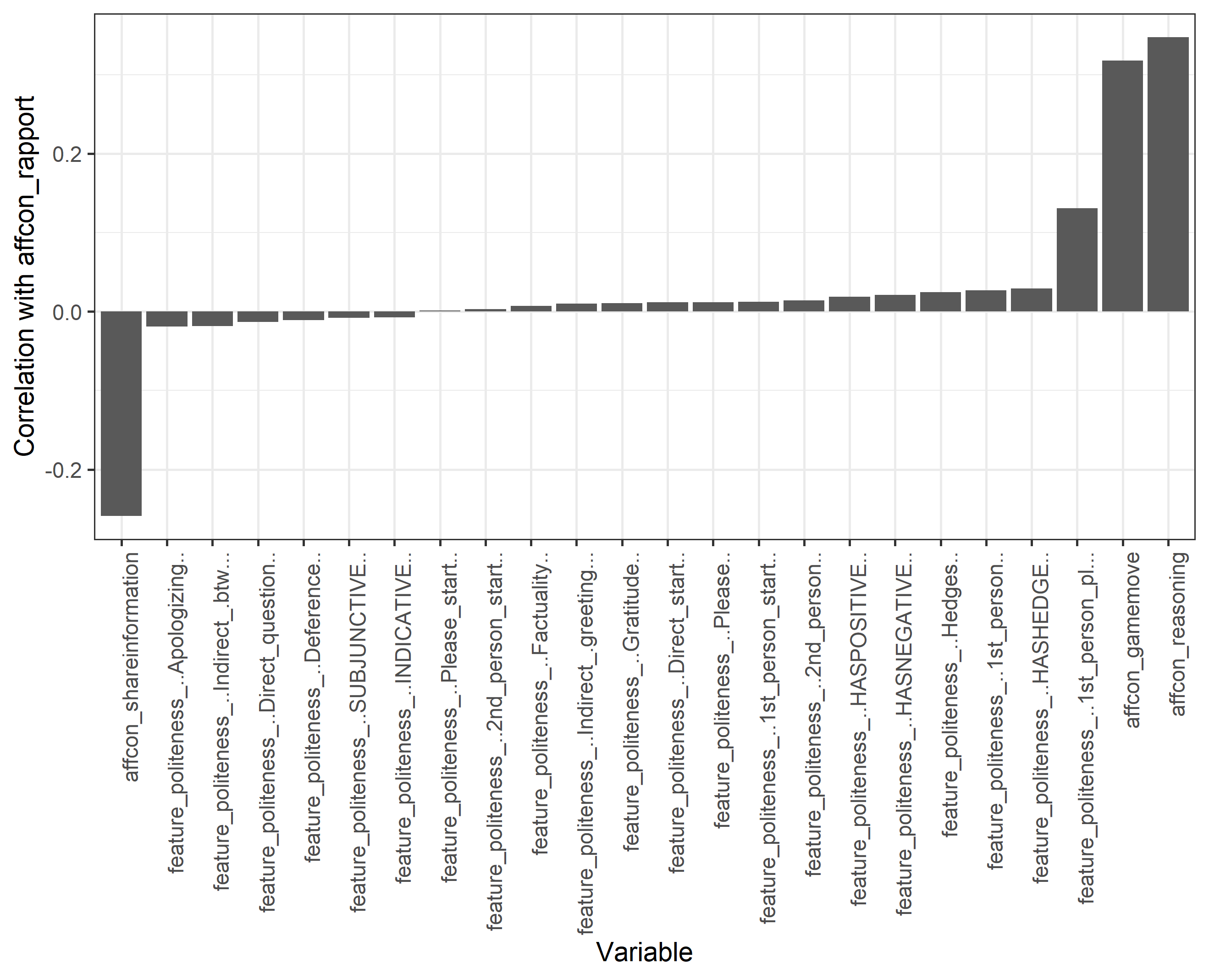
**4-grams**

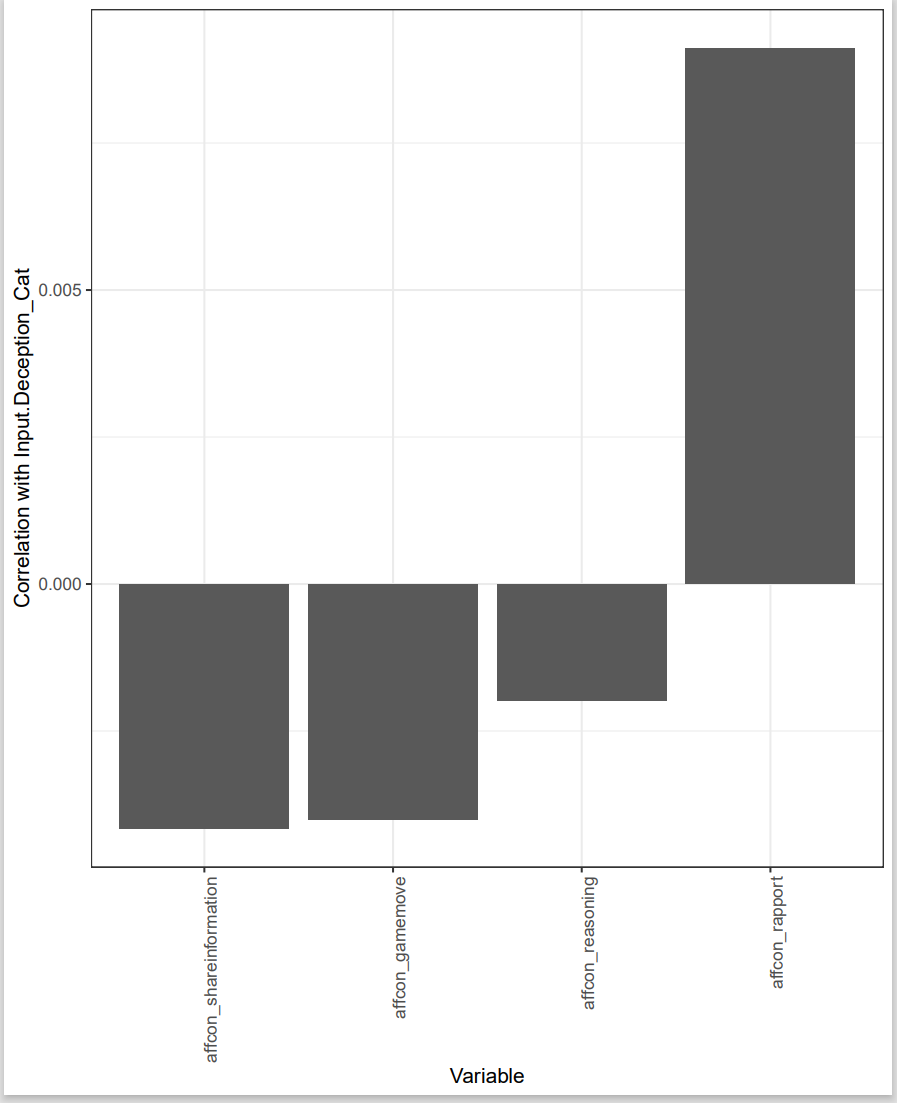
|  |  |
| --- | --- |
| **Affcon\_rapport = True** | **Affcon\_Rapport = False** |
| Move east to counter  Mostly relying on the  Move into Serbia keeping  Most likely to be  Move combinations if  Move either direction  Move cautiously  Move by  Move either  *Could we say a lot on game movement discussion, and gives choices (“either”), and adjectives (“cautiously”)* | Mind terribly if moved  Minute they are asking  Military lined up at  Might be more worth it  Mind moving into  Misclicked when placing  Might move  Mind terribly  Miss it  Misery yknow  *Could we say a lot of minding, trying to be nice but actually shooting a dagger HA* |

**Pronouns**

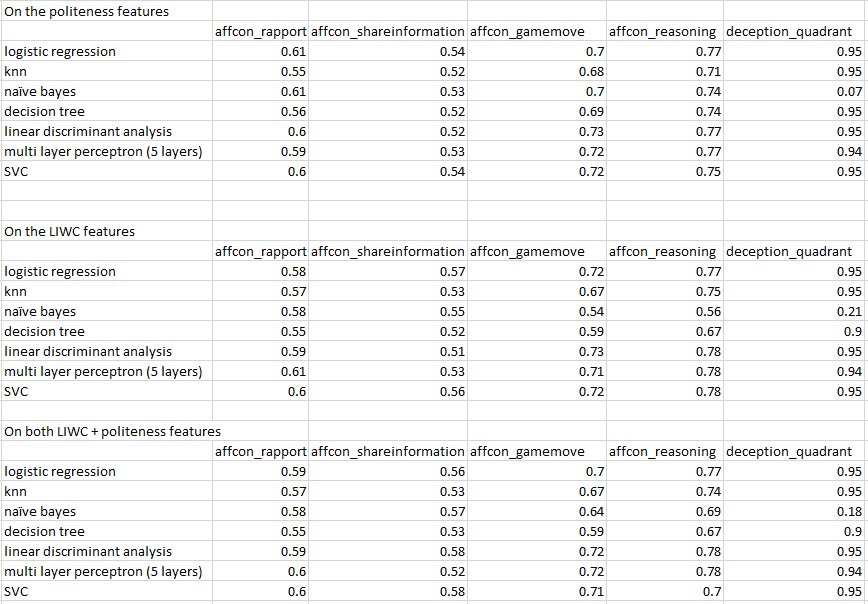
|  |  |  |  |
| --- | --- | --- | --- |
|  | **First person %** | **Second Person %** | **Third Person %** |
| Average |  |  |  |
| Affcon\_rapport=True | 3.93847 | 3.325816 | 1.024625 |
| Affcon\_rapport=False | 4.15769 | 3.41641 | 1.0598 |
|  |  |  |  |
| Per text |  |  |  |
| Affcon\_rapport=True | 5.7562  Std 5.9032 | 3.7712  Std 5.2289 | 1.0721  Std 3.057 |
| Affcon\_rapport=False | 4.8014  Std 5.7312 | 4.0585  Std 5.6784 | 1.0823  Std 3.1328 |

**Correlation Plots**





**Lexical Features**



**Against affcon Rapport (only Text) + Early stopping**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Number of epochs | Batch Size | Precision | Recall | F1 Score |
| ROBERTA + Roberta Transformer | 3 | 64 | 0.586 | 0.586 | 0.586 |
| BERT + 3 layers BiDirectional LSTM | 32 | 64 | 0.5063 | 0.91485 | 0.6517 |
| BERT + 3xCNN | 32 | 64 | 0.62 | 0.62 | 0.62 |
| BERT + RNN | 32 | 64 |  |  |  |
| BERT + CNN-LSTM | 32 | 16 | 0.7422 | 0.5393 | 0.6183 |
| BERT + LSTM | 32 | 64 | 0.4999 | 0.9999 | 0.6665 |
| BERT + BERT Transformer | 32 | 64 | 0.4956 | 0.4956 | 0.4956 |
| ConvLSTM | 32 | 64 | OUT OF MEMORY GG |  |  |
| GPT2 + GPT transformer | 32 | 64 | 0.3957 | 0.3957 | 0.3957 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Number of epochs | Batch Size | Precision | Recall | F1 Score |
| Numerical MLP | 32 | 64 | 1.00 | 0.586 | 0.4061 |
| Combined  Text LSTM +  Numerical MLP | 32 | 16 |  |  | 0.50 |

**References To Consider:**

<https://arxiv.org/pdf/2004.13609.pdf>

**Our Paper Structure: +cool title name**

Introduction: Intro to the Diplomacy game, rapport (+sample texts)

Related work: Past Diplomacy papers by Cornell lab, some stuff on Rapport/affcon

Dataset description:

* Constructing dataset for annotation, i.e. splitting them into sentences, deduplication those that are similar (eg moves)
* Explain Rapport/ Game Move/ Reasoning/ Share Information – paraphrase from how we put in MTurk + examples
* When we construct this dataset, we seek to maximize the following objectives… OR what is our objectives of the dataset
* Annotation using MTurk, rater interface, summary statistics

Data Analysis

* Summary statistics + table/ distribution graphs
* Interrater correlation
* Correlation among annotated categories
* Performing LIWC + correlation among categories (correlation plots)
* Rapport vs talkativeness

Modelling

* N-grams of each annotated category (reference Table3 of GoEmotions)
* Pronouns
* Language model architecture + parameters + experiments + results
* **Novelty experiment**: fusion of data

Conclusion

* Conclusion, future work
* Data disclaimer/ limitations

**Language Models Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model/accuracy type** | **gamemove** | **reasoning** | **rapport** | **shareinformation** |
| Embedding + LSTM (Train acc) | 0.7186 (Epochs: 3/15) | 0.7602 (Epochs: 2/15) | 0.6108 (Epochs: 2/15) | 0.6400 (Epochs: 3/15) |
| Embedding + LSTM (Val acc) | 0.7268 (Epochs: 3/15) | 0.7772 (Epochs: 2/15) | 0.6092 (Epochs: 2/15) | 0.5985 (Epochs: 3/15) |
| Embedding + CNN (Train acc) | 0.7201 (Epochs: 3/15) | 0.7631 (Epochs: 3/15) | 0.6199 (Epochs: 3/15) | 0.6406 (Epochs: 3/15) |
| Embedding + CNN (Val acc) | 0.7077 (Epochs: 3/15) | 0.7736 (Epochs: 3/15) | 0.5790 (Epochs: 3/15) | 0.5766 (Epochs: 3/15) |

**\*\*Note to KY: Use all Validation Accuracy? Fill in Precision/ Recall/ F1 too**

Pr – Precision, R – Recall, F1 – F1 Score

**Neural Network Models**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Gamemove** | | | **Reasoning** | | | **Rapport** | | | **ShareInformation** | | |
|  | Pr | R | F1 | Pr | R | F1 | Pr | R | F1 | Pr | R | F1 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |

**Using Lexical Features (LIWC + Politeness)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Gamemove** | | | **Reasoning** | | | **Rapport** | | | **ShareInformation** | | |
|  | Pr | R | F1 | Pr | R | F1 | Pr | R | F1 | Pr | R | F1 |
| Logistic Regression |  |  |  |  |  |  |  |  |  |  |  |  |
| KNN |  |  |  |  |  |  |  |  |  |  |  |  |
| Naïve Bayes |  |  |  |  |  |  |  |  |  |  |  |  |
| Decision Tree |  |  |  |  |  |  |  |  |  |  |  |  |
| Random Forest |  |  |  |  |  |  |  |  |  |  |  |  |
| Linear Discriminant Analysis |  |  |  |  |  |  |  |  |  |  |  |  |

**Comments on 4 Dec from Nyati**

1. Rapport annotation update

broad strategy: go for novelty not rigor in one month

1. Something for lynnette to consider: predict trust not rapport.
2. IDEA 1

features add karna : enough variations to try, hard to claim novelty.

techniques: this would need a good understanding of prior art and a lot of implementations in one month

1. IDEA 2

fusion: in input vs output space.

multitask/transfer learning.

alongwith where to add features.

put it as, for one level of questions the confidence is low. second is high. learn based on the specific classifiers. to predict the major class.

this year's emnlp/aaai literature survey

group truth: honest annotators of other labels.

honesty/confidence in annotator. reinforcement learning reward dependent on that.

train on high-confidence annotations

previous work: summarization, recommendation of clickstream.

simple reward function: sigmoid/ whatever

[1:20 PM, 12/4/2020] Dr Kokil Jaidka: I think the core message is go for an exciting problem since we don't have enough time to do extensive lit reciew.

[1:21 PM, 12/4/2020] Dr Kokil Jaidka: She suggested use a fusion idea/approach as a solution to that exciting problem. And exciting problem could be how to deal with shitty annotations

[1:22 PM, 12/4/2020] Dr Kokil Jaidka: Just think about her message and we can figure out a novel problem to attack. Something that we can defend on novelty at least

**6 Dec 2020**

Data with quality info is at: <https://www.dropbox.com/sh/0wkyaa91cs7fwtk/AAAAypnp2f6hIC1D5V3x9Epna?dl=0>

I discarded all the intermediate cleaning steps I have done before to prepare this data, so that we can stay true to the research problem for this paper.

* The labels are based on at least 3 people agreeing. As you know, there is always a tie breaker. So the labels are in the “\_label” column in both files.
* The only quality characteristics we can use is the agreement on the label (provided in the “\_*pc*\_agree” column) or the time taken on the hit.
  + Based on the time taken, I removed those which took too little time on the hit (less than 1 standard deviation), which leads us to suspect a bot. So by removing anyone who took less than 80 seconds and removing those HITs with less than 3 annotations, we get 4000 good quality hits (“conf\_good\_agg\_withpc.csv”)
* Although there are more fine-grained annotations on this full data, I think we should stick with gamemoves, reasonsing, shareinformation, and rapport (not use the other finegrained ones which could still be dirty)

Let me know if you see any improvement to your models esp with the conf\_good data! I really appreciate how much time and effort yall are putting into this 😊

Talk to you as planed, I’ll send an invite for thurs evening.