**AffCon2021 Log**

### 24 Dec 2020

- Cleaned pipeline-onehot.ipynb so that all variants of Throughput, Worktime, PC agreement are elaborated, and wrote 1 function to allow an easier way to choose variants, and to run the entire notebook.

- Added upright V/U-shape throughput (TP4 option) in pipeline-onehot.ipynb.

- Added textlength variant for weighted onehot encodings in pipeline.ipynb.

- Added new notebook pipeline-keras\_concat.ipynb and transferred all cells revolving Keras Concatenate layer from pipeline-onehot.ipynb over. Worked a little on getting the code to work by “re-wiring” how Keras Concatenate layer should pull the 3 (4) models together (so that we are predicting on 3 (4) vectors from the 3 (4) annotations, then predicting 1 variable, either rapport (or shareinfo, respectively)). Results: not good, 6% macro F1 accuracy, worse than pipeline-onehot.ipynb models. Did not work on weighting the Keras Concatenate inputs by TP, WT, or PC, since results aren’t good.

######################################

## Weighted Onehot Encoding options ##

##############

# Throughput #

##############

# TP1: weighted by 1 average per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

# TP2: weighted by 1 linear variance per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

# TP3 + k: weighted by 1 inverted k-power U-shaped variance per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

# TP4 + k: weighted by 1 upright k-power U-shaped variance per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

# (For TP3 & TP4, k=1 results in V-shaped variance, and as k>1 increases, sides will curve into U-shaped variance)

############

# Worktime #

############

# WT1: weighted by 1 average per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

# WT2: weighted by 1 linear variance per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

################

# PC agreement #

################

# PC1: weighted by 1 average per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

# PC2: weighted by 1 linear variance per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

# PC3: weighted by 1 PC agreement weight per annotation in each OHE, i.e. (a, b, c, d) -> (w1\*a, w2\*b, w3\*c, w4\*d)

#####################

# Input text lenght #

#####################

# TL1: weighted by 1 normalised number of characters per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

# TL2: weighted by 1 normalised number of words per set of OHE, i.e. (a, b, c, d) -> (w\*a, w\*b, w\*c, w\*d)

### 22 Dec 2020

Low dimensionality to input features for lexical methods vs throwing neural models => at what point would the information be useful?

Eg. if we keep the best possible data, what’s the best possible performance vs if we keep the worst possible data, what’s the best possible performance?

Worktime: Get a bunch of people who did it very quickly vs number of people that did it at average speed

Variance vs U-shaped variance: do hand-picked quality check

Text length: really short observations vs really long observation texts (poor?) vs average observation texts (good?)

Lexical features pick out false negatives more

Scikit learn metrics

'macro':

Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.

'weighted':

Calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters ‘macro’ to account for label imbalance; it can result in an F-score that is not between precision and recall.

* Should use weighted (or default)

To do:

* Quality checks: Get a bunch of people at extreme ends of worktime/ variance/ text length/ pc agreement etc vs the average/central to see the F1 scores
* Change one-hot encoding to ~~percentage encoding for pipeline models~~

+ keras concatenate the models

* Finish up running all the models + other neural network models - to report everything in supplement and the best models in main text   
  + Use n-grams as features

PAPER STRUCTURE

* Evaluate quality features, and show distribution of quality features in dataset, validate how our selection functions are correct
* Show downstream prediction, which we did for one dataset
* Show this ideas work for another dataset, which is blind to us

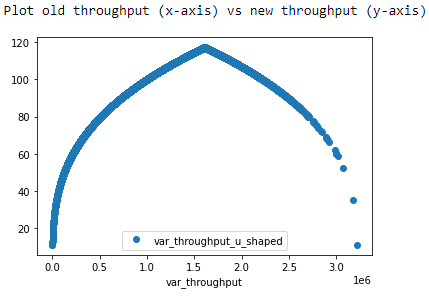
### 21 Dec 2020

Feature Extraction Penn Treebank features to yield relevant features

* For all meta data: Pearson Correlation yields 0 features, Wrapper method with sm.OLS yields 34 features (and faster), embedded method yields 1 feature

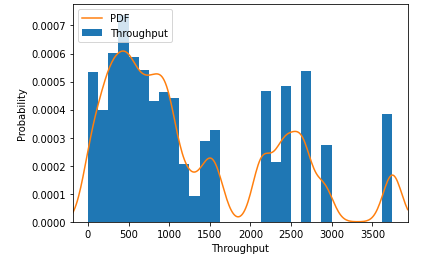
Feature selection doesn’t make a difference in the F1 scores, but at least shows which are the salient features

Throughput weighted by inverted U-shaped variance

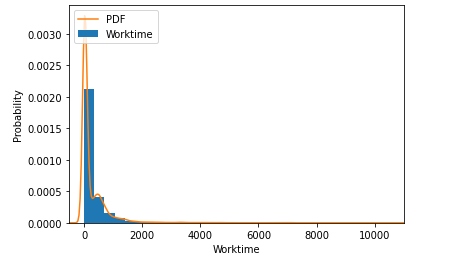


### 17 Dec 2020

Generally two peaks for Probability vs Throughput



Worktime has a very nice shape



### 15 Dec 2020

* Higher throughput, better the results. Or else a flattening
* Worktime: people going too fast, might be bots. -> Inverted U instead of linear
* U shape: x-axis time taken, y-axis number of people who have that time. People who are at average are most likely to be right
* Smoothing with Log takes care of things when we plot stuff

TO DO:

* Use variance instead of average for sigmoid
* Use logistic regression stacked + sigmoid
* Use BERT features for logistic regression
* More NN models (Basic VAE, XLNET, BERT)
  + Implement all <http://users.umiacs.umd.edu/~jbg/docs/2020_acl_diplomacy.pdf> report scores + improvement with meta
  + Fine tune bert (we are doing that)
* Add more features (bag of words, semantic features, bert, Roberta)
* Use table in <https://www.aclweb.org/anthology/2020.coling-main.511.pdf> (semi-supervised)

### 12 Dec 2020

1. Hyperparameter tuning – robustness + some curve
2. Quality as input: velocity & inter-annotator agreement

Quality matters for small dataset, doesn’t for big ones

\*velocity filter\*: everything one SD below average time to be a bot

And percentage agreement information

Weigh different things differently in your LSTMs

Either input at LSTM layer

Either at combination layer

Weight encoding – velocity + pc

Weight embedding one

When there is no tie breaker

Throughput

Different sigmoid functions

Other datasets: 2 other datasets – Kokil has other dataset

This idea works if you have a dataset of a certain size

1. Different models per annotation

Different models per annotation – show each annotation, which is the best consistently + convidence labels

Different aspects attention

Finally combined model attention to each specific annotation, because one particular LSTM architecture may not capture

LSTM might capture reasoning vs CNN captures shared information

Needs to show generalizability:

Difficulty: requires domain knowledge

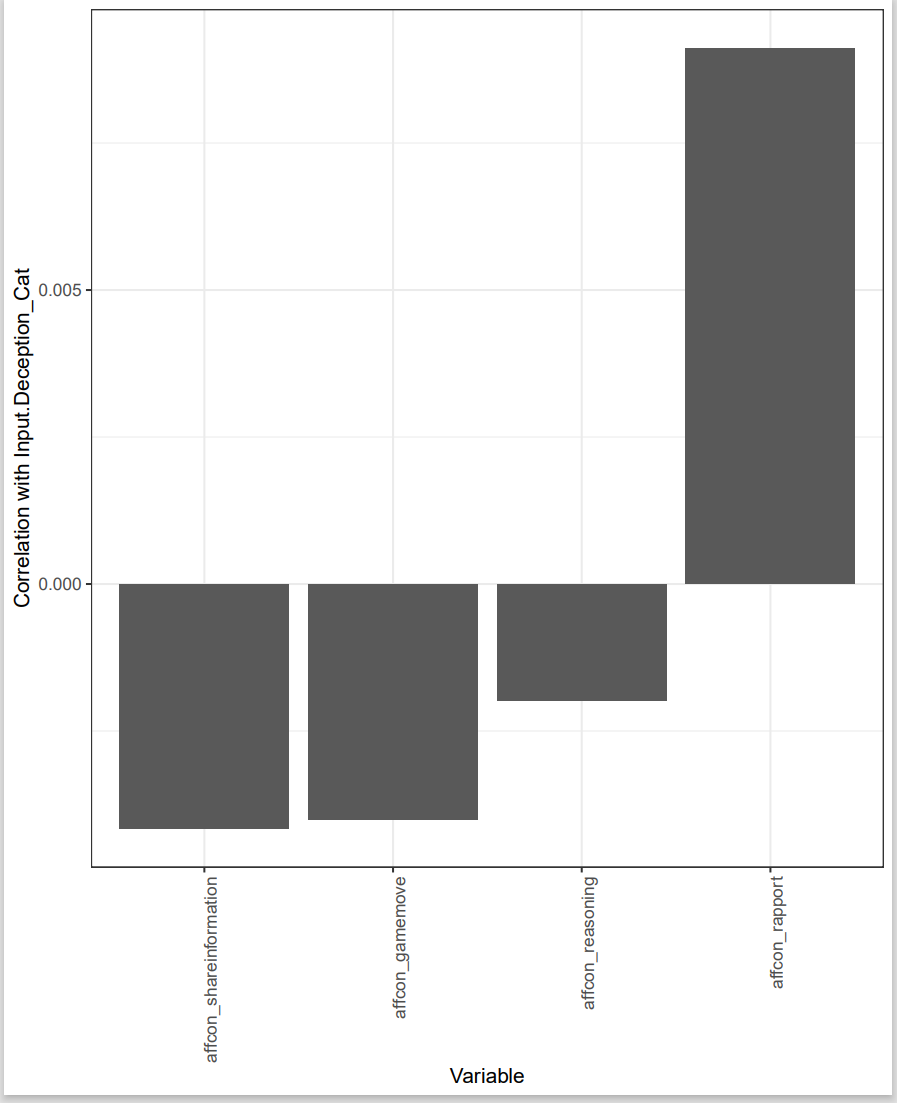
May not generalize unless we have another dataset – use betrayal dataset

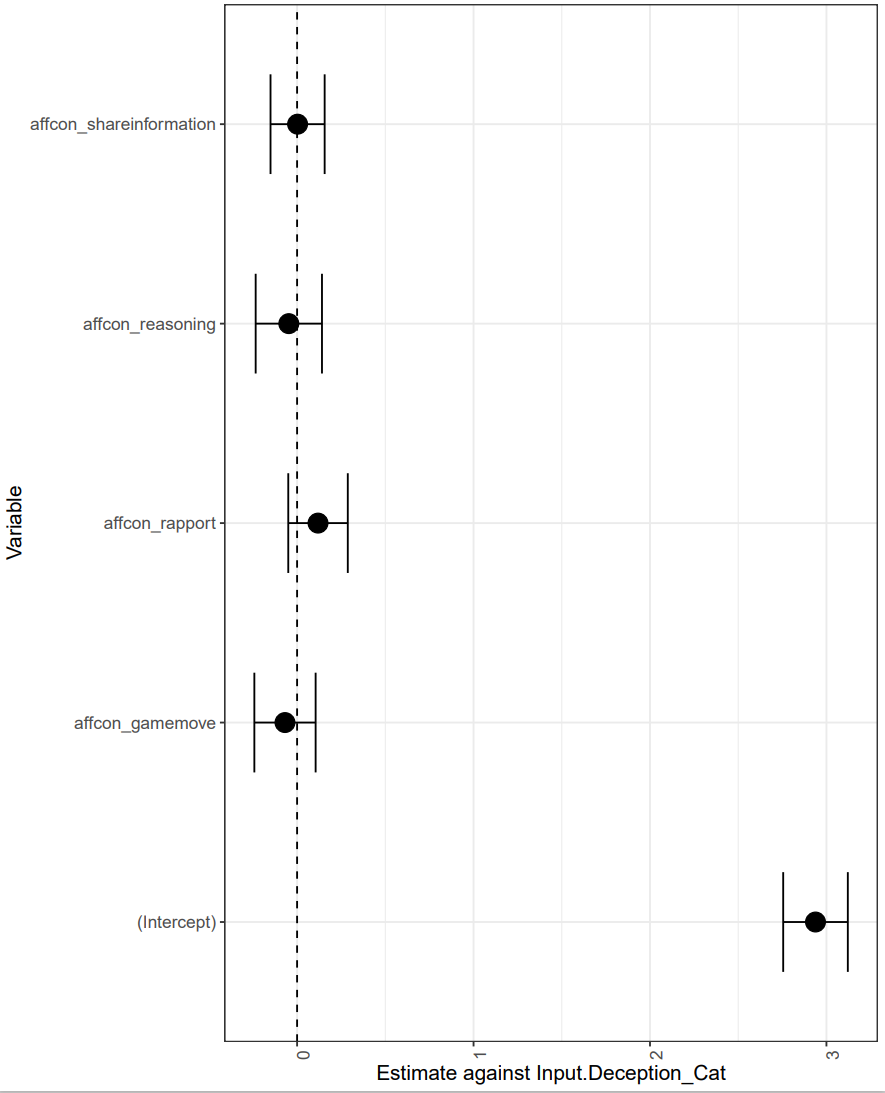
Any papers look at semantic/ discourse features/ part of speech information? => These models are able to give semantic differences better than other models?

Use pos-tag features for explainability

### 6 Dec 2020

1 is straightforward, 0 is cassandra





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)

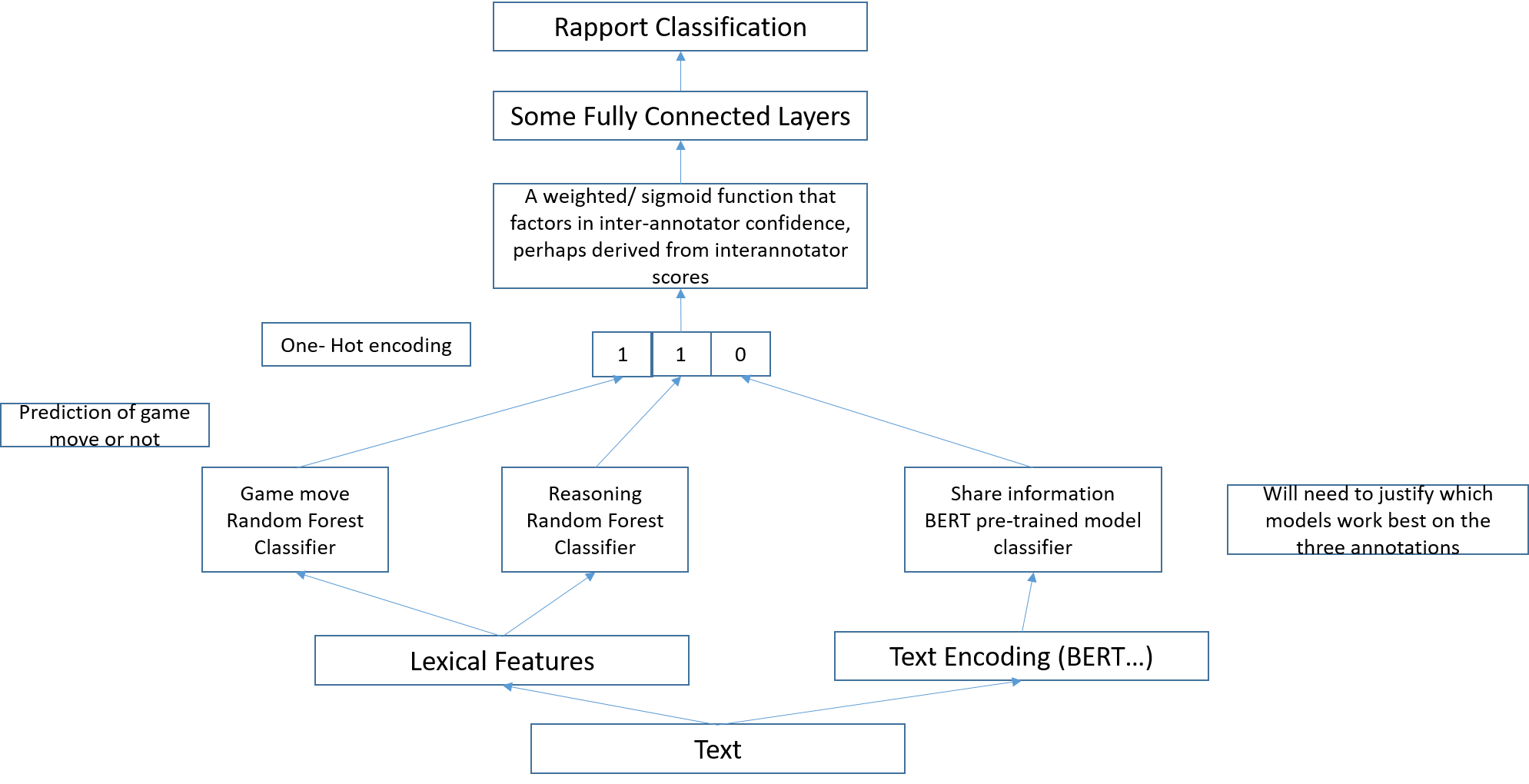
Let me know if you see any improvement to your models esp with the conf\_good data! I really appreciate how much time and effort yall are putting into this 😊

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

### 4 Dec 2020

Lynnette’s proposed structure

Not predicting Trust as I don’t know how to map Rapport to Trust yet…



### 4 Dec 2020

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.

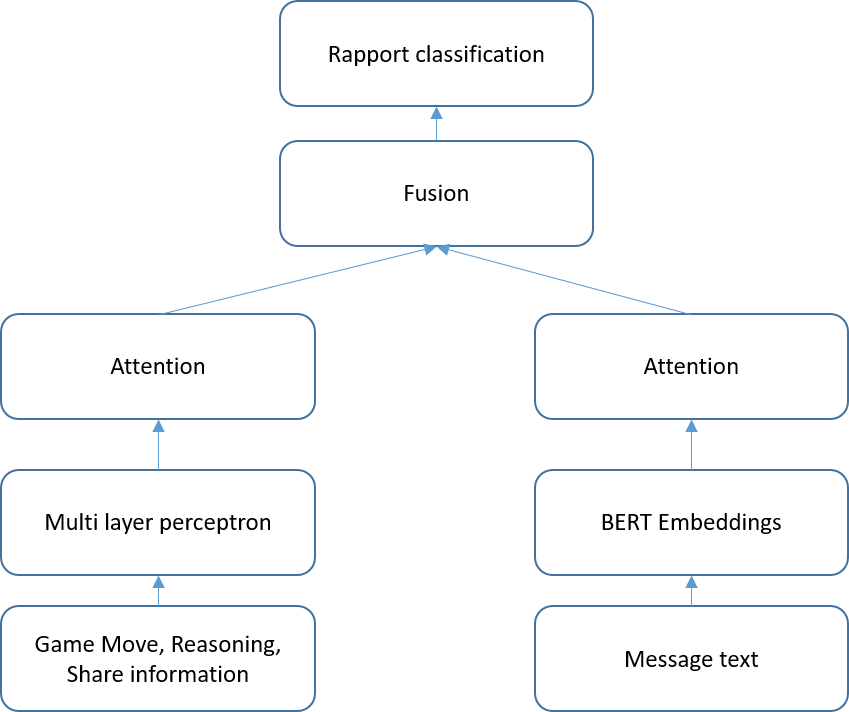
train on high-confidence annotations

previous work: summarization, recommendation of clickstream.

simple reward function: sigmoid/ whatever

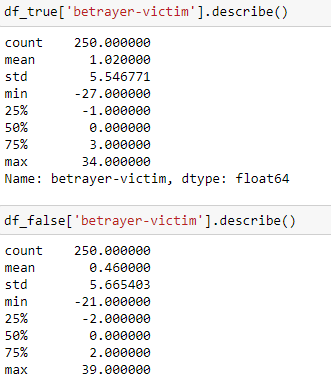
### 2 Dec 2020

What we can do…



Running a trained affcon\_rapport model on Betrayal dataset

Made each season a sequence, and annotated each sentence with rapport or not, then totaled up victim/ betrayer scores



### 1 Dec 2020

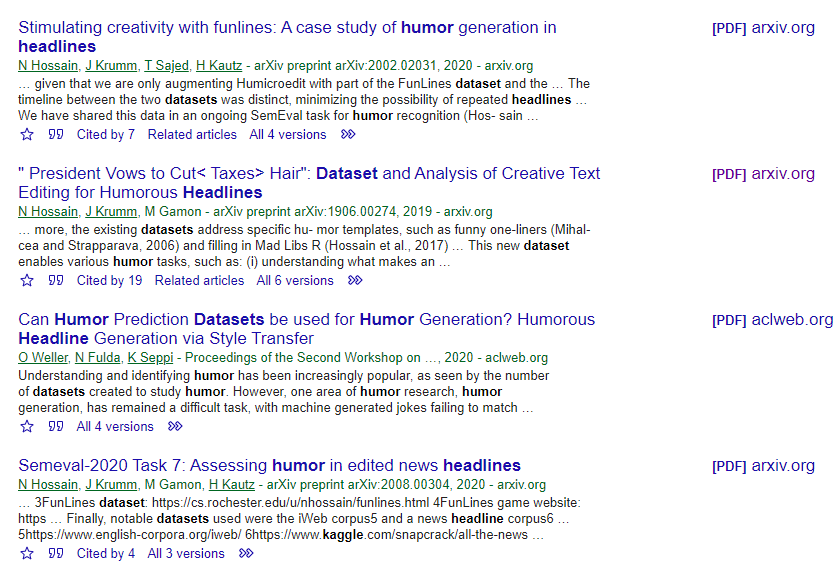
|  |  |
| --- | --- |
| **Affcon\_gamemove=True** | **Affcon\_gamemove=False** |
| Mostly looking to reposition  Move forward but we  Mostly just need Germany  Mostly worried that they  Mostly because concerned  Mostly just straight  Mostly you | Misclick earlier haven lied  Missing orders think he  Mind double check through  Might have an issue  Mere technicality gives  Might have helped  Might get  mistake |
| **Affcon\_reasoning=True** | **Affcon\_ reasoning =False** |
| Most natural allies on  Most likely outcome and  Most over est overdog  Most of the day  Most likely to  Most logical  Most natural | Moving out of the  Much interesting to say  Moving my fleet down  Much naval power to  My moves planned  Much better to  Much better for  Moving away |
| **Affcon\_shareinformation=True** | **Affcon\_shareinformation=True** |
| Means of gaining ground  Mean you could possibly  Mean you decide to  Mean no truce  Mean you and | Much prefer that we  Much prefer playing this  Much longer when  Much suspicion before  Much higher side  Much stronger |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Average |  |  |  |
| Affcon\_reasoning=True | 6.8131 | 4.5499 | 1.4327 |
| Affcon\_reasoning=False | 5.20708 | 4.1776 | 1.13518 |
|  |  |  |  |
| Per Text |  |  |  |
| Affcon\_reasoning=True | 5.4780  Std 5.7796 | 3.6071  Std 5.123 | 1.0990  Std 3.0834 |
| Affcon\_reasoning=False | 5.0539  Std 6.0794 | 4.7869  Std 6.1776 | 1.0017  Std 3.0986 |
|  |  |  |  |
| Average |  |  |  |
| Affcon\_shareinformation=True | 6.23964 | 4.25879 | 1.564909 |
| Affcon\_shareinformation=False | 6.177136 | 4.1756 | 1.43787 |
|  |  |  |  |
| Per Text |  |  |  |
| Affcon\_shareinformation=True | 4.8816  Std 5.4977 | 3.7039  Std 5.2521 | 1.3234  Std 3.3234 |
| Affcon\_shareinformation=False | 5.2632  Std 6.1739 | 4.0819  Std 5.5764 | 0.8072  Std 2.7831 |
|  |  |  |  |
| Average |  |  |  |
| Affcon\_gamemove=True | 5.3842 | 3.6566 | 1.1263 |
| Affcon\_ gamemove =False | 5.3979 | 3.6048 | 1.1351 |
|  |  |  |  |
| Per Text |  |  |  |
| Affcon\_ gamemove =True | 5.3954  Std 5.7895 | 3.9349  Std 5.3361 | 1.0886  Std 3.0706 |
| Affcon\_ gamemove =False | 5.3339  Std 6.0171 | 3.7571  Std 5.6038 | 1.0441  Std 3.1292 |

### 29 November 2020

Action items:

1. Lynnette: Target/priority is to find a paper whose structure we can rip
   1. Proceedings paper
      1. We basically actually have only six weeks to get this done.
      2. Dataset paper would be nice, we can get away with less method novelty. Example dataset paper from NAACL: <https://arxiv.org/pdf/1906.00274.pdf> but they have written other follow up papers I think (screenshot)



* + 1. Still, they like insights, so look at the message attention paper below (<http://wwbp.org/papers/acl-2020.pdf>)

1. Lynnette: Second priority is to find the novelty we want to push
   1. The ideas below
2. Lower priority:
   1. Lynnette: Cites of the datasets and the trust papers [done]
   2. Lynnette: Prepare the full dataset for annotation.
   3. Lynnette: to explore betrayal dataset
3. Kokil: tutorial dataset

Novelty directions

 We can use social and agency labels as intermediates in the latent space

1. Hierarchical model : <http://wwbp.org/papers/acl-2020.pdf>
2. To help you with the implementation, have a look at this -  [https://www.aclweb.org/anthology/N18-1092.pdf](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.google.com%2Furl%3Fq%3Dhttps%3A%2F%2Fwww.aclweb.org%2Fanthology%2FN18-1092.pdf%26sa%3DD%26source%3Dhangouts%26ust%3D1587101375340000%26usg%3DAFQjCNHm0gRvC0hROAwYCG4U9CN1qyMzEw&data=02|01|nchhaya@adobe.com|148787f4091942ca714908d7f2a0d8d0|fa7b1b5a7b34438794aed2c178decee1|0|1|637244644983523941&sdata=6zPRKlKhzVMRX9ZI98k7W5IIXclNSLlErDmDNwlyJgo%3D&reserved=0). Here is the code - [https://github.com/Asa-Nisi-Masa/Embed-Align-NLP](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fgithub.com%2FAsa-Nisi-Masa%2FEmbed-Align-NLP&data=02|01|nchhaya@adobe.com|148787f4091942ca714908d7f2a0d8d0|fa7b1b5a7b34438794aed2c178decee1|0|0|637244644983533932&sdata=qGSZcbrwCUGizxG7cEU2PqA%2BSViX9tegvTeXvz8wkEc%3D&reserved=0)
3. If we are looking at introducing other label information into Bert, first it may be a good idea to append the labels to the Bert embeddings and see if the prediction improves.
4. Latent universal task specific bert - [https://arxiv.org/pdf/1905.06638.pdf](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Farxiv.org%2Fpdf%2F1905.06638.pdf&data=02|01|nchhaya@adobe.com|148787f4091942ca714908d7f2a0d8d0|fa7b1b5a7b34438794aed2c178decee1|0|1|637244644983503949&sdata=HpEw%2BxG9WOvha2bpwBwjiKsDKZHxx45z45U43mD%2F8M8%3D&reserved=0)
5. Few shot learning – latent space - [https://arxiv.org/pdf/1910.04176.pdf](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Farxiv.org%2Fpdf%2F1910.04176.pdf&data=02|01|nchhaya@adobe.com|148787f4091942ca714908d7f2a0d8d0|fa7b1b5a7b34438794aed2c178decee1|0|1|637244644983503949&sdata=eKOT3wjzBBLF%2F5MpBinCoL6E8SwtMsiSxCPj7%2F7X%2FH8%3D&reserved=0)
6. Projected attention layers - [https://arxiv.org/pdf/1902.02671.pdf](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Farxiv.org%2Fpdf%2F1902.02671.pdf&data=02|01|nchhaya@adobe.com|148787f4091942ca714908d7f2a0d8d0|fa7b1b5a7b34438794aed2c178decee1|0|1|637244644983513941&sdata=jtsMGuKpNmkA0QkOKQp47ARs11cyFZbhILLVw%2BM8W7M%3D&reserved=0) - here, they create a multi-task setup that enables the language model to learn the adjacent space better.
7. Multiview data - [https://arxiv.org/pdf/2002.12573.pdf](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Farxiv.org%2Fpdf%2F2002.12573.pdf&data=02|01|nchhaya@adobe.com|148787f4091942ca714908d7f2a0d8d0|fa7b1b5a7b34438794aed2c178decee1|0|1|637244644983513941&sdata=lVtELRZKLUwnrVVOxr0fxOghyEQWV%2BWeCaVrgRR%2Fiw0%3D&reserved=0) – this one is around video data but it may be good to explore this concept for our implementation.
8. The paper presents a method for cross-lingual classification across a common subspace e that may be learnt – can we map it to our setup where instead of language being the differentiating aspect, we have the agency/ social tags in there. – [https://arxiv.org/pdf/1206.6481.pdf](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Farxiv.org%2Fpdf%2F1206.6481.pdf&data=02|01|nchhaya@adobe.com|148787f4091942ca714908d7f2a0d8d0|fa7b1b5a7b34438794aed2c178decee1|0|1|637244644983523941&sdata=qcpHrf9AGha%2BhmMPp4pG9guNuvWDA3ZxvRMQxMipICU%3D&reserved=0)

* It may be a good idea to check 1 & 2 quickly and see if their code is available – scan through it may be
* 3 will be most straight forward to implement in my opinion – let me know if need more references or guidance there.
* 4 and 5 are where, we may need to think about how to map between their problem setup and ours – you can scan through the paper(s) if you find them interesting, then we can dig deeper.
* For a quick implementation, can you check and see if any of 1, 2, or 3 are something you may be comfortable with ?

### 17 November 2020

Potential papers and ideas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Idea** | **Data required** | **Experiments** | **Potential Venues** | **Deadline** |
| Building up a successful lie  [Keep for future, but could annotate all sentences] | [Requires more annotations]  Annotate affcon\_rapport for all data  And the sub- rapport items | Figure out the temporal/ language/ affcon\_ sequence that results in the first Lie/Truth, hence where receiver gets deceived  Naïve approach:  Build a text classifier on past texts of a game/ person against Deceived  Not so naïve approach:  Build a rapport ranking score for each person in each game against Deceived | CSCW  CHI |  |
| A model for Trust | Current set of annotations  Kokil suggests that we collect a second dataset.  (eg the ones on Reddit selling stuff) | Understanding The Trust Equation | Trusted Advisor  Some widely used leadership Trust equation  Build a trust matrix for each person/ game  **Credibility** – learn word vectors from a credibility dataset  Eg Twitter disinformation/truth/lies (<https://www.data.gouv.fr/en/datasets/credibility-corpus-with-several-datasets-twitter-web-database-in-french-and-english/#resource-7f14efd9-0016-4b45-9ff5-519f70e4dcf9>)  Because Twitter is also short texts/ conversational  <https://github.com/pushpendra-1104/Credibility_Detection>  **Reliability** – number of times receiver thinks it is truth vs number of times it is truth  Intimacy/rapport – sum up rapport scores for all sentences in a message, and take the mode. If >1 then the message attempts to build rapport  **Self-orientation** – difference in game score because ultimately it’s a single winner game  **Validation**: If Trust Score > threshold, on the next iteration, the receiver should think the speaker is telling the truth  Further than this: generative lies  [https://journals.plos.org/plosone/article?id=10.1371/journals.pone.0115844] | Ultra cool venue for ultra cool idea |  |
| A dataset paper | Current set of annotations where speaker tells the truth  [Can consider using new annotation set]  Reasoning  Game move  Rapport  Share Information | Build language models of each annotation set  Run LIWC/ NRC VAD and build lexical models  Build concatenated language model + numerical model  Identify frequent words, how words differ in each set  Enrich features with diplomacy unlabeled training data – Learn what is Betrayal?  Use other labels to create a subspace/ weigh different features differently | ACL 2021  Dec 1-11 – Find novelty  Dec 12 – 30 – Experiments  Jan – Write up  Feb – Paper writing | 25 Jan – Abstracts due  Feb 2021 |
| Ethos Pathos Logos  **Kokil’s idea** | Annotations where speaker tells the truth  Reasoning – Logos  Ethos – Game move  Pathos – Rapport | ? | Some communications journal |  |

### 15 November 2020

1. Kokil to create clean data file (TODAY)
2. Formalize the papers/research problem per paper. Lynnette to identify the first one she wants to lead
   1. Kokil to create an excel file for that, put into the same folder (TODAY)
   2. For the first one, Lynnette to find the conference, the deadline, and the research design
3. Introduce Lynnette to annotators, give them the codebook examples to start with
4. Follow up and ensure Lynnette/all others are onboard Sesami. Lynnette payment S$900.
5. Maybe after 24th Nov: Lynnette to give Kokil a tutorial about setting up GPU stuff on her machine

Lynnette’s timeline:

* December is a good time to work
* Flying end of Jan
* Feb should be pretty busy

### 14 November 2020

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 |

Brilliant idea from past work to do data augmentation… use translation

<https://github.com/GT-SALT/AAAI_CLF>

<https://arxiv.org/pdf/2004.10972.pdf>

**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** (calculated as a whole)

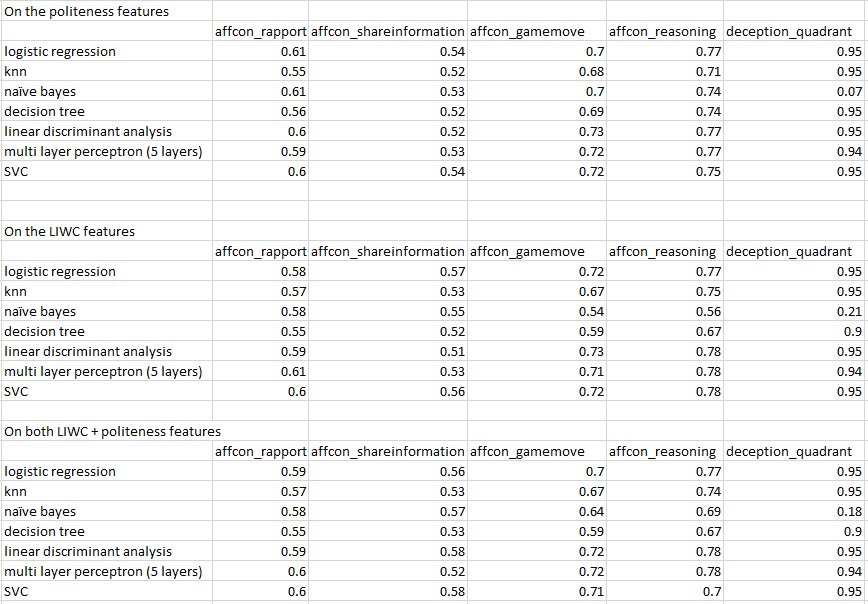
|  |  |  |  |
| --- | --- | --- | --- |
|  | First person % | Second Person % | Third Person % |
| Affcon\_rapport = True | 39.3847 | 33.25816 | 10.24625 |
| Affcon\_rapport = False | 41.5769 | 34.1641 | 10.598 |
|  |  |  |  |
| Per text |  |  |  |
| Affcon\_rapport=True | 5.7562  Std 5.9032 | 3.7712  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 |

### 8 November 2020

Generated Politeness features using Cornell’s ConvoKit & LIWC Features.

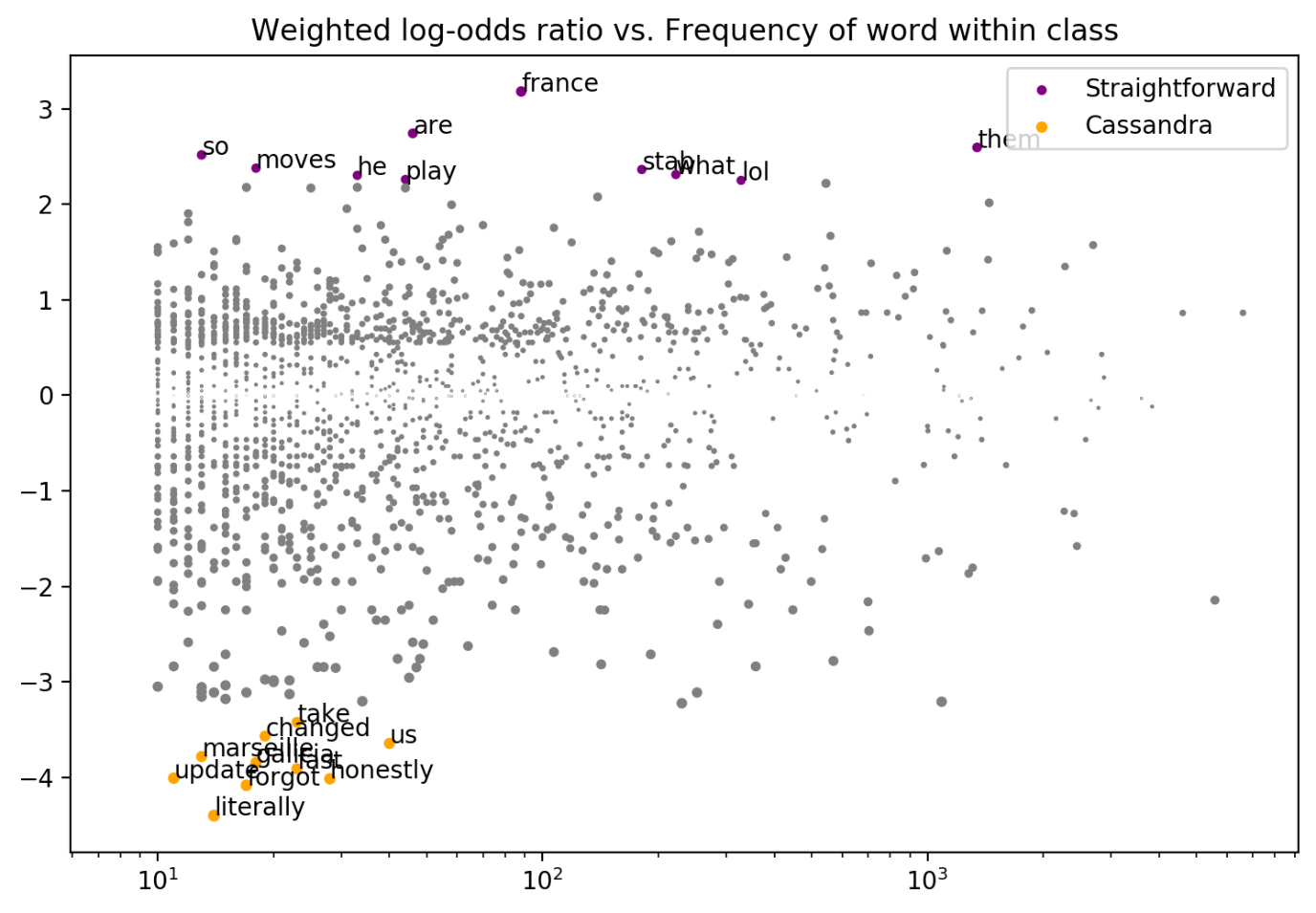
Performed some regression analysis against the annotated dependent variables. Looks like combining both LIWC + politeness features does as well as only the LIWC features. The Multi Layer Perceptron does the best, suggesting we do need a neural network solution. However, the other values I generated could be used as baseline numbers.

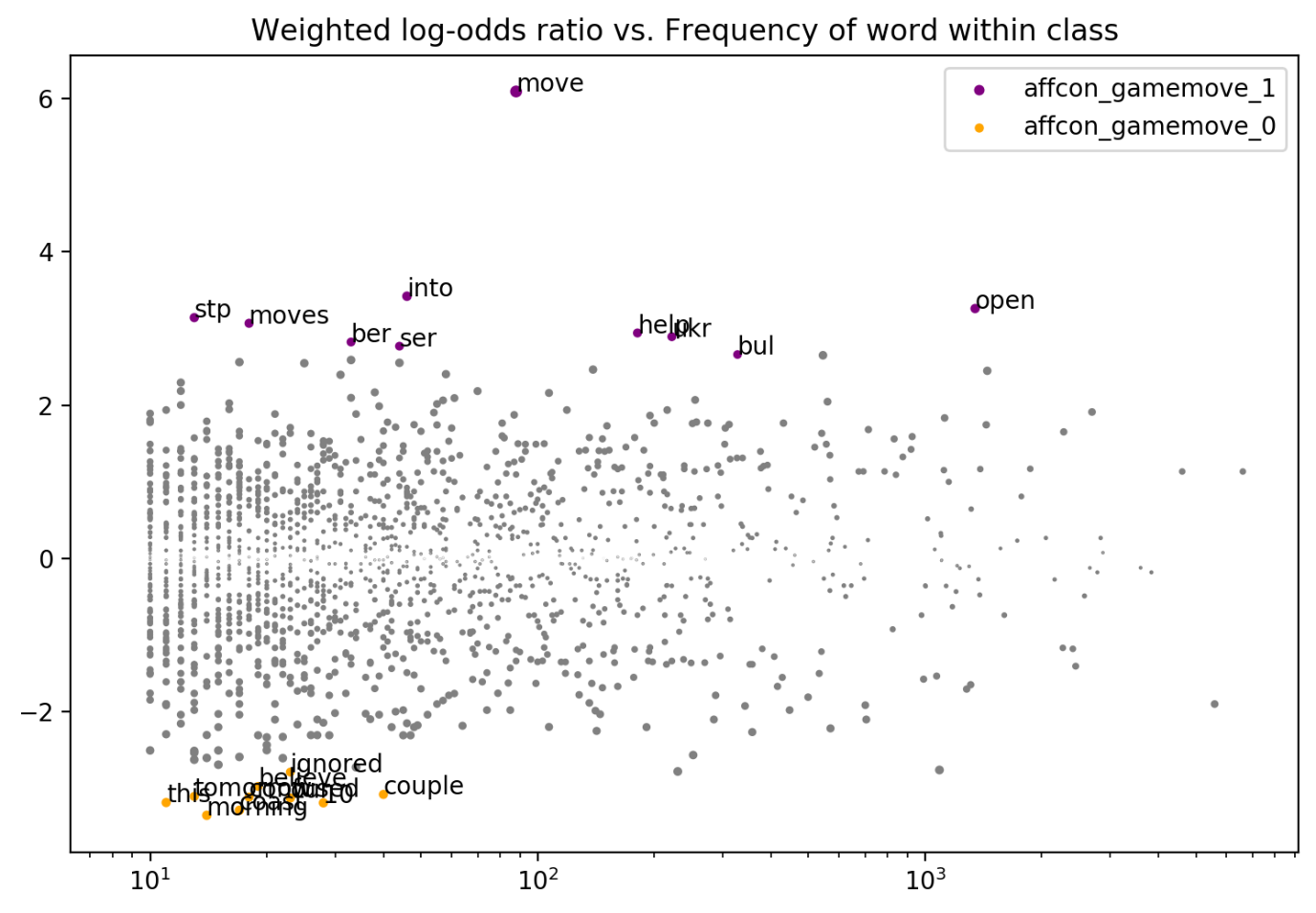
For reference, the table is as below.

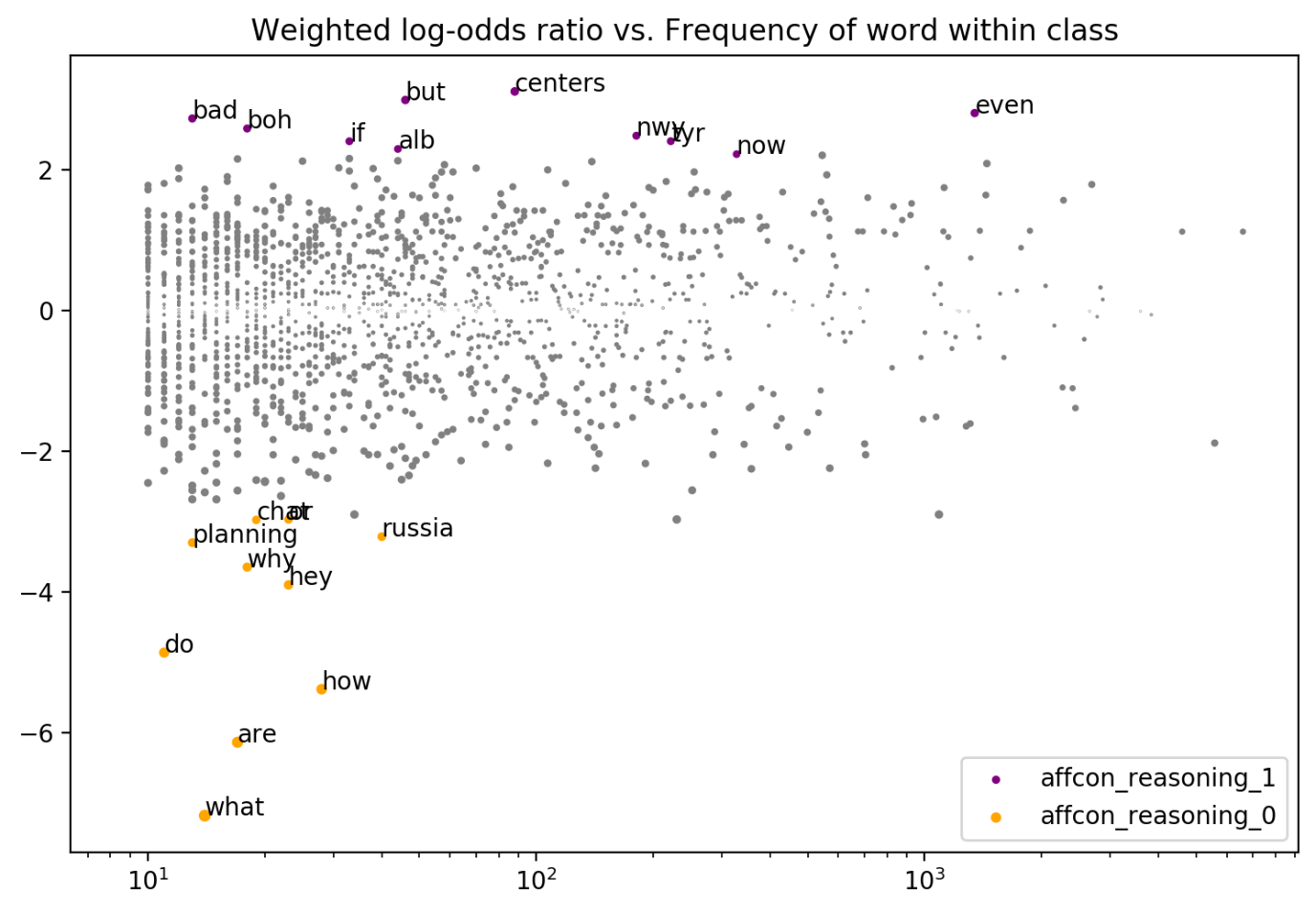


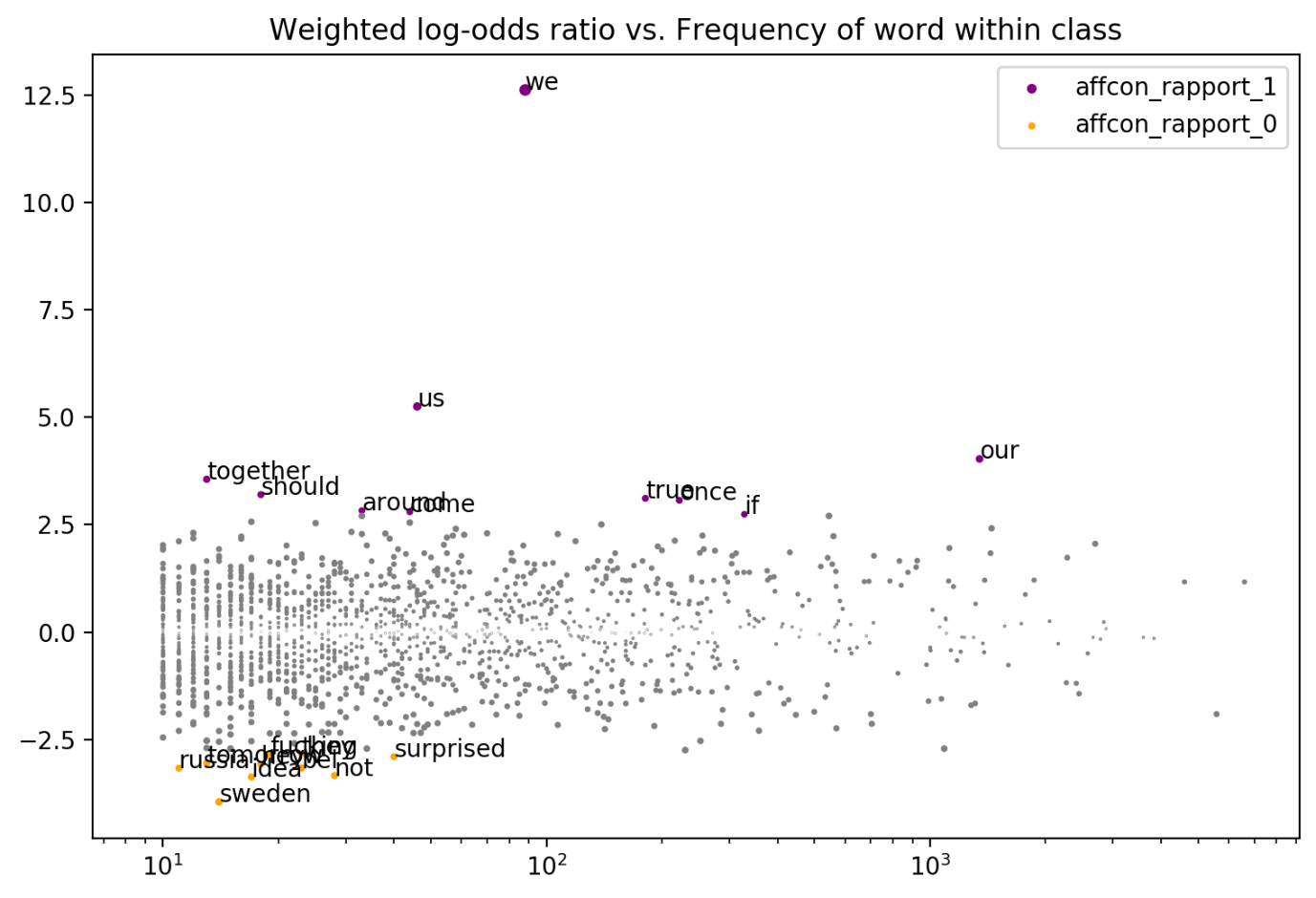
### 5 November 2020

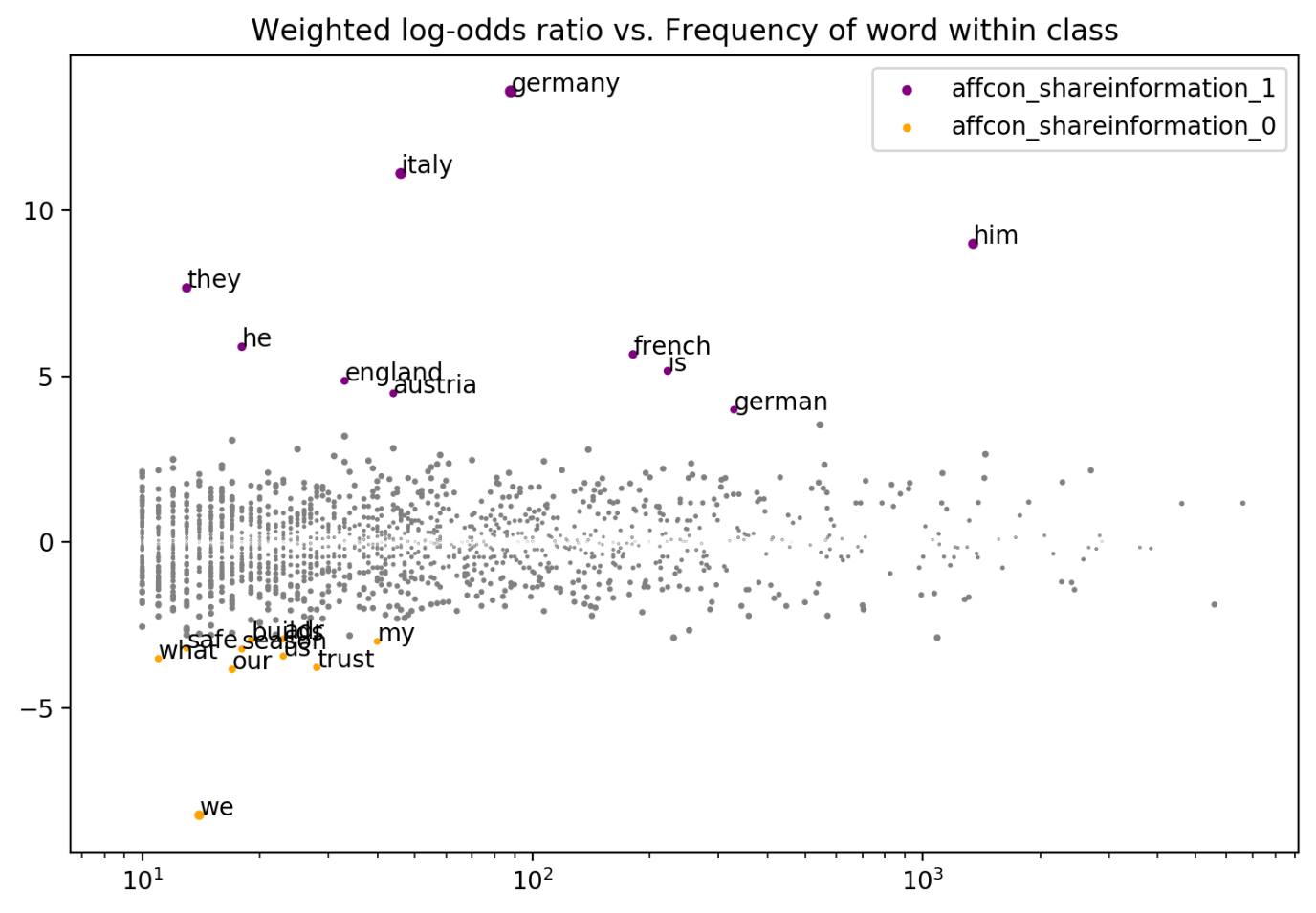
**Fighting Words from Cornell Convokit (**[**https://github.com/CornellNLP/Cornell-Conversational-Analysis-Toolkit/blob/master/convokit/fighting\_words/demos/fightingwords\_demo.ipynb**](https://github.com/CornellNLP/Cornell-Conversational-Analysis-Toolkit/blob/master/convokit/fighting_words/demos/fightingwords_demo.ipynb)**)**











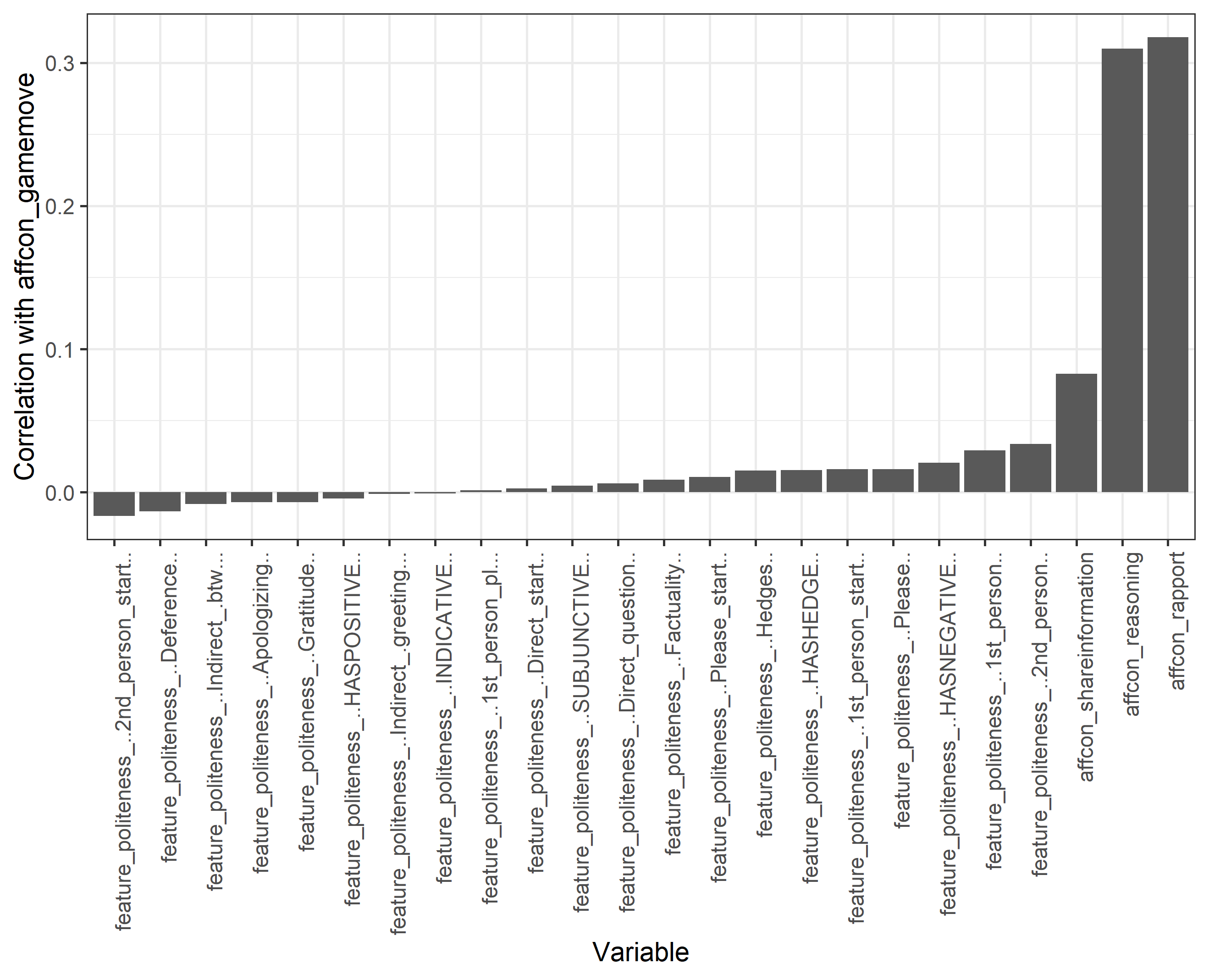
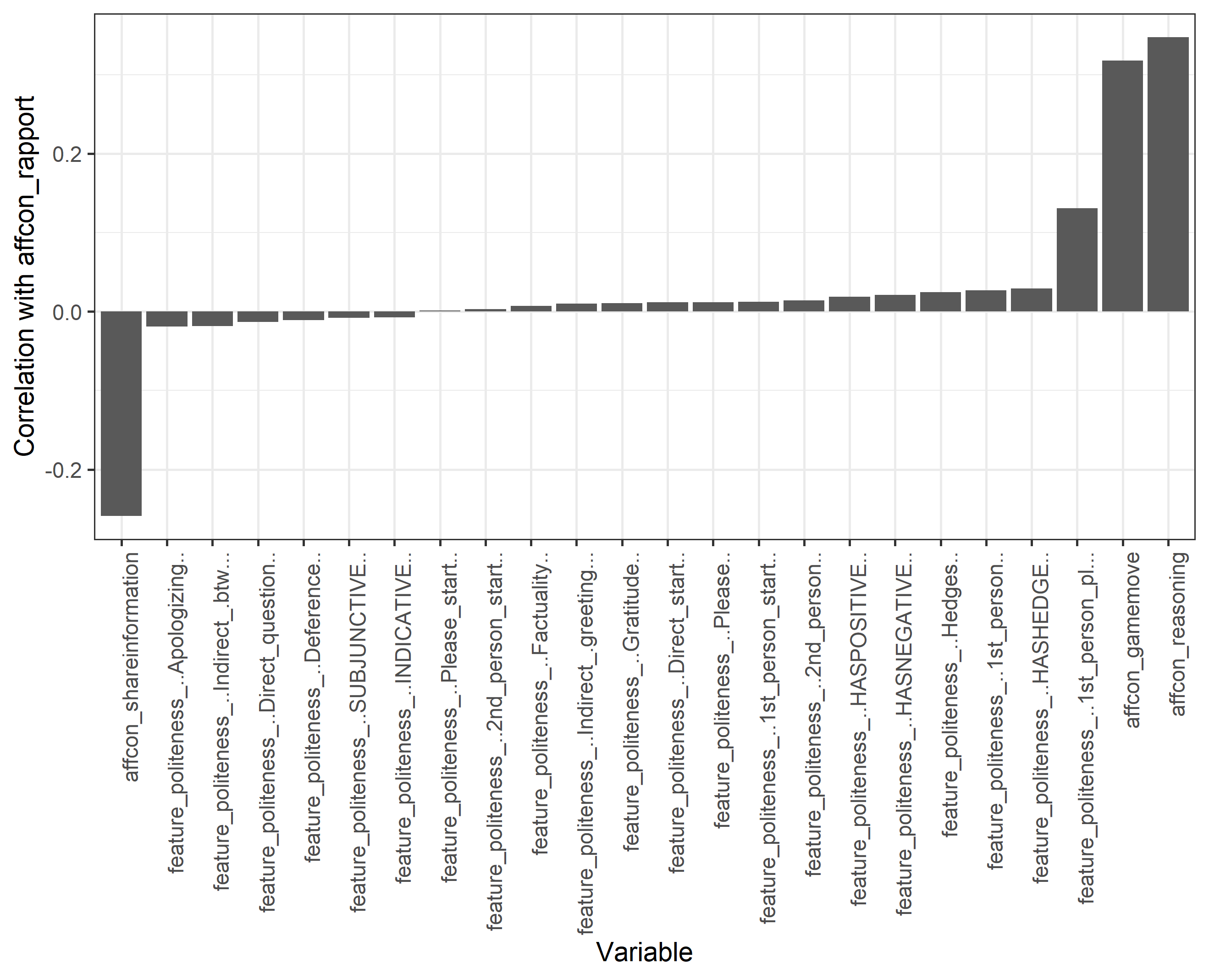
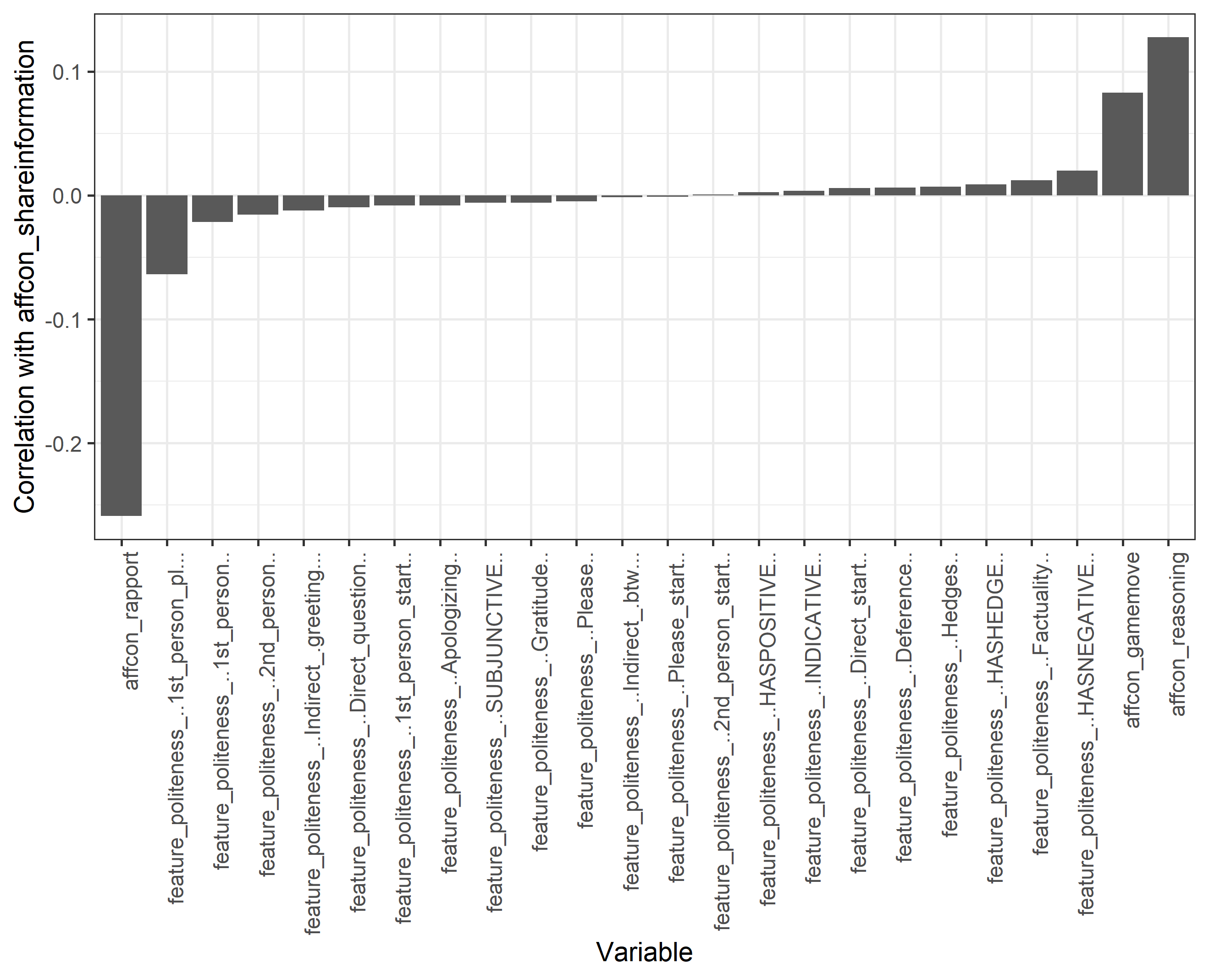
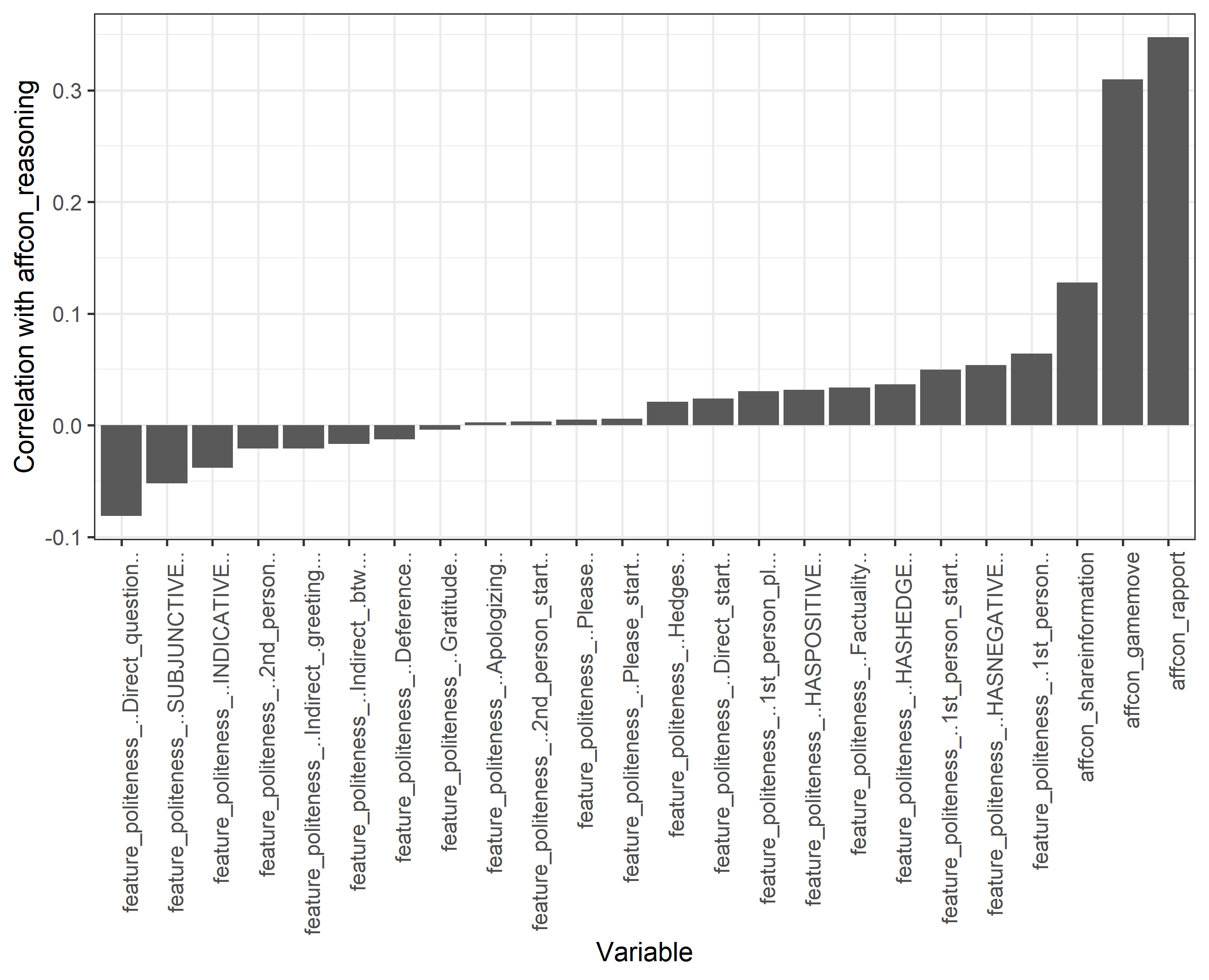
**Convokit Bag of Words Classifier**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| Affcon\_rapport | 0.69 | 0.70 | 0.69 |
| Affcon\_reasoning | 0.76 | 0.79 | 0.75 |
| Affcon\_gamemove | 0.73 | 0.75 | 0.72 |
| Affcon\_shareinformation | 0.69 | 0.69 | 0.69 |

**Politeness Strategies against Rapport**

Politeness strategies generated with Convokit and correlation plots

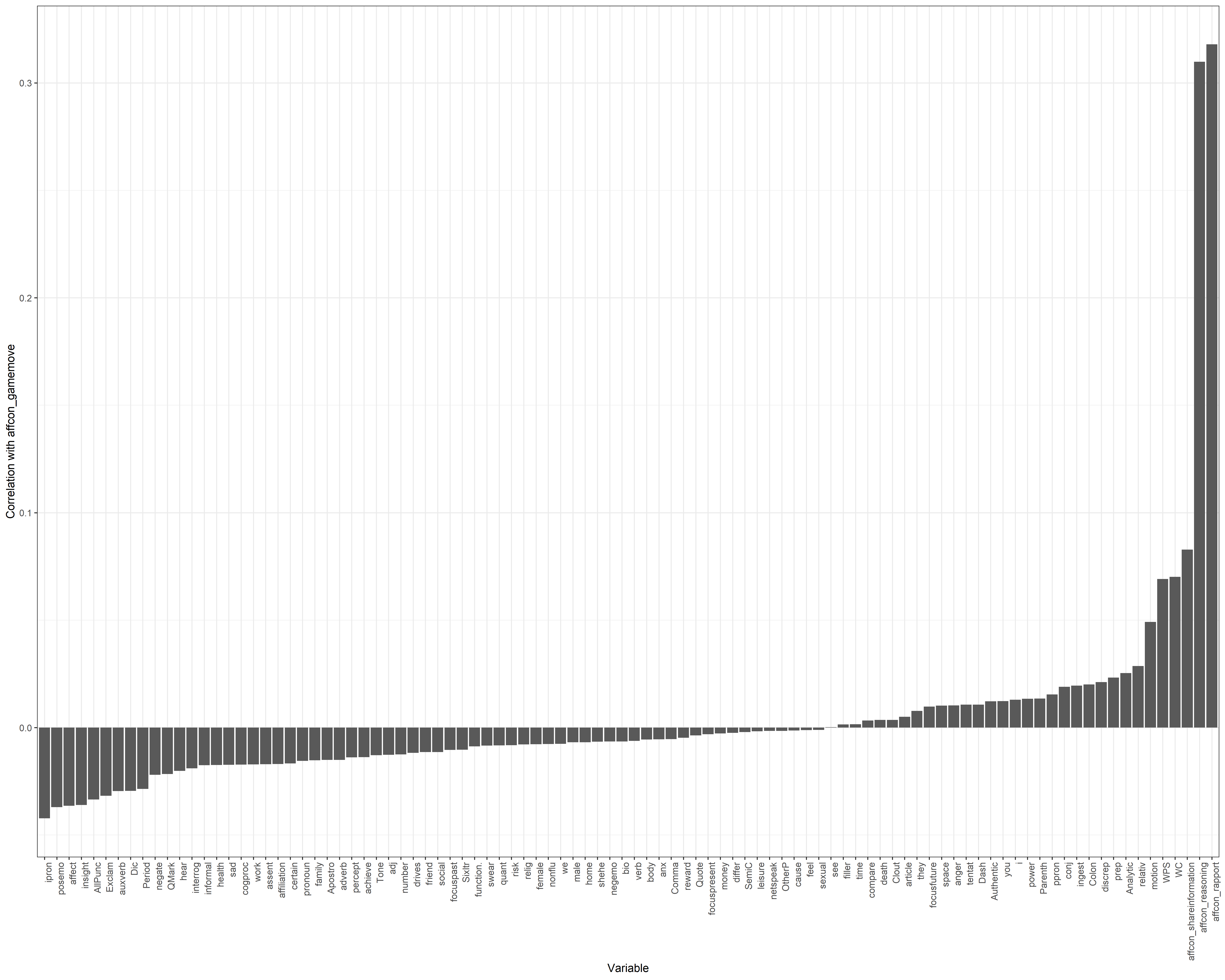
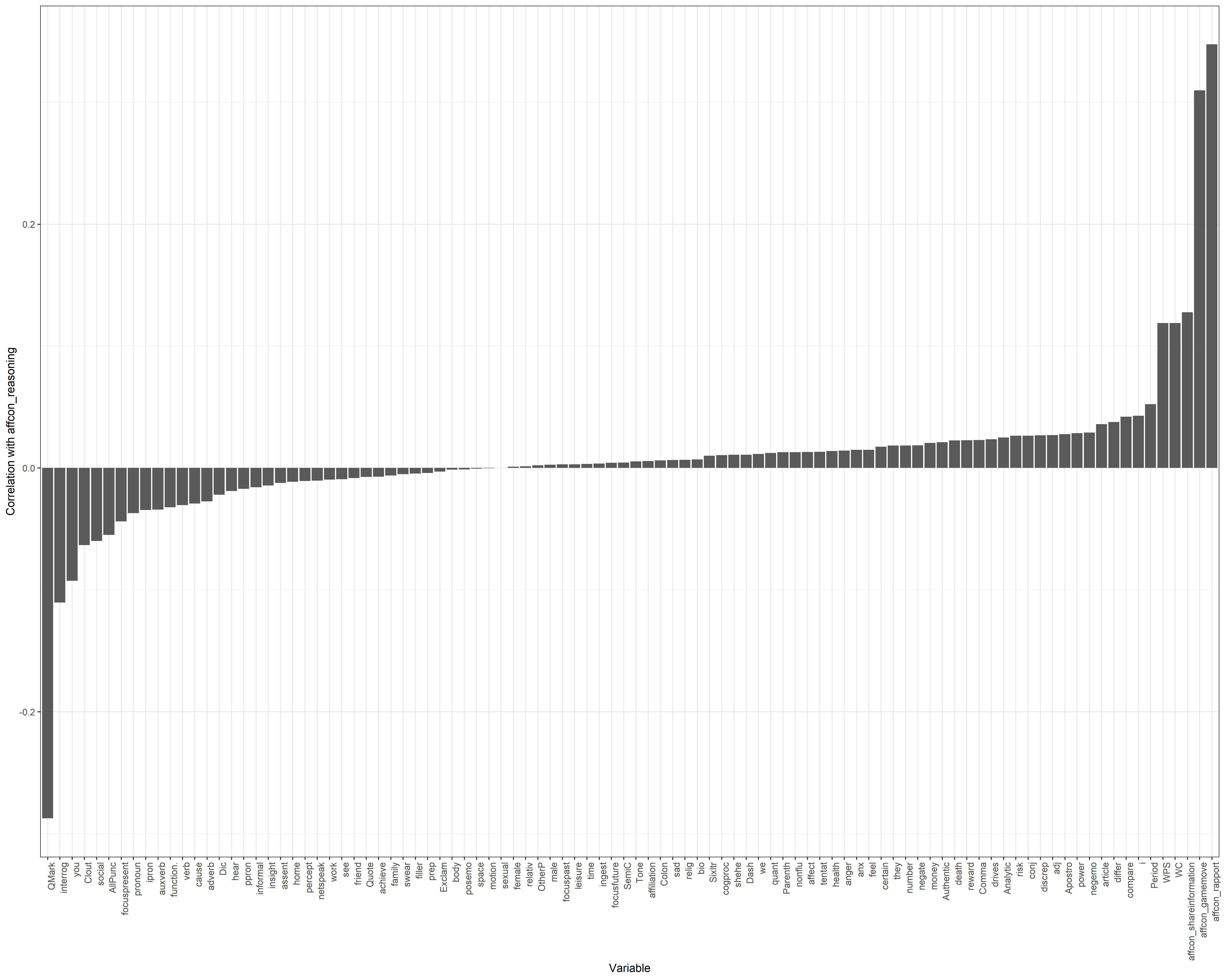
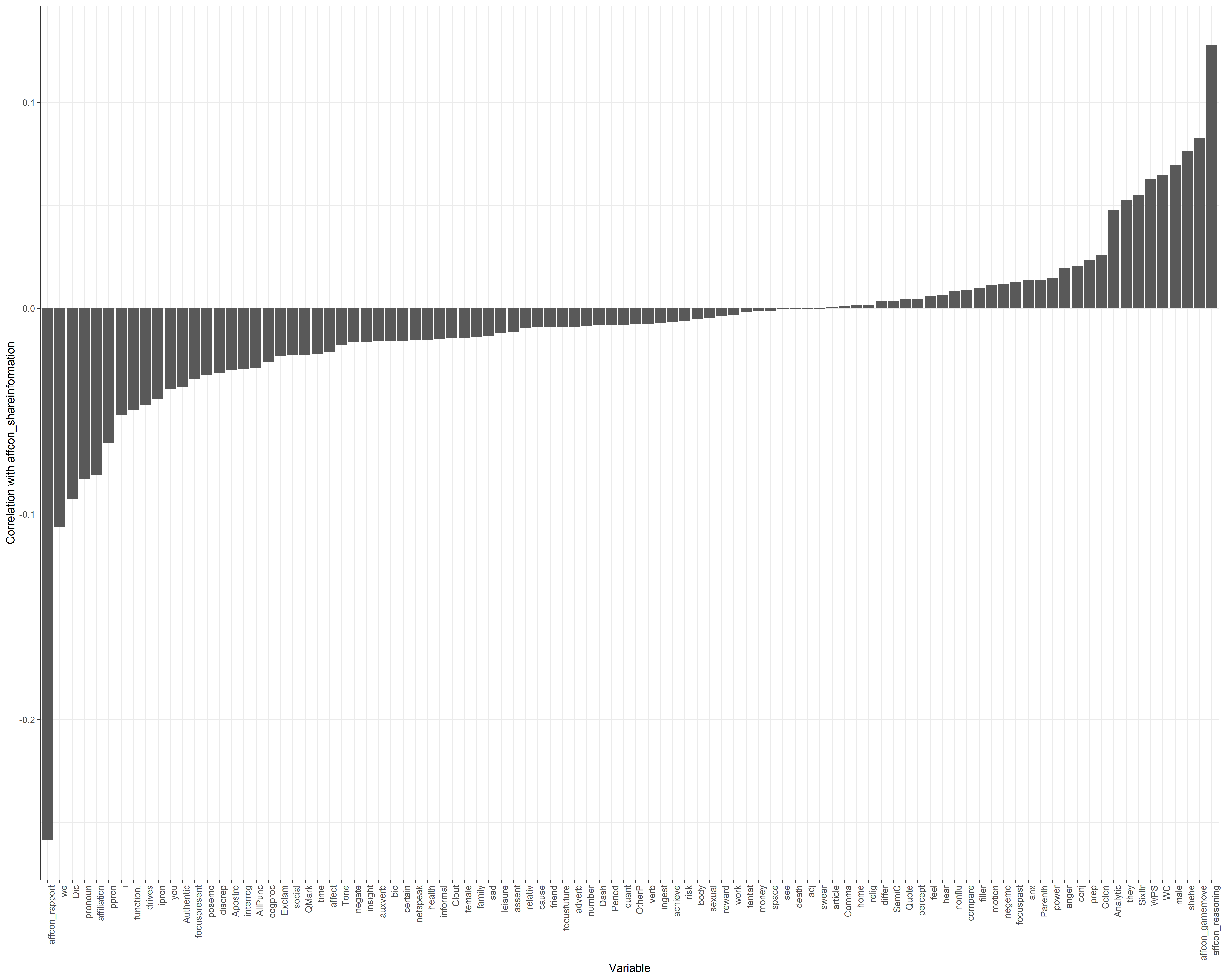
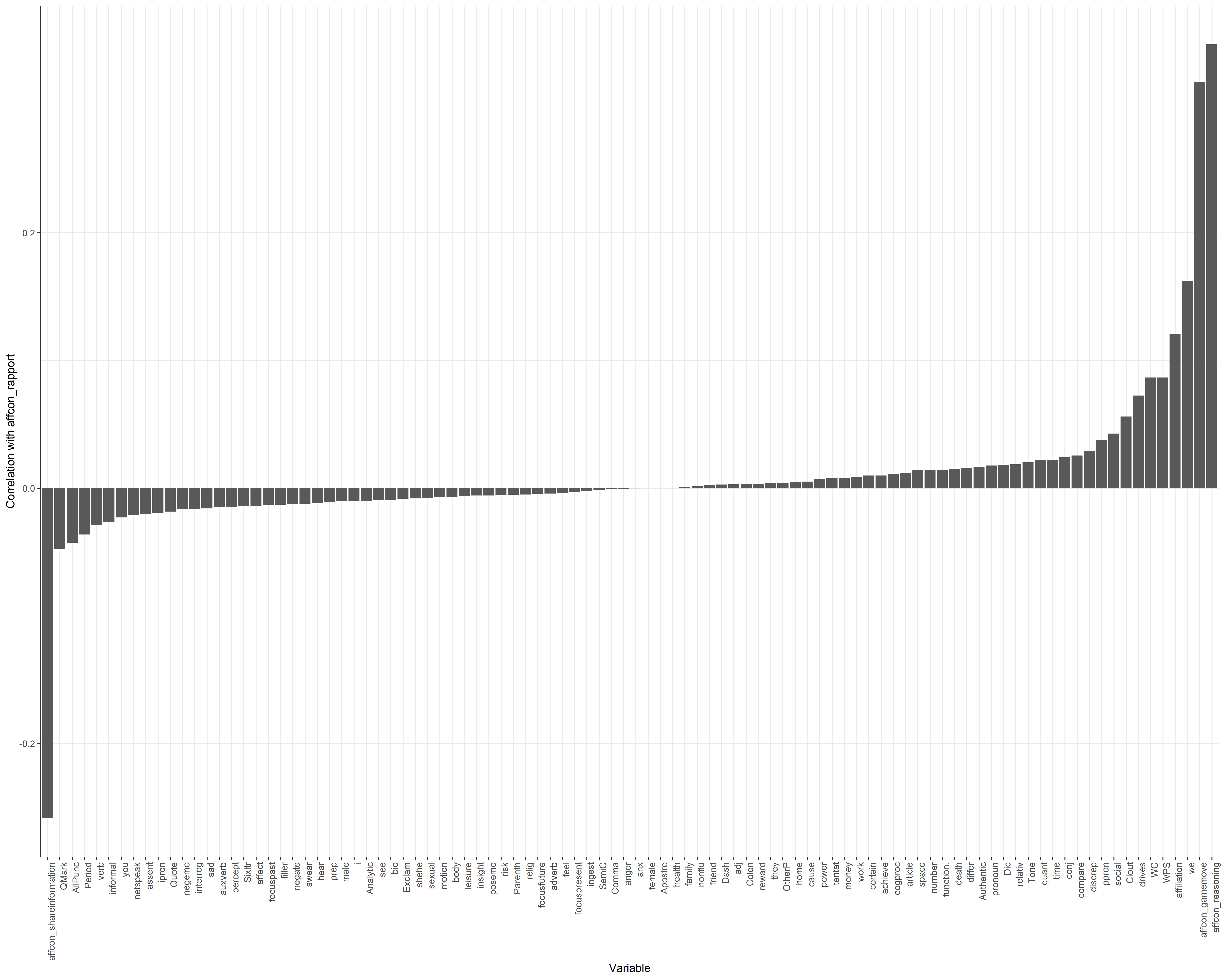




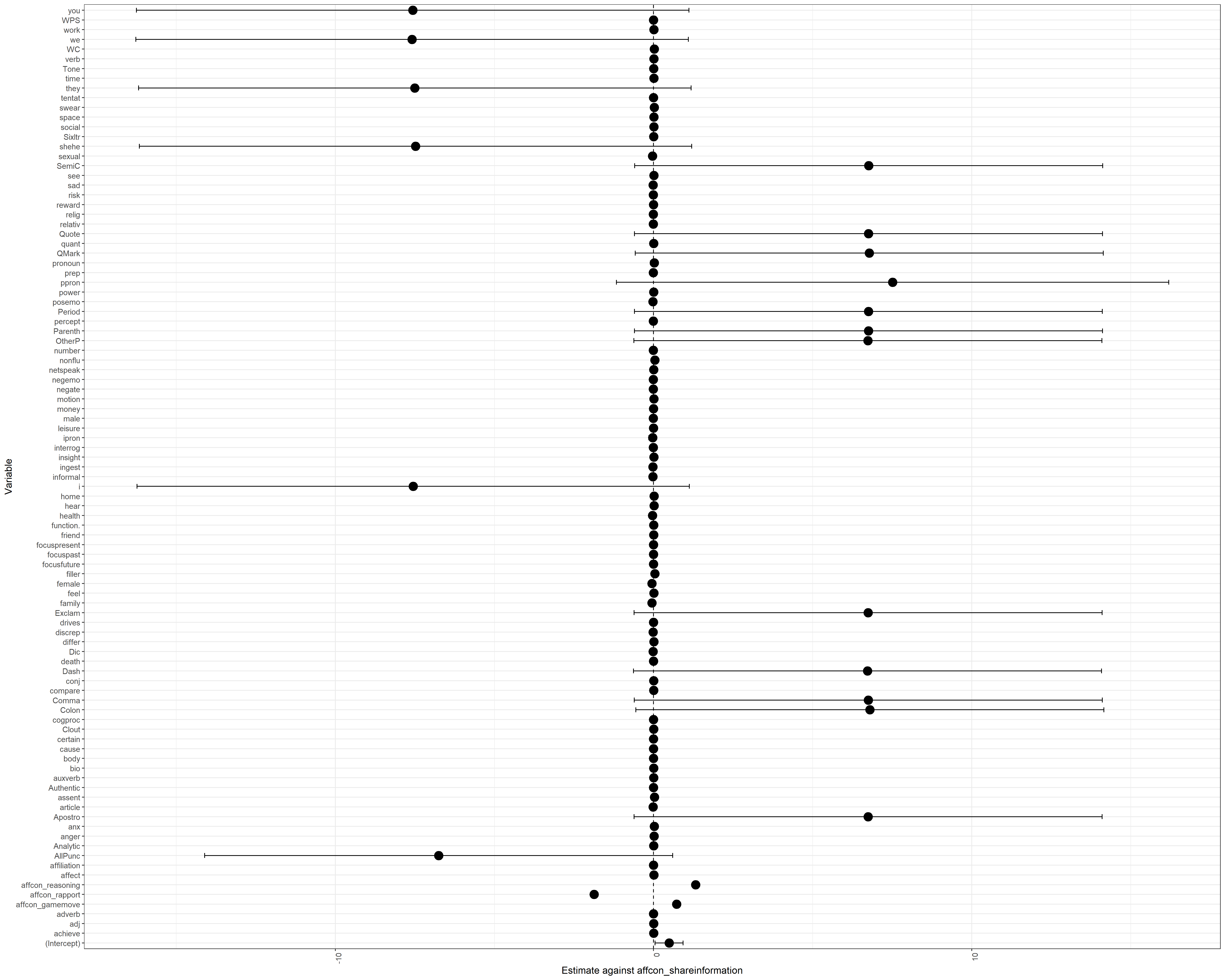
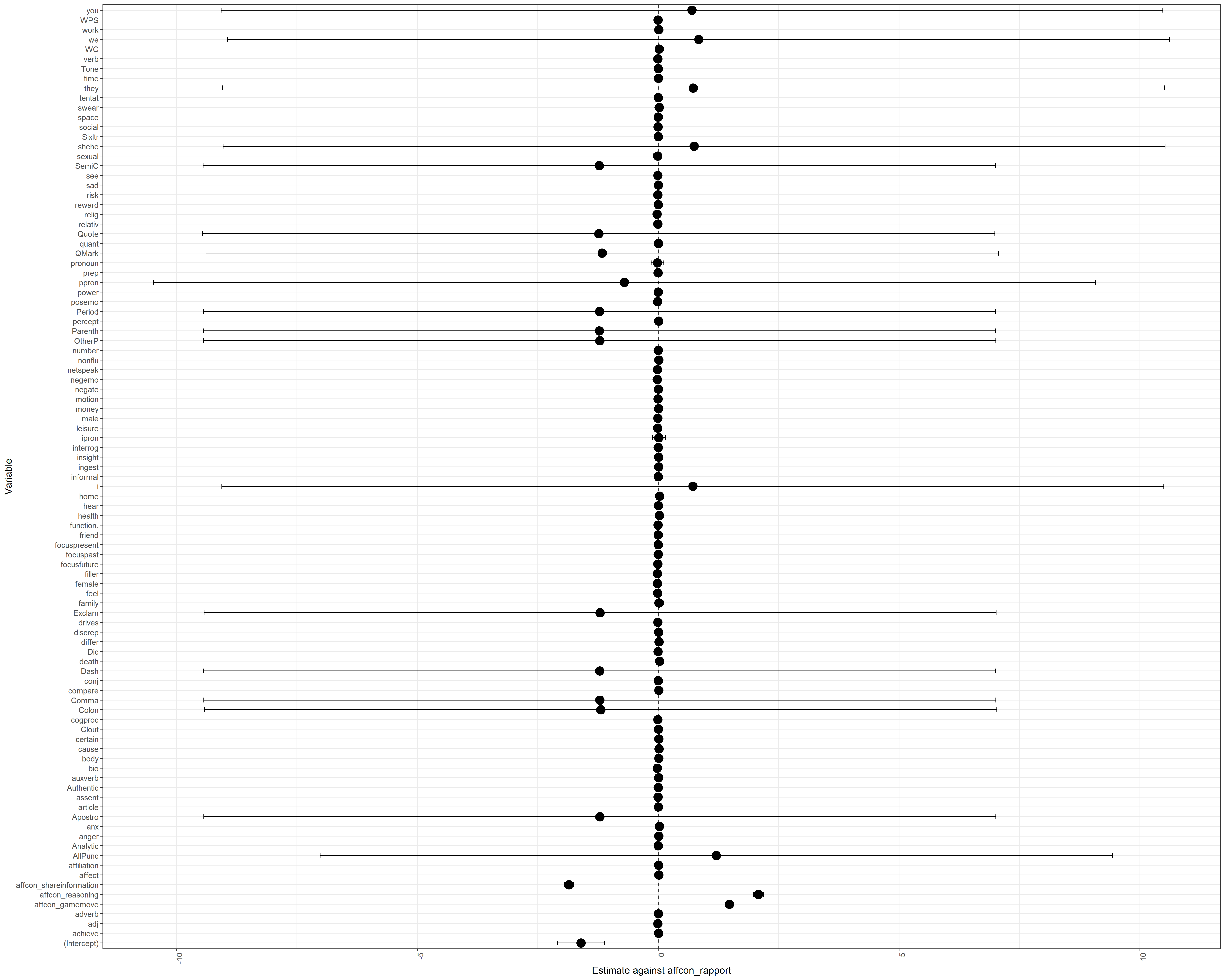
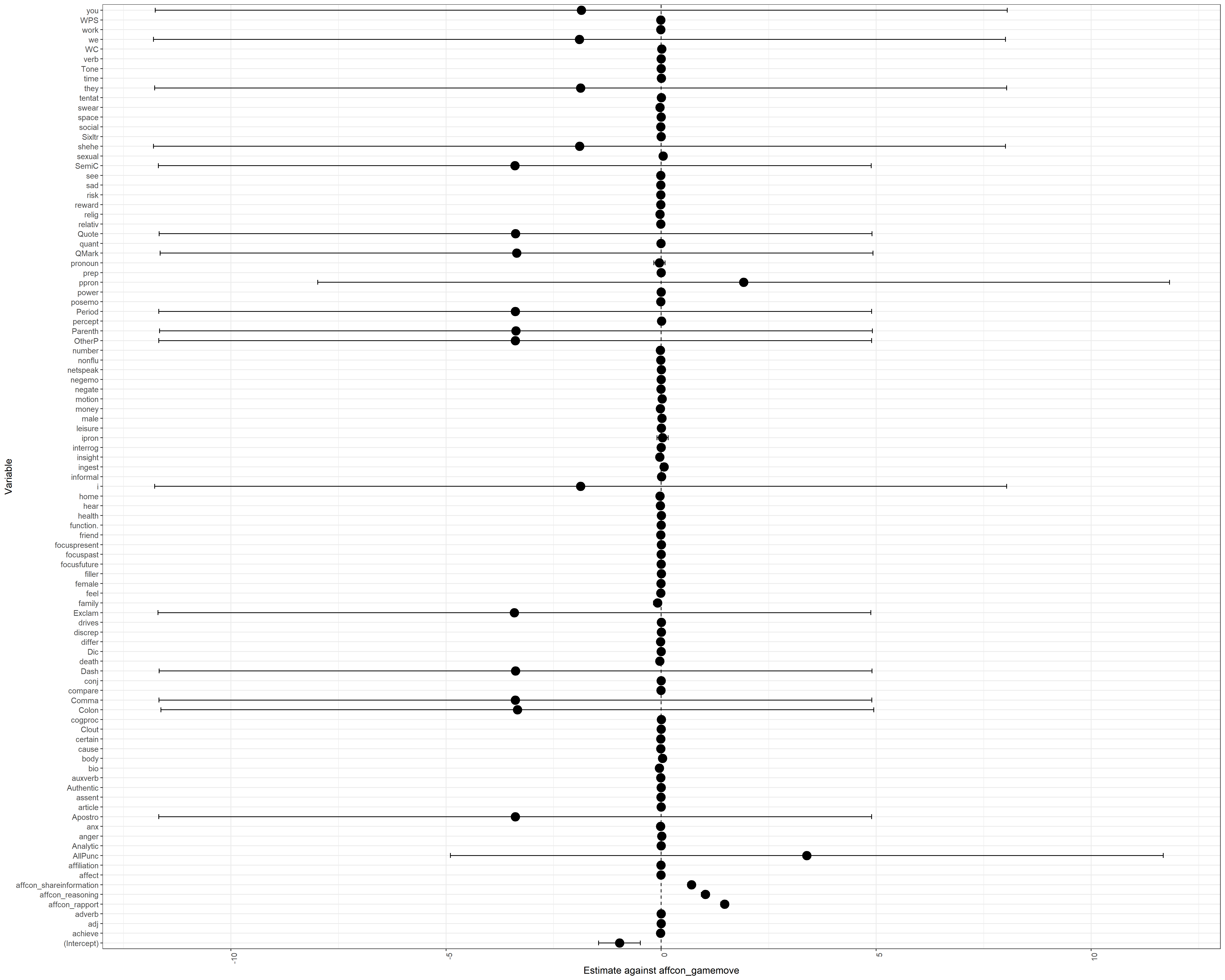
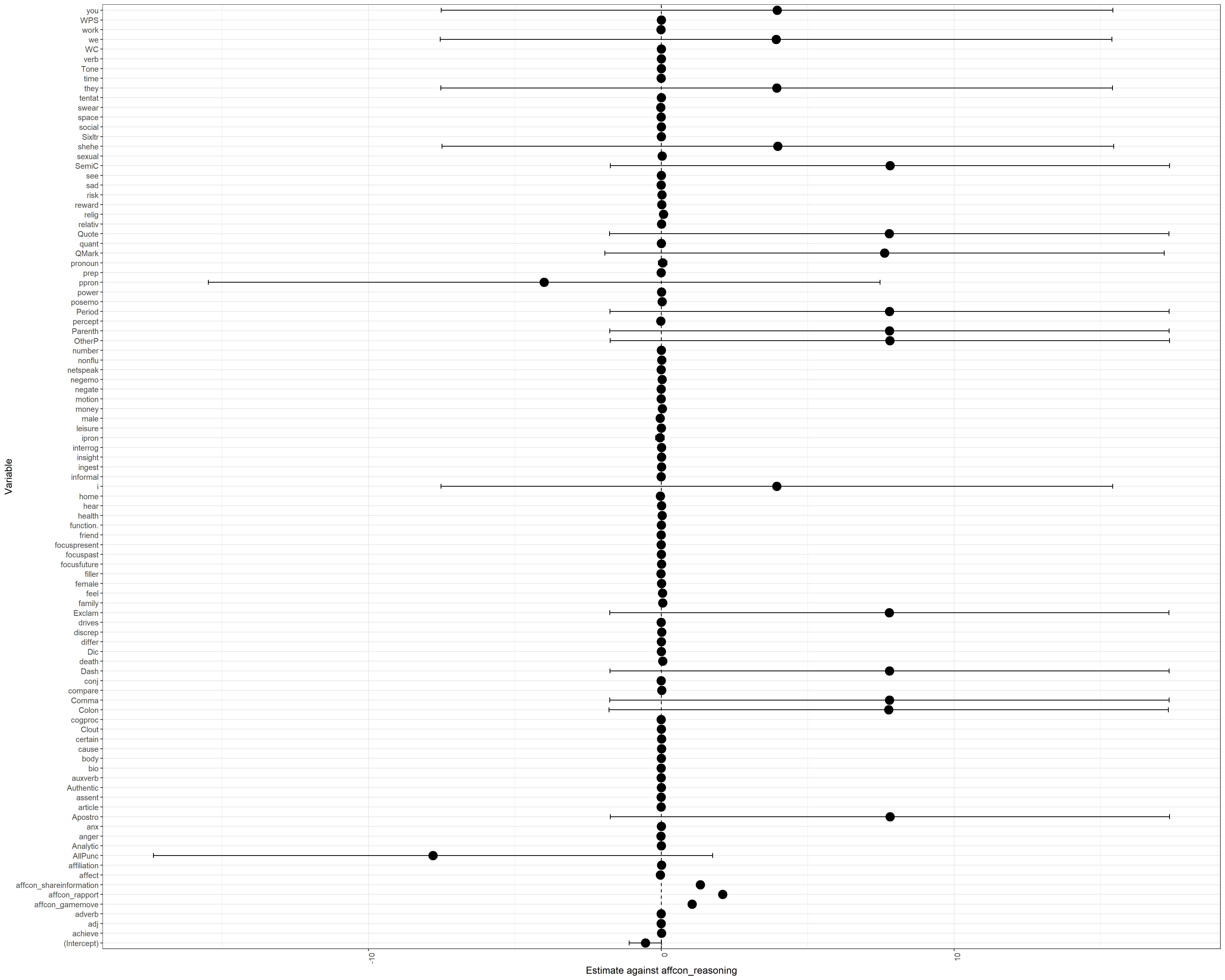
### 31 October 2020

**Logistic Regression/ Correlation between LIWC values and affcon annotations**

Correlation



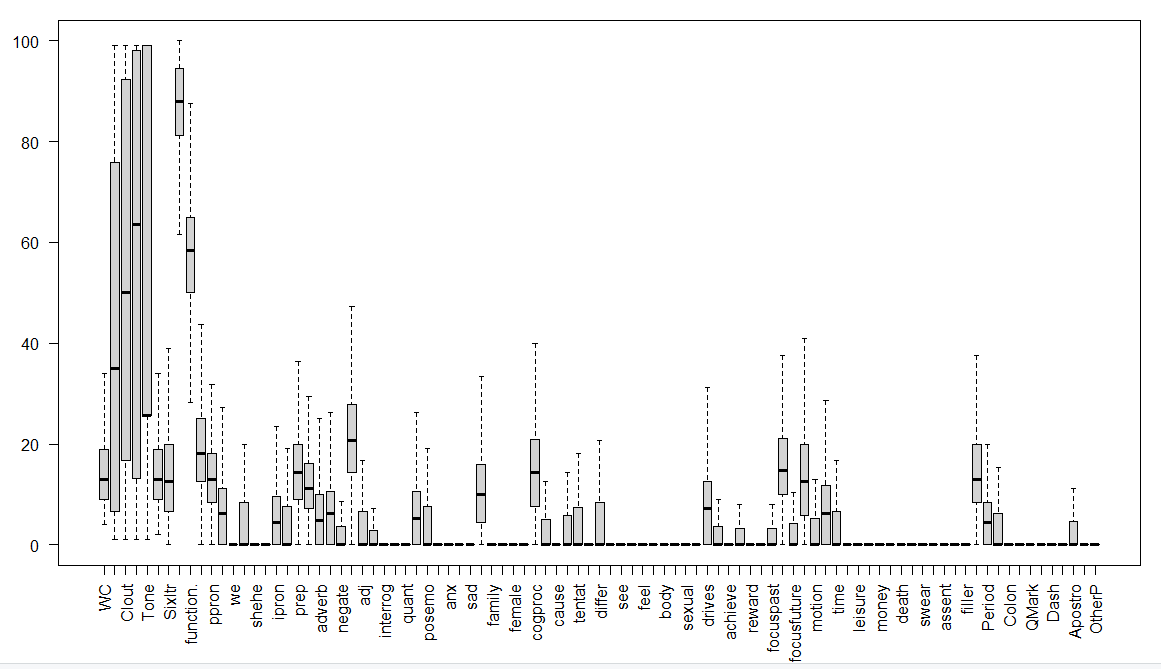
Logistic Regression

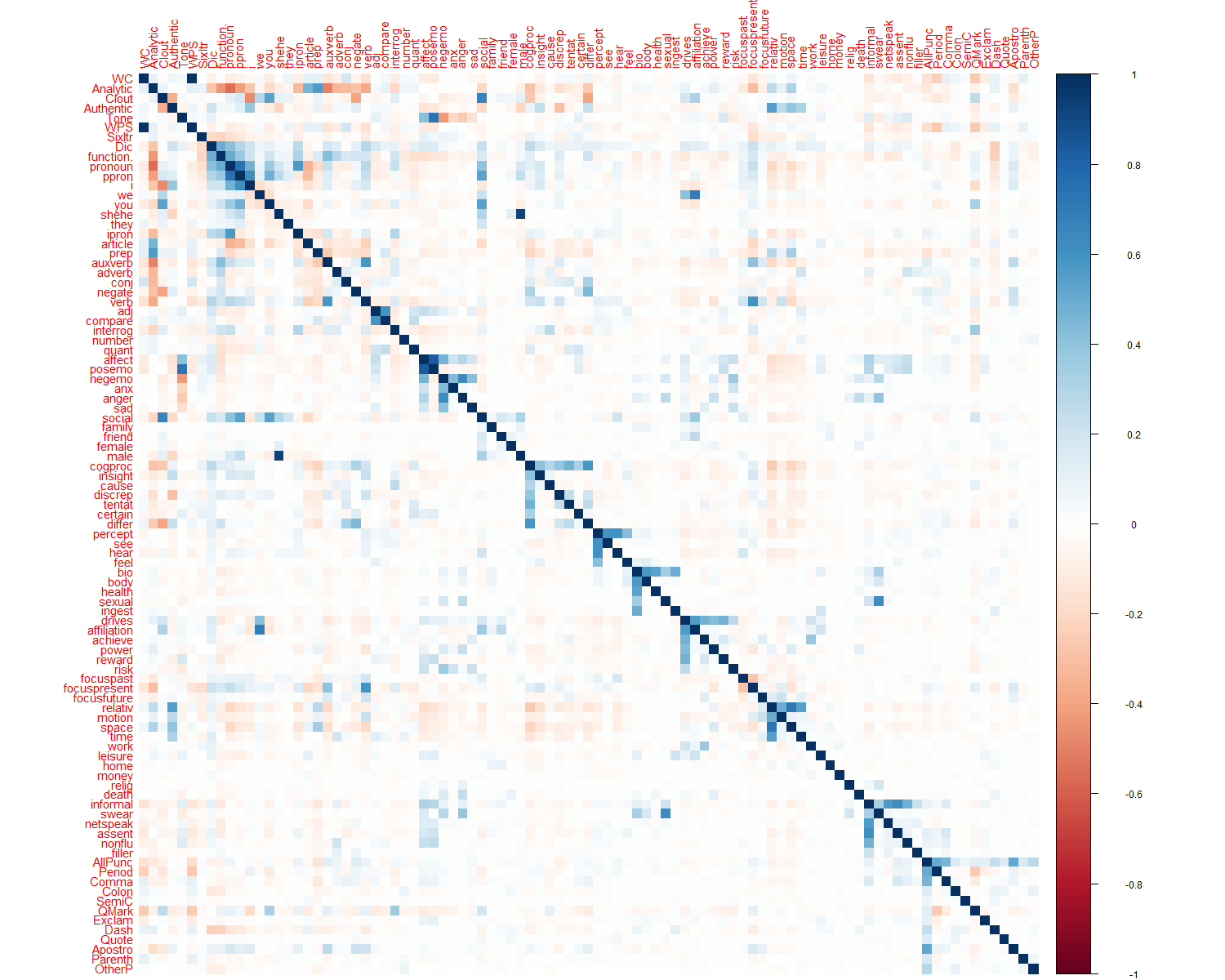


### 30 October 2020

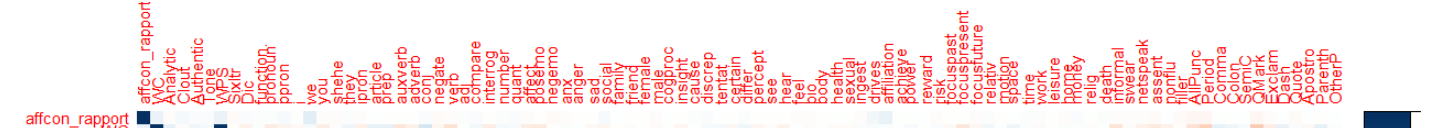
Looked at logistic regression and correlation between LIWC values vs affcon\_rapport.

**For Training set with LIWC values**





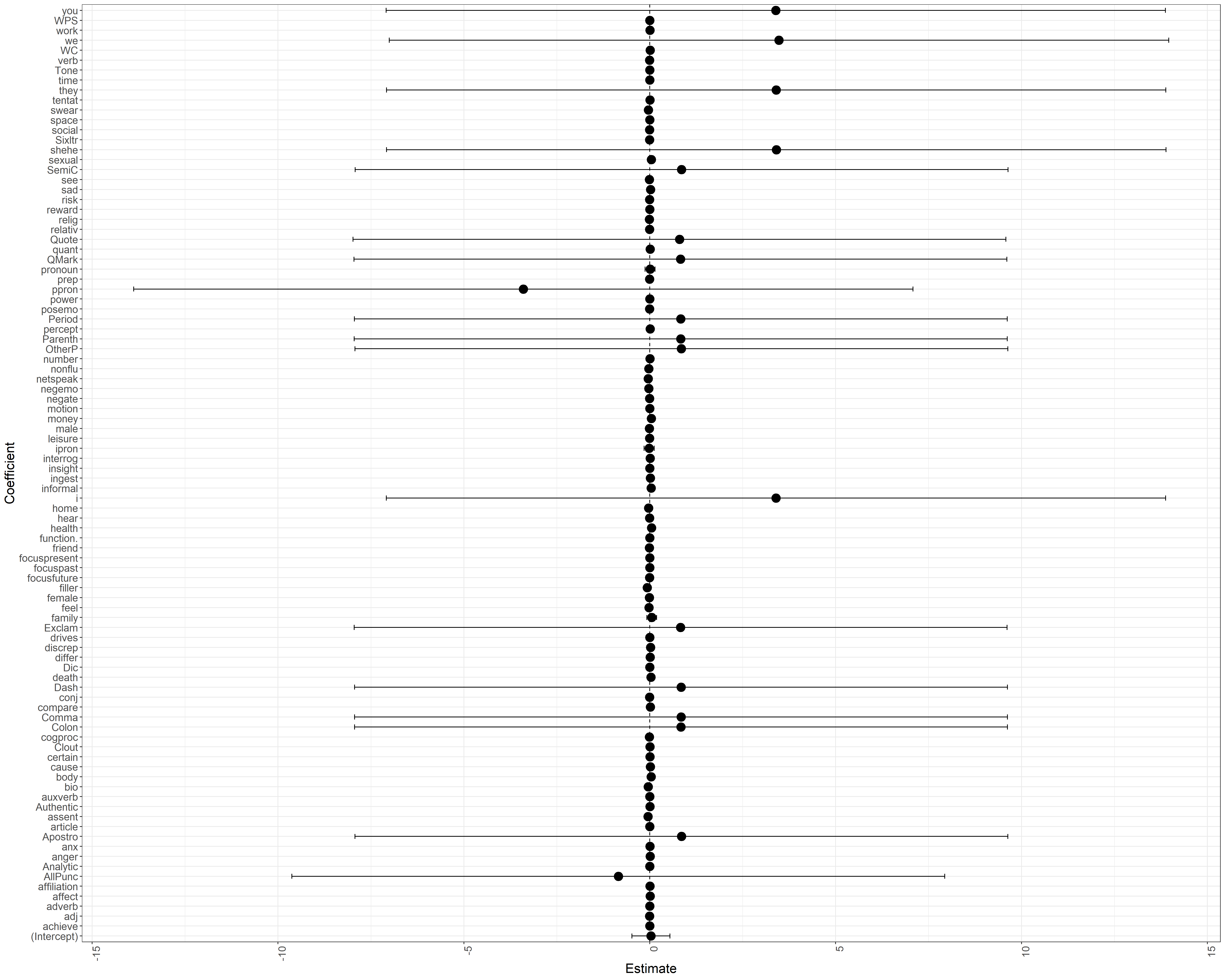
Specifically to affcon\_rapport: we/ drives/ affiliation/ WPS are strongly correlated



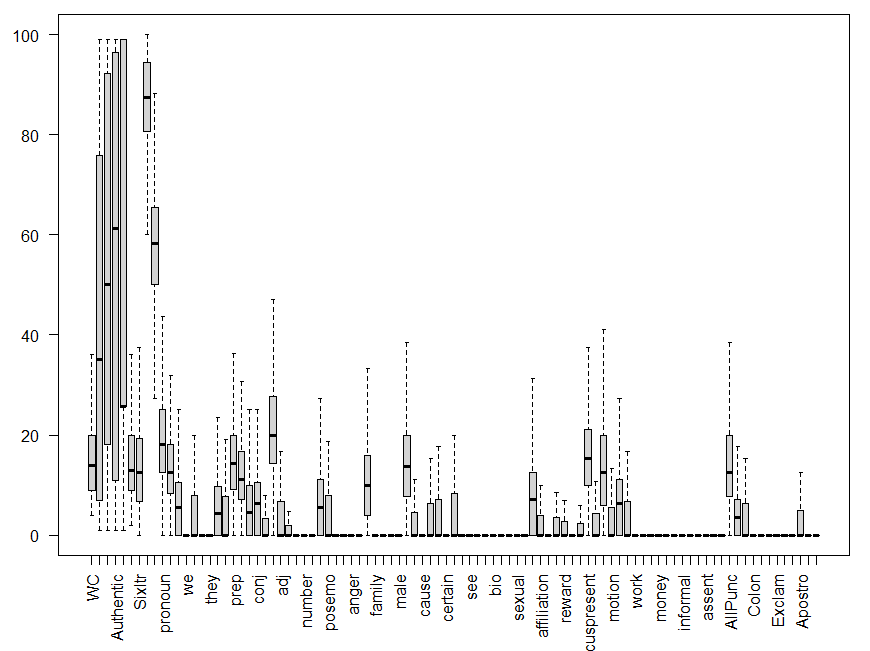
Probably see it better here

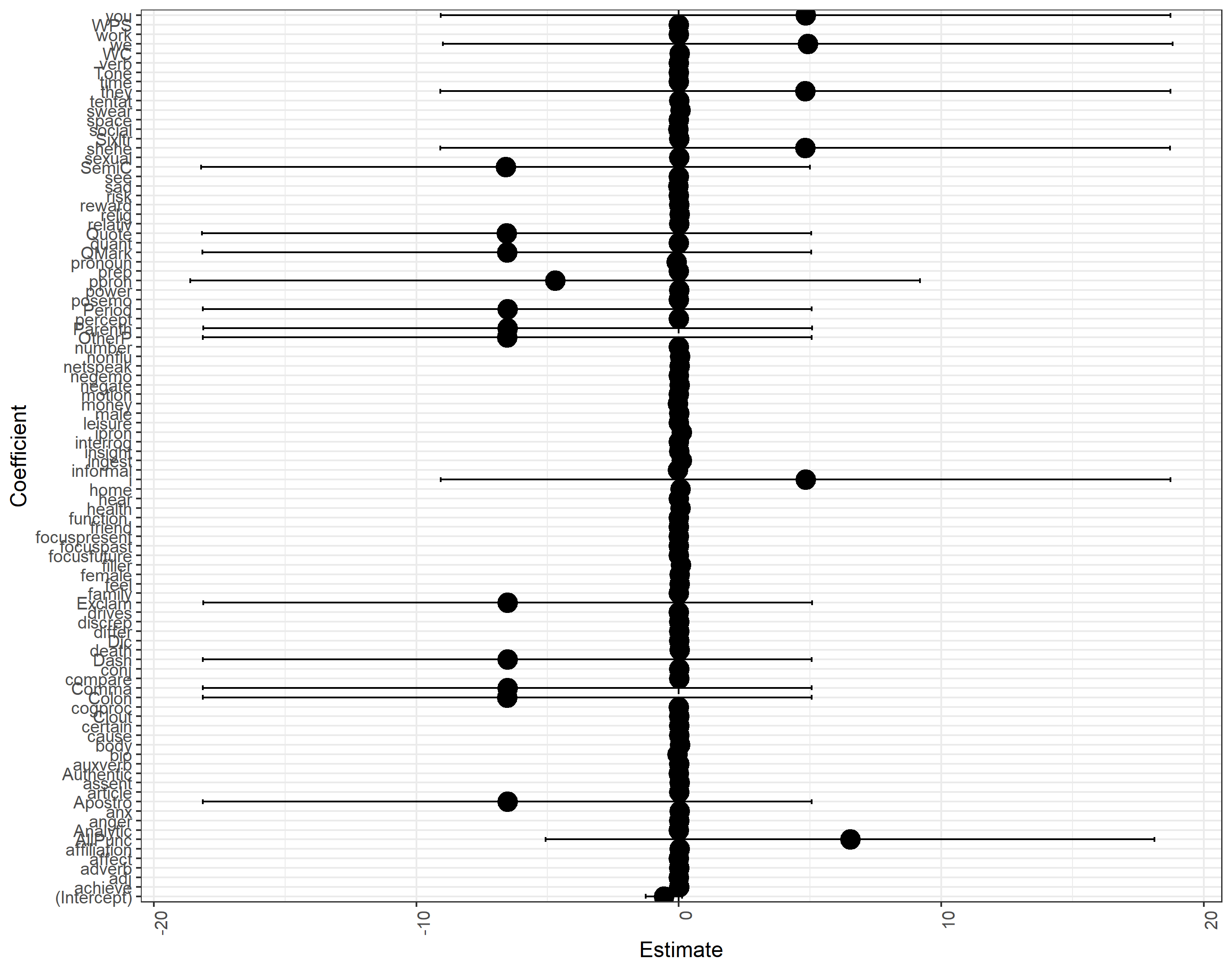
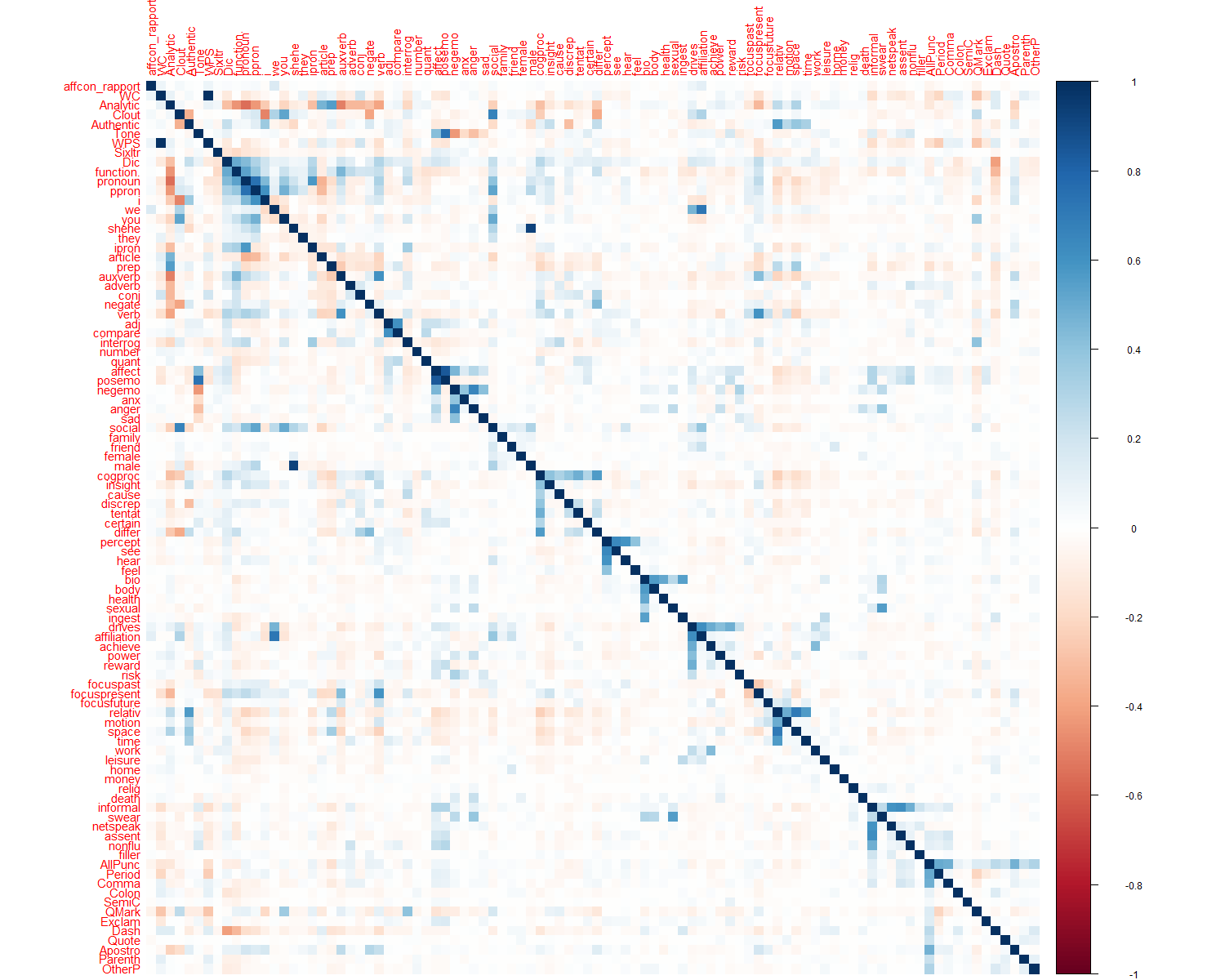


Logistic Regression and its coefficients



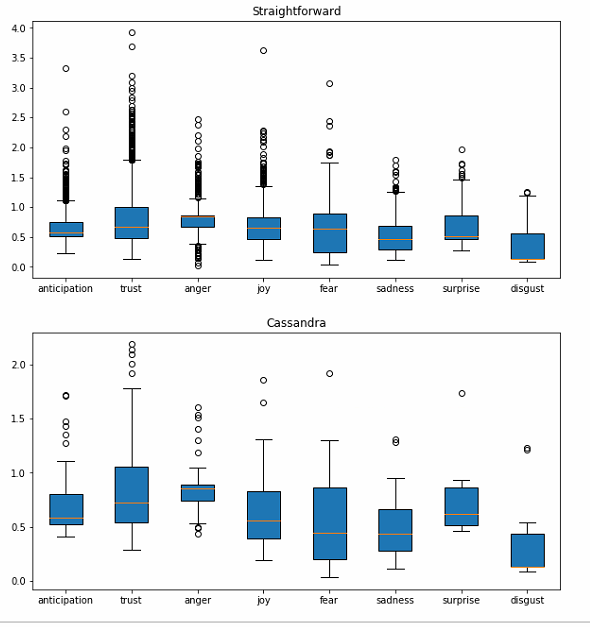
**For Testing set with LIWC**



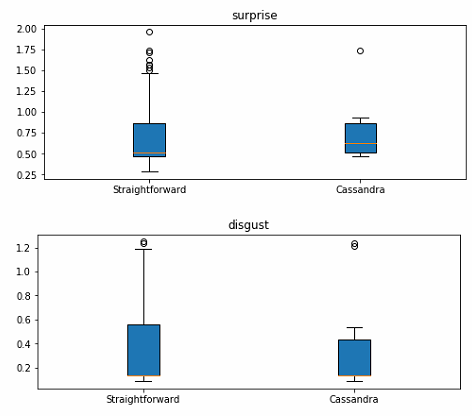
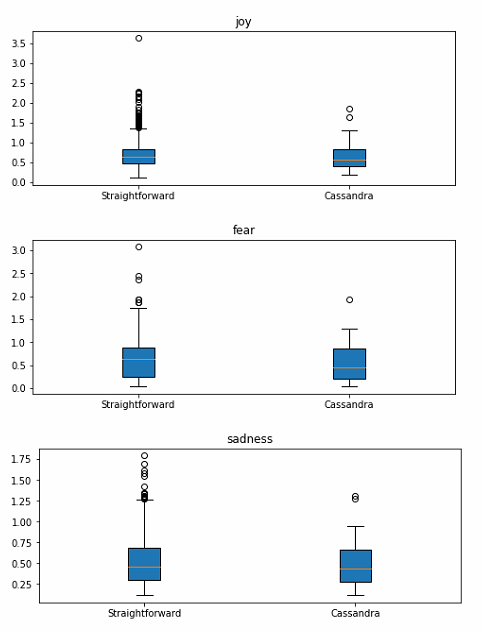
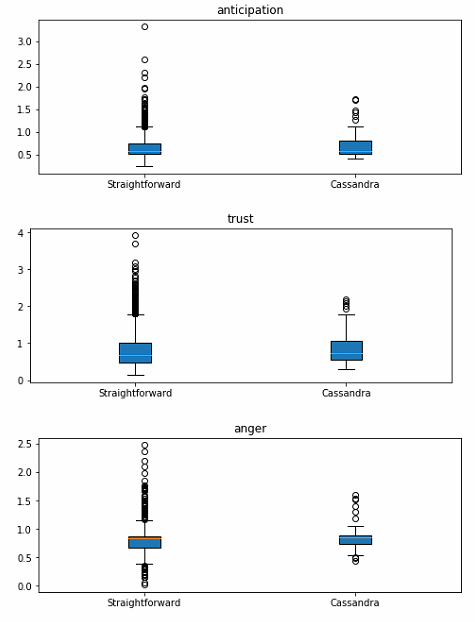


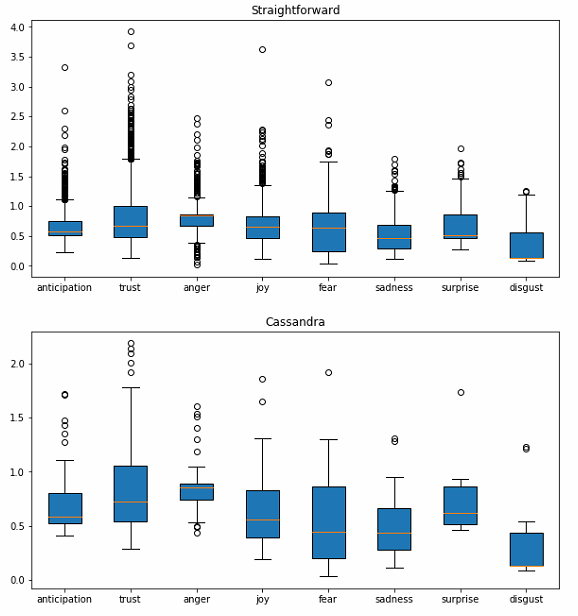
### 22 October 2020

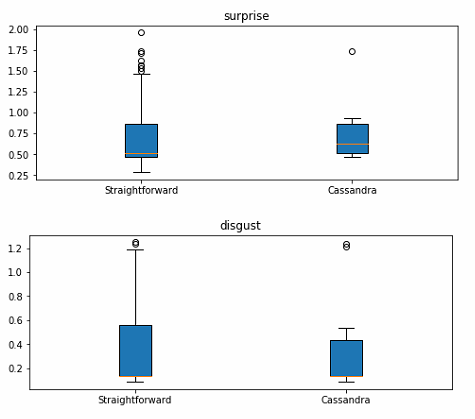
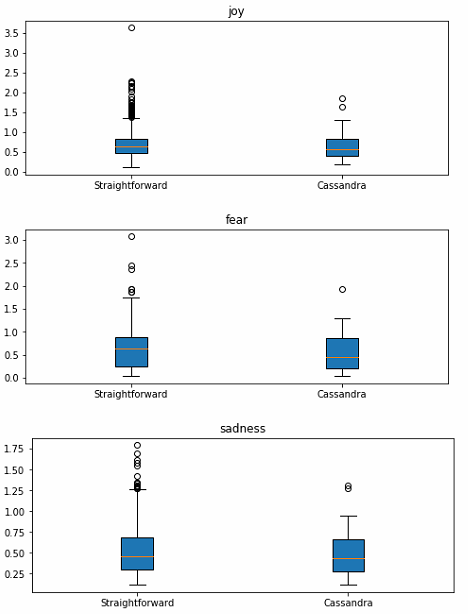
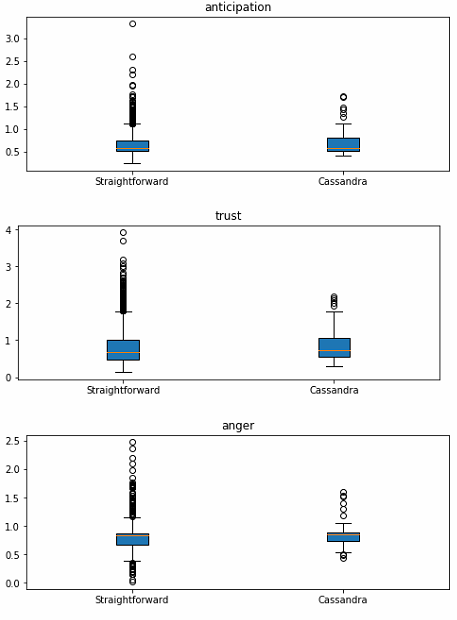
**Training set VAD**



**Test set VAD**







### 16 October 2020

Number we discard (3/5): 0.2981192

File is in AMT Oct 14 Conf/processed\_data/processed\_data.csv

Organize as 1 row is one observation (right now you have each hit as each row)

Put “1” if true, “0” if false, “NA” if you discarded the label because of 3/5 agreement.

5/5 - 1

4/5 - 1

3/5 - NA

2/5 - 0

1/5 - 0

0/5 - 0

Added a data\_readme.md in AMT Oct 14 Conf/processed\_data/data\_readme.md

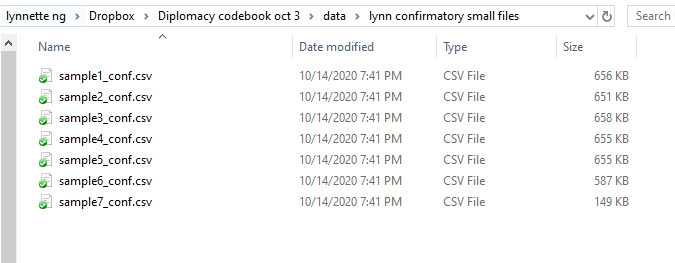
File 2: AMT Oct 14 Conf/processed\_data/processed\_data\_truefalse.csv

### 14 October 2020

7 October pilot: 73% agreement

12 October pilot: 75% agreement

Split files into 2500 lines



Action Items for Lynnette:

* Prepare data from second Pilot (Pilot Oct 12)
* Use only rapport categories
* Anything less than 3/5 is 0; above is 1
* File should contain original columns from diplomacy dataset and this label
* Create a README explaining what each column is

### 10 October 2020

First pass trip in data/df\_trimmed\_20201009.csv

Second pass trim data/df\_trimmed\_20201010.csv

Basically used genism sentence similarity module, representing each sentence as a tf-idf vector, and if vectors are at least 80% similar, I take the longer sentence.

Then I just ran these a few times.

### 9 October 2020

File to use for training and test set: five\_words\_sentences\_20201003

Can you help to deduplicate this file?

Now we want to run the annotation task but I realized that running just task 1 will cost me a bomb in the current format. If you can figure out a way to deduplicate really similar messages, it could help a lot. What we want is

1. Probably a simple method that figures out if two sentences differ only by a few words, and then keep only one of them.
2. Sounds like tfidf vectors could help here.

### 7 October 2020

Task 1 pilot scores

Question 1: Game Move

0 1 percentage 0.61 cohen -0.11683848797250862

0 2 percentage 0.69 cohen 0.08500590318772139

0 3 percentage 0.63 cohen -0.09208972845336483

0 4 percentage 0.68 cohen 0.04076738609112707

1 2 percentage 0.74 cohen 0.18800749531542793

1 3 percentage 0.64 cohen -0.12429731417863832

1 4 percentage 0.67 cohen -0.04961832061068705

2 3 percentage 0.64 cohen -0.16959064327485374

2 4 percentage 0.71 cohen 0.038461538461538436

3 4 percentage 0.65 cohen -0.16047745358090193

Question 2: Reasoning

0 1 percentage 0.61 cohen -0.11683848797250862

0 2 percentage 0.69 cohen 0.08500590318772139

0 3 percentage 0.63 cohen -0.09208972845336483

0 4 percentage 0.68 cohen 0.04076738609112707

1 2 percentage 0.74 cohen 0.18800749531542793

1 3 percentage 0.64 cohen -0.12429731417863832

1 4 percentage 0.67 cohen -0.04961832061068705

2 3 percentage 0.64 cohen -0.16959064327485374

2 4 percentage 0.71 cohen 0.038461538461538436

3 4 percentage 0.65 cohen -0.16047745358090193

Question 3: Rapport

0 1 percentage 0.61 cohen -0.11683848797250862

0 2 percentage 0.69 cohen 0.08500590318772139

0 3 percentage 0.63 cohen -0.09208972845336483

0 4 percentage 0.68 cohen 0.04076738609112707

1 2 percentage 0.74 cohen 0.18800749531542793

1 3 percentage 0.64 cohen -0.12429731417863832

1 4 percentage 0.67 cohen -0.04961832061068705

2 3 percentage 0.64 cohen -0.16959064327485374

2 4 percentage 0.71 cohen 0.038461538461538436

3 4 percentage 0.65 cohen -0.16047745358090193

Question 4: Share information

0 1 percentage 0.61 cohen -0.11683848797250862

0 2 percentage 0.69 cohen 0.08500590318772139

0 3 percentage 0.63 cohen -0.09208972845336483

0 4 percentage 0.68 cohen 0.04076738609112707

1 2 percentage 0.74 cohen 0.18800749531542793

1 3 percentage 0.64 cohen -0.12429731417863832

1 4 percentage 0.67 cohen -0.04961832061068705

2 3 percentage 0.64 cohen -0.16959064327485374

2 4 percentage 0.71 cohen 0.038461538461538436

3 4 percentage 0.65 cohen -0.16047745358090193

### 4 October 2020

Task 1 done! <https://requester.mturk.com/create/projects/1678686>

If we are going to discard *Providing information related to other game players (Yes/No) (discard*) for Task 1, can we also don’t put it in for annotation?

Because for A and B, they are looking at moves from speaker/ receiver point of views. While Providing Information doesn’t.

I presume, if we leave this out, they will be classified under General Banter? So maybe we can just have an OTHERS category?

Ans:

In Task 1, we want to identify and separate any text that’s not about speaker or receiver’s moves, or about general banter.

But sharing information is not general banter, it’s a very specific way to build trust that’s close to logos. We want to remove it so that we can have a more specific and easier Task 2 and 3, but it might constitute an implicit justification. So it might be still useful, and we don’t want to throw it completely out.

(in the worst case, sharing tactical info about others and giving justification might be useful as knowledge signals together)

**Action items (10 pm)**

Annotation instructions for (a) Task 1 and then (b) Task 3

In formatting the data, can you

(a) split into sentences keeping unique sentence ids for each row.

(b) also store postids. Post id should be same for the same utterance even if multiple sentences

(c) add a length column and a number of words column. and

(d) make a separate file of only sentences with at least 5 words. Finally

(e) create a sample of 100 sentences for first batch of annotation from the file you created in d.

**Data (10pm)**

* data/diplomacy\_corpus\_sentences\_20201003.csv: data consisting of (a) to (c)
* data/five\_words\_sentences\_20201003.csv: data for (d) and (e)

**Ethos/pathos/logos**

\* Use the dataset split by sentences

[https://requester.mturk.com](https://requester.mturk.com/)

username: dot\_tee@yahoo.com

**Task 1**

1. Is the sentence talking about moves? (Yes/No) (use it to get ethos)

[Personal stance]

[Counter-offer, push-back, clarification, reciprocity]

[Accept proposal]

Providing information related to other game players (Yes/No) (discard)

[Sharing tactical information related to others]

1. Giving reasons yes no (all of this is logos)

[Justification for POA, speculation, hindsight]

1. General banter yes no (use it to get pathos)

[Reassurance]

[Personal thoughts but not strategy, sharing emotions]

**Task 2**

**Take the positive annotations from A & B**

If yes, is it

From A:

1. Personal stance/focusing on self-interest/exerting authority, expertise, qualifications (Ethos)

(yes/no)

1. Counter-offer, push-back, clarification, reciprocity (yes/no)
2. Accept proposal (yes/no) (pathos)

From B: Justification for POA, speculation, hindsight (logos) *(might not need to run this and just use all of this)*

*(anything that’s not a b or c is probably a proposal)*

**Task 3 General banter** Yes/No

**Take the positive annotations from C**

1. Reassurance. Intention to calm anger, reduce fear and anxiety or explain a surprise. (Pathos)
2. Personal thoughts but not strategy. Appealing to emotions by sharing emotions.  (pathos)

-------------------------------------------------------------------------------------------

Task 3

Take just general banter - reassurance

Clarifications/ help needed from Kokil:

1. Should “Talk informally/all other” be under “General Banter”?
2. Did up a quick annotation instructions for mturkers – perhaps could start trial run? I have not done this before, so I need some help!
3. Did code up sentences
4. Should we split each item by sentences or by newlines? Currently it’s split by new lines, mimicking when the players type an enter on their keyboard – ie, like a chat room, players may type multiple sentences before they hit the enter key, or a single one
5. Open the csv file with UTF-8 encoding then you won’t have the strange symbols@#?

### 3 October 2020

New coding scheme at:

<https://www.dropbox.com/sh/ivsyubfb5gcpwwb/AAAARD9Q6Jeczltz5DrkpdW3a?dl=0>

Action items for Lynnette:

1. Read all my examples in the word document (Diplomacy codebook)
2. See if you can code up your first 30 sentences following this codebook
3. Try around 100 more.
4. Do we need to simplify further? (Probably yes) Set up a time to discuss with me.
5. **Can we create simple instructions? Start drafting instructions for mturkers. Examples from last time are in the same dropbox folder and also** [here](https://www.dropbox.com/s/shw9mlvaga3ulbv/example_Annotation%20instructions%20for%20Disclosure.docx?dl=0)
6. **Simplify sentences (Some splits have more than one sentence)**

**Kokil’s proposed categories (sentences can be in more than one category)**

|  |  |
| --- | --- |
| Sharing tactical information | Sharing tactical information about others |
|  | Personal stance/focusing on self-interest |
| Discuss game moves | Propose a plan of action |
|  | Counter-offer, pushback, clarification, invite their point of view |
|  | Accept proposal |
|  | Reject proposal |
| Discuss strategy | Justification for POA |
|  | Speculation |
|  | Hindsight and clarification after a move |
| Build rapport and trust (or conversation positive) | Personal thoughts but not strategy |
|  | General banter |
|  | Reassurance |
|  | Greet, thank, and compliment |
|  | Talk informally/all other |
|  | Apologize, apologize and explain |
| Disagreement (or conversation negative) | Didn’t find any examples |

### 26 Sept 2020

Split messages into sentences and uploaded to Github (<https://github.com/quarbby/special-umbrella/blob/master/diplomacy_corpus_sentences.csv>)

To do:

* One column for game
* Figure out ranking to each of the themes of who won the games

### 21 Sept 2020

**is this about the game or is this general conversation?**

1. **yes: coding scheme for negotiation**

1. 1994 negotation coding scheme paper (attached)

1. Propose for acceptance

2. Acknowledge receipt

3. Reject proposal

4. Acknowledge reject

5. Accept proposal

5. Counter

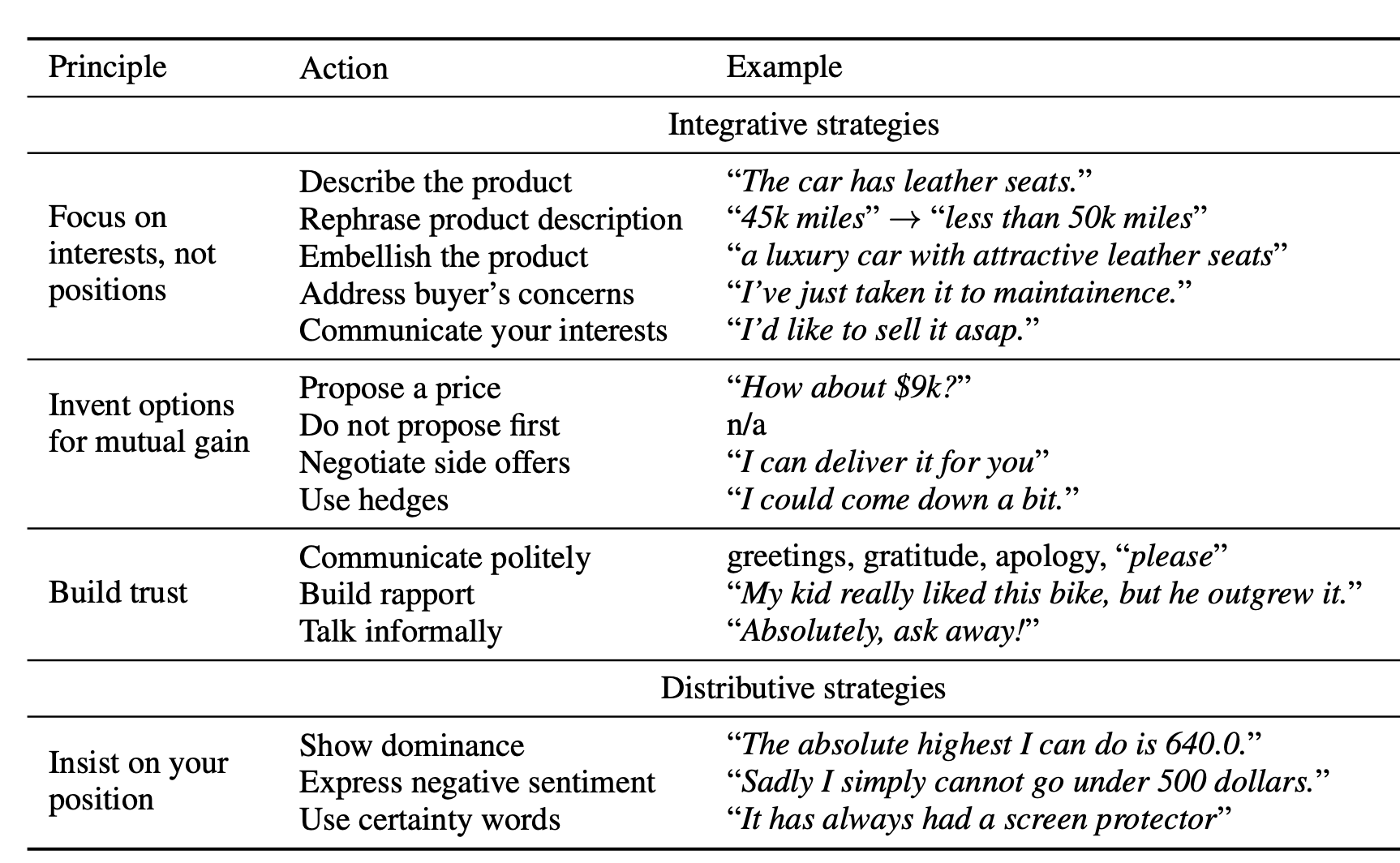
6. Retract proposal

9. Propose replace

2. Charity donation coding scheme…

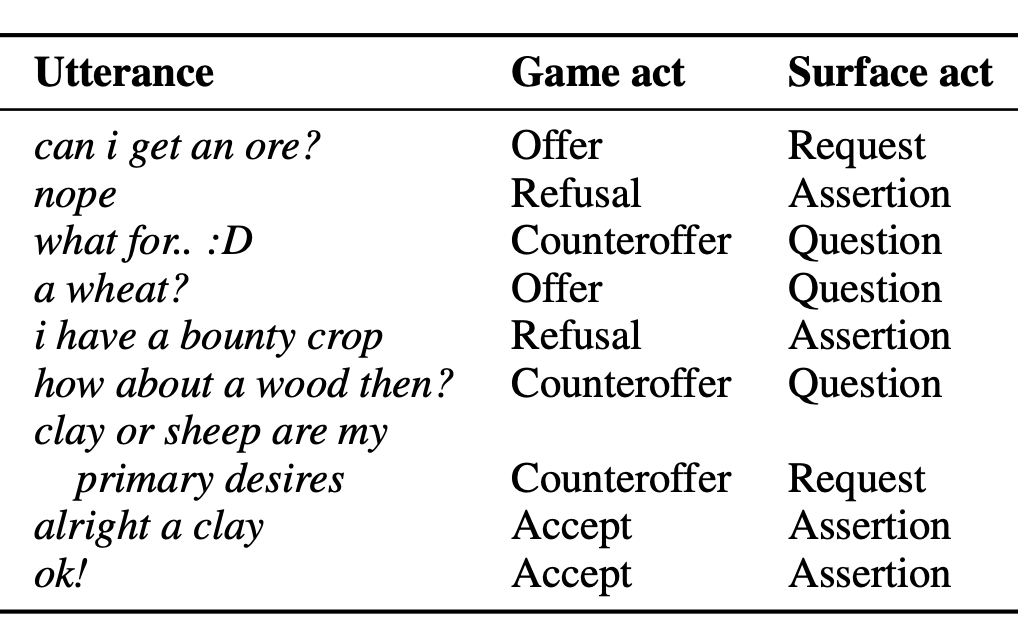
3. Negotiation scheme

<https://www.aclweb.org/anthology/W19-5943.pdf>



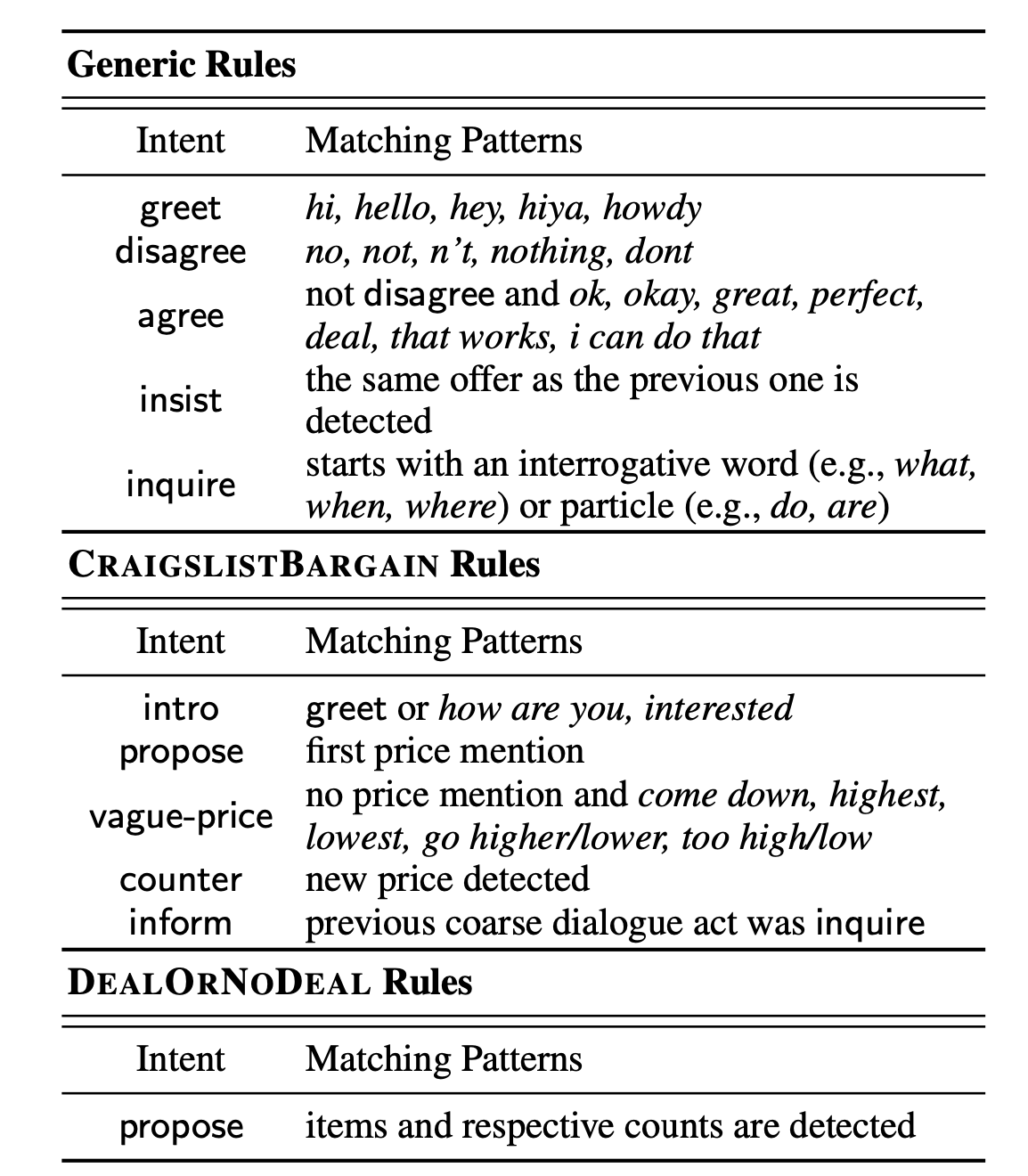
4. Settlers of Catan Negociation Strategies

<https://sorinmd.github.io/downloads/EACL-2017.pdf>



5. Decoupling negotiation

<https://www-nlp.stanford.edu/pubs/he2018decouple.pdf>



1. **no: Still useful to understand interpersonal relationship. <Do we need to find a coding scheme?>**

Examples:

Whatever I did, I'm sorry.

Hey. Want to talk?

Old friend! Sure, why not?

Good morning

**[Lynnette’s proposal]: Combine (5) with (1)**

Because (1) is a bit difficult to understand and categorise. (5) also takes care of generic rules like “insist” and “inquire”. So something like this…

|  |  |
| --- | --- |
| **Negotiation Rules** | **Examples** |
| **General Conversation** |  |
| Greetings | Germany!  Hi, are you there?  Just woke up. |
| Discourse, General | 👍  Okay, thanks for telling me  Just an FYI: I’ve now had both England and France suggest to me that I should move to Tyrolia and France will support me to Munich in the Fall. One saying that to me is not a big deal, but with both mentioning it, my alarm bells are going off. I am concerned about an E/F.  England did return my message, but he did not tell me anything substantive so I really don’t know what he’s doing. I’m planning to move towards Turkey. |
| Discourse, on Trust  (too implicit)  cant separate trust from negotiation | I am pretty conflicted about whether to guess that you were telling the truth or lying about the “brewing lagers” thing. I am going to take it literally and say 👎 even though I don’t think you meant it deceptively. 😉  But I think I can get over “Lagergate” and we can still be friends.  As of right now, I think Austria may be my most reliable ally.  Well, at least I have an idea on who to trust. Obviously, my ideas are subject to change.  I understand your desire to watch behavior before committing to anything. I, personally, am a partner player. I look carefully early in the game for a small group to work with, and then I value loyalty and collaboration. I like to work closely with a tight-knit alliance.  Did France tell you he was moving to Burgundy, or was that a stab? |
| Discourse, on cooperation | If you prefer to hop and back and forth, or play more of an individual game, then we might not be a good match.  I’m looking for a loyal ally or two that I can coordinate with and make awesome moves with.  So, um, no alliance then? [Or is this Question proposal?]  I’m certainly not moving to Tyrolia. But I just want you to be cautious here. I feel like England and France are working together. |
| Discourse, on proposed ideas | I'm trying--I just offered to assist with taking Sweden in exchange for some assistance into Belgium...not sure if they'll go for it… |
| Question, inquire | So… do i suggest it?  What is your appetite like for unusual and crazy?  You’ve whet my appetite, Italy. What’s the suggestion?  Any thoughts?  ...what are you and Austria planning for this year, then? I'm willing to tell you my plans in exchange as a gesture of trust.  Okay, I can dig it. What do you want me to do? |
| Agree |  |
| General | Sorry Italy i’ve been away doing, um, German things. brewing Lagers? |
| **Negotiation** |  |
| Intro, propose | Just the person i want to speak with. i have a somewhat crazy idea that i’ve always wanted to try with I/G, but i’ve never actually convinced the other guy, and it might make you suspicious of me |
| Propose | I’m thinking that this is a low stakes game, not a tournament or anything, and an interesting and unusual move set might make it more fun. That’s my hope anyway  I think you need to get England on board to attack France.  I suggest that you order:  Kiel Support Berlin holding  Berlin Support Munich holding  Helg to Holland  Munich Support Berlin holding  I do think you’re best off moving to Burgundy. And there is some chance that we fail this turn. But I think we just take a guess and hope for the best. We’ll get him next turn if not this one. |
| Acknowledge proposal  Keywords: “consider proposal” | I’ll consider this proposal, but, basically, I’m not going to expose myself to risk from either of you until i’ve seen a bit of your behaviour  Hi Germany, I’ll certainly consider that. Though, I’ll note: traditionally, Germany would help Italy to Marseilles if the two of them work together there. The reason is that: if I help you to Marseilles, I’m basically cut off from going west and getting anything myself. So, usually, Germany would help Italy into Marseilles to encourage Italy to come west and Germany would plan to take Paris, Belgium and Brest. |
| Reject proposal/  Question proposal  Keywords: “I don’t think I’m ready” | It seems like there are a lot of thiways that could go wrong… i don’t see why France would see you approaching/ taking Munich -- while i do nothing about it and don’t feel skirttish  Yeah, i can’t say i’ve tried it and it works, cause i’ve never tried it or seen it. But how i think it would work is (a) my Spring move looks like an attack on Austria, so it would not be all surprising  I don’t think i’m ready to go for that idea, however I’d be down for some good ol fashioned Austria-kicking?  I don’t think I can afford to move to Piedmont this season. I don’t really trust Austria to avoid walking through that door if I leave it wide open. |
| Acknowledge Reject of Proposal | That's valid. And actually I was conferring with England and we concluded that it's not really gonna be possible for me to help you take Marseilles this year anyway. |
| Retract proposal/ Amend proposal | All that said, i’ve literally never done it before, and it does involve risk for you, so i’m not offended or concerned if it’s just not for you. i’m happy to play more conventionally. Up to you.  I am just sensing that you don’t like this idea, so shall we talk about something else? That was just a crazy idea I’ve always wanted to try. I’m happy to play more conservatively. |
| Accept proposal | I have good news! (1) I am finally attacking France this turn. (2) I will be supporting Munich to hold from Tyrolia.  I agree completely--although I didn't know that a country could hold \*and\* support at the same time! Thanks! |
| Counter Proposal | I do think you’re best off moving to Burgundy. And there is some chance that we fail this turn. But I think we just take a guess and hope for the best. We’ll get him next turn if not this one. |
| Threats |  |

**[Cool list of resources and papers from CL literature]** <https://github.com/rowatc/Diplomacy-AI>

**[Game Strategies** <http://www.diplomacy-archive.com/resources/strategy.htm> **- but i dont know how to codify/ use it]**

### 20 Sept 2020

Action items:

1. Lynnette to check coding schemes in CL conf proceedings
2. Kokil to check coding schemes in discourse and linguistics literature
3. Kokil to update github page to explain training set
4. Kokil to send out CFPs

--------------------Detailed notes below

Sender > Focus only on Truth (Straightfoward & Cassandra)

Reciever > Focus on Lie & Truth

Explore linguistic facets that predict trust

First part, get text annotated for negotiation strategies

Game score + game score delta to create a game\_score\_recipient = game\_score – game\_score\_delta

Created a column on which game it was

Want:

* One column for game
* One column for end of number of reply\_to for sender/ recipient)
* Figure out ranking to each of the themes of who won the games

Predictive performance: Negotiation strategies

Insights: Use Truth

Look at 10-15 of these and see if they map to the negotiation strategies.

Is this about the game or is this general conversation?

Yes: coding scheme for negotiation

1. Propose for acceptance
2. Acknowledge receipt
3. Reject proposal
4. Acknowledge reject
5. Accept proposal
6. Counter
7. Retract proposal
8. Propose replace

No: irrelevant

## Criteria for a good coding scheme:

1. Can annotate at sentence level.
2. Easy
3. Meaningful
4. Great coverage

Role of questions? A lot of them have questions.

Explore ACL literature eg. MTurk task (Persuasion for Good)

## Is this about the game or is this general conversation?

{"PA","ACK", "UNK"}

### yes: coding scheme for negotiation

A1. 1994 negotation coding scheme paper (attached)

1. Propose for acceptance

2. Acknowledge receipt

3. Reject proposal

4. Acknowledge reject

5. Accept proposal

5. Counter

6. Retract proposal

9. Propose replace

A2. Charity donation coding scheme…

### B. no: Still useful to understand interpersonal relationship.

Examples:

Whatever I did, I'm sorry.

Hey. Want to talk?

Old friend! Sure, why not?

Good morning

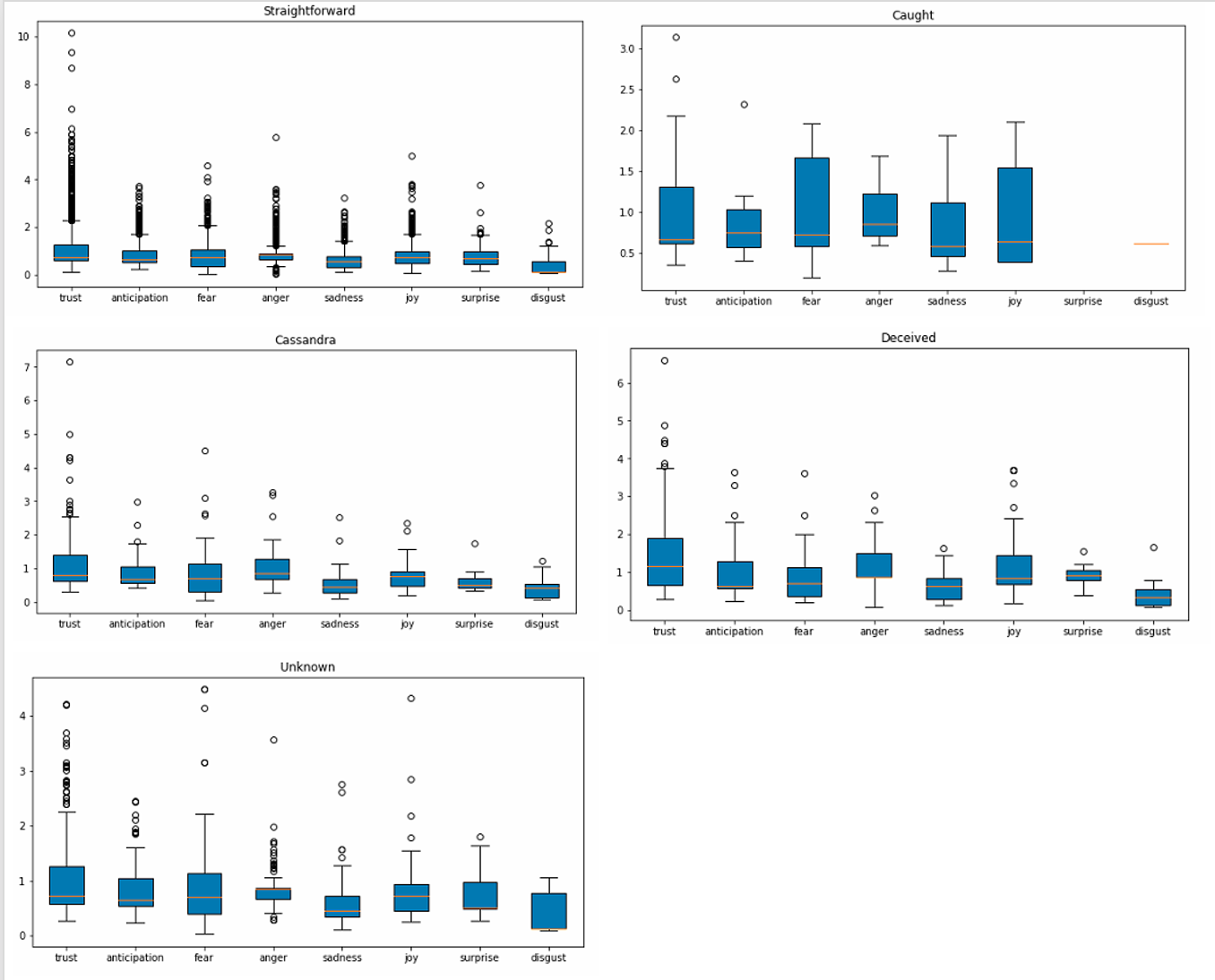
**18 Sept 2020**

Ran emotion intensity lexicon from  <http://sentiment.nrc.ca/lexicons-for-research/>. Chose it because most of the words are inside, word-sense and word lexicons do not cover all the words

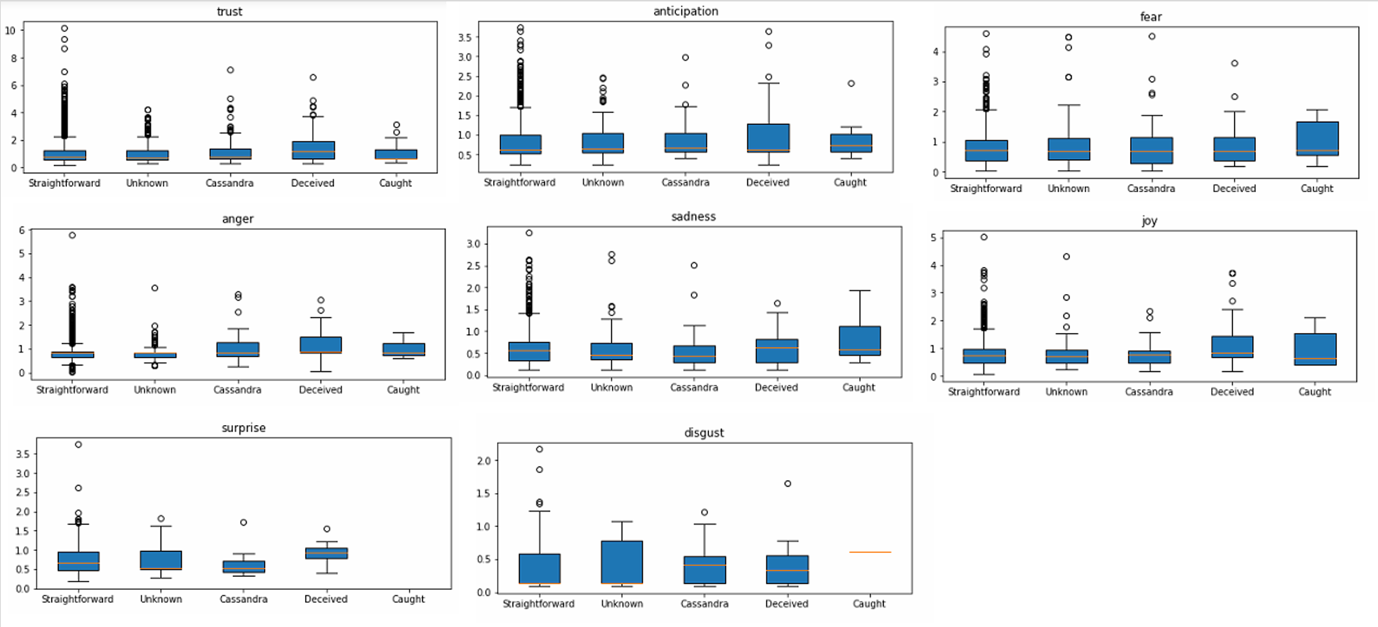
* For each word in the sentence, tokenize it, and find its match on the emotion intensity lexicon
* If matched, it has an emotion and a corresponding emotion intensity.
* For each sentence, sum aggregate the emotion intensities. The emotion with the highest intensity is a dominant emotion. Dominant emotion and its corresponding intensity is used to plot the box charts.

Can also run Emotion VAD lexicon but a bit unsure how you’d want it to be ran. I can extract for the words in each sentence, the corresponding VAD, but how might we aggregate them?

Emotion Intensity box chart per Deception type

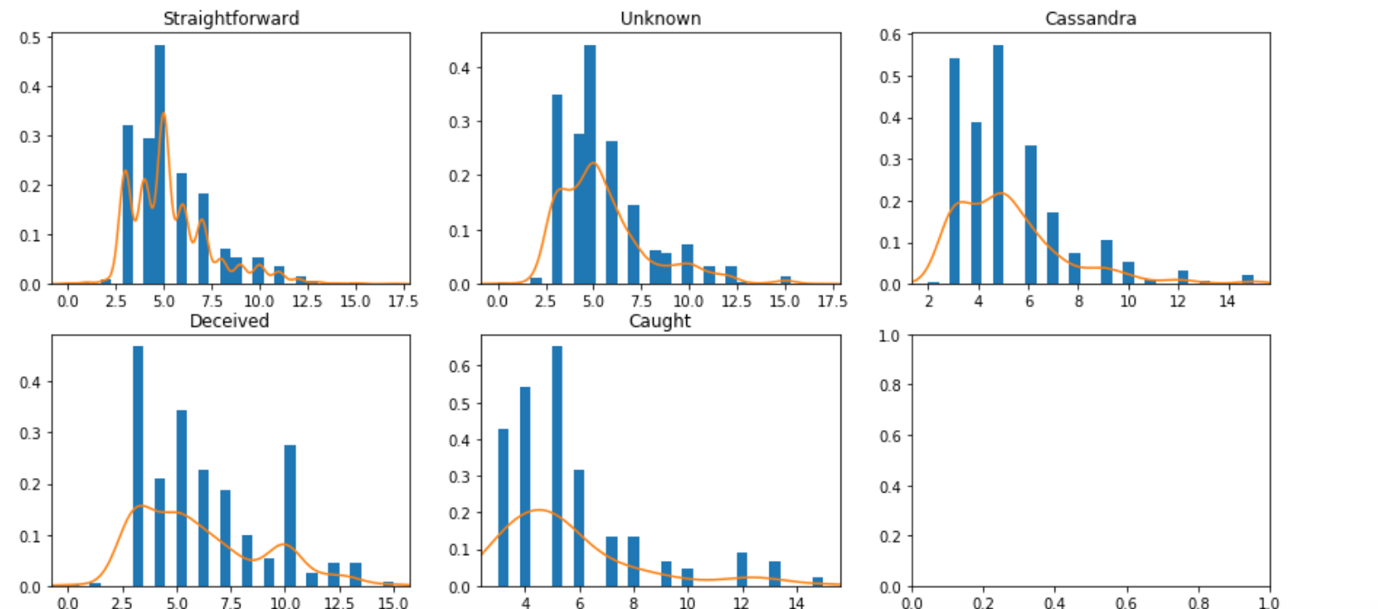
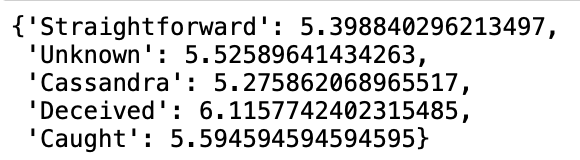


Emotion Intensity per Emotion type

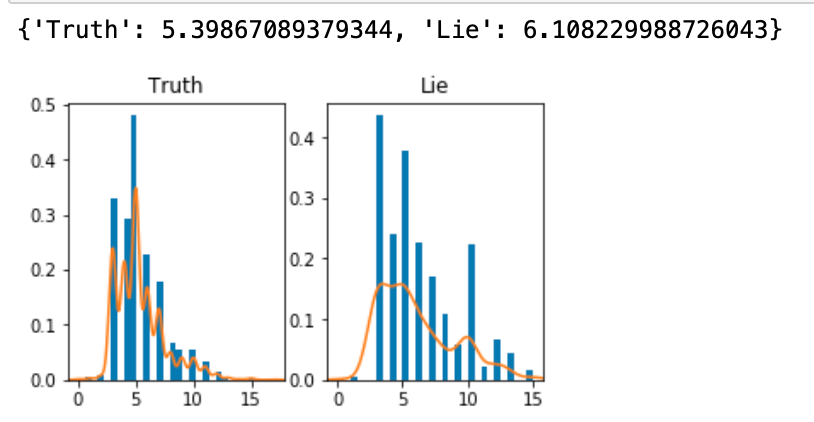


**10 Sept 2020**

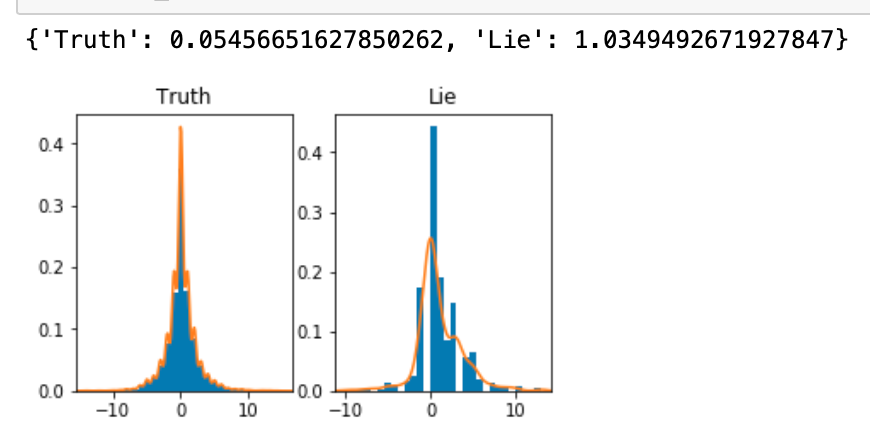
Average game score vs whether the receiver was deceived



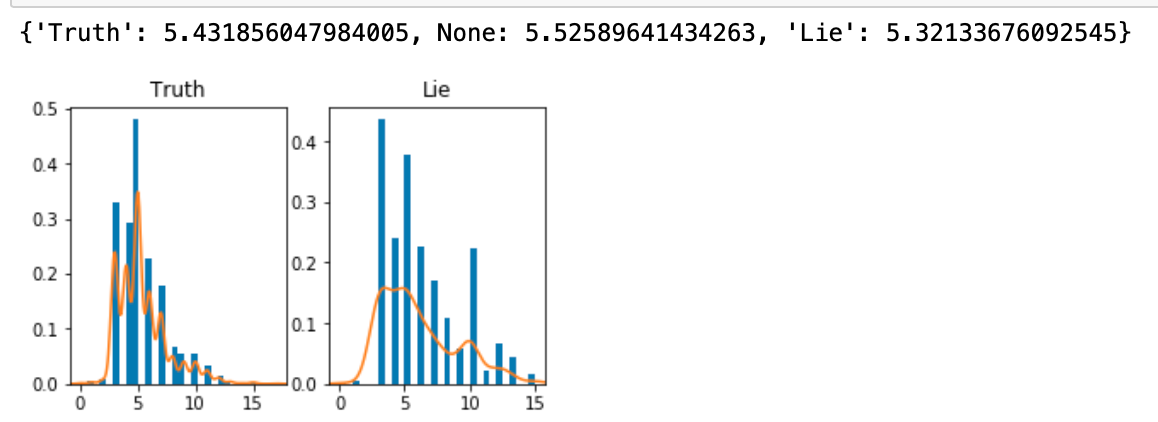
Average game score vs sender intention



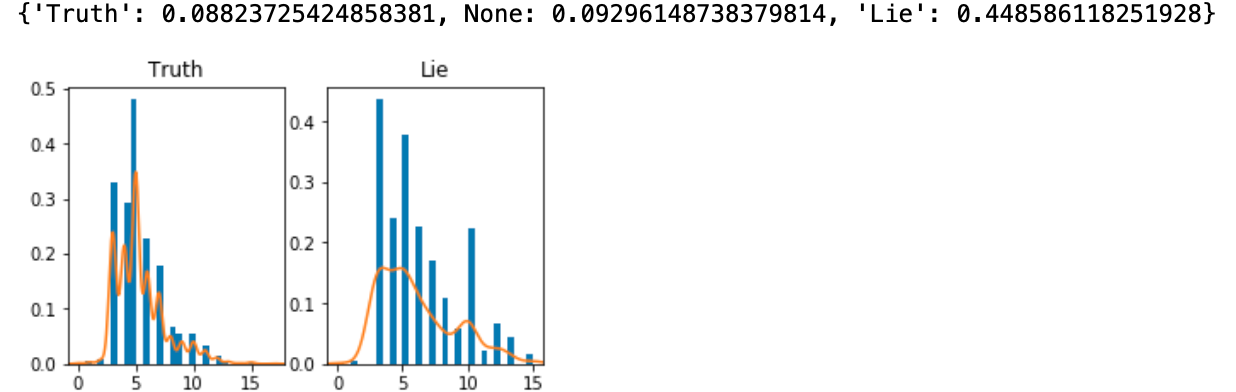
Average game score delta vs sender intention



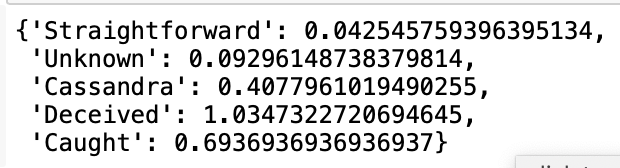
Average game score vs receiver perception

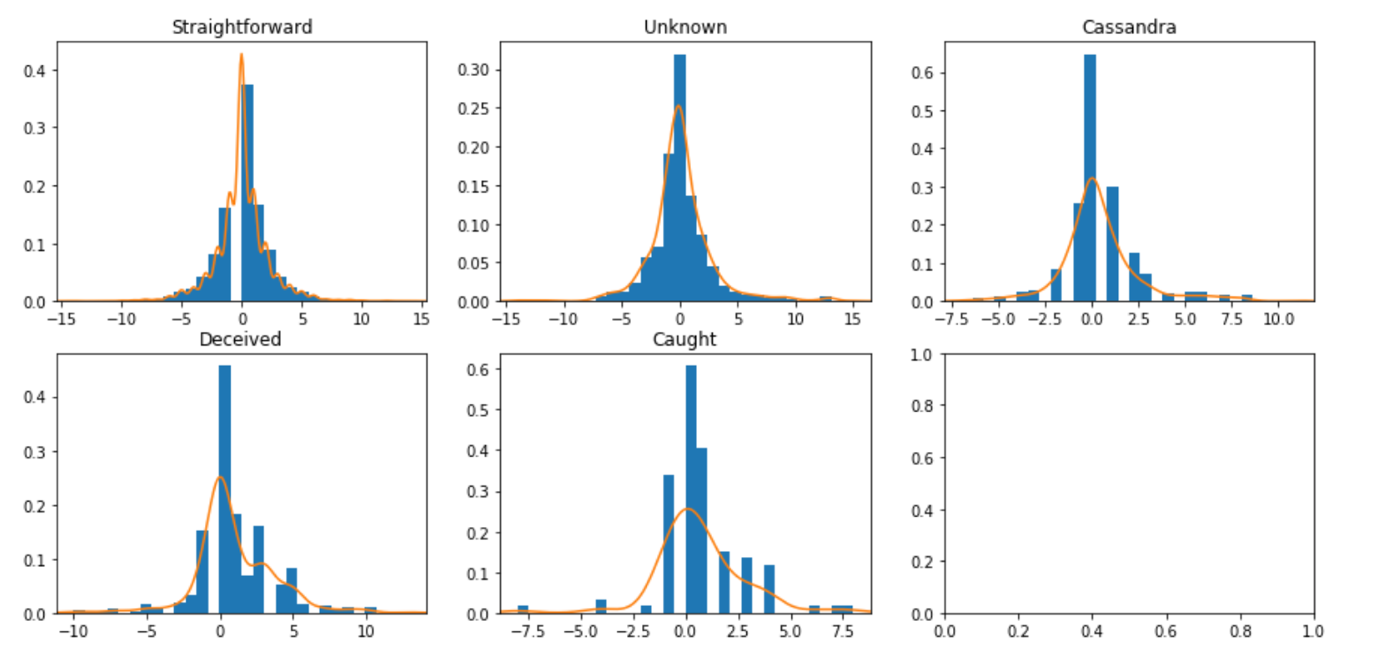


Average game score delta vs receiver perception



Average Game score delta vs whether receiver was deceived





**6 Sept 2020**

Made use of the Diplomacy dataset within Christian’s Convokit library (<https://convokit.cornell.edu/documentation/diplomacy.html>) -> Relevant ACL paper <https://www.cs.cornell.edu/~cristian/Deception_in_conversations.html>

From Kokil:

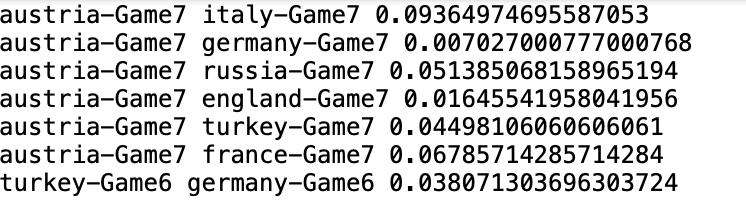
Can you use the game\_id field to try and figure out (a) who are the players, and (b) which ones are working together? The ones that work together won’t be deceiving each other.

Even if you are able to do (a) and calculate some average score of the pairwise exchanges, we would know which ones are working together.

Finding Coordination

Made use of Coordination from Christian’s Convokit Library (<https://github.com/CornellNLP/Cornell-Conversational-Analysis-Toolkit/blob/master/examples/coordination/examples.ipynb>)

Which could calculate within a game, some coordination scores between a person to another. That means, linguistically, these people start to talk like each other, so that means they are likely to coordinate, or work together. I guess this is one interpretation.



However, there are issues with computing pairwise\_scores, which is what we want. It does not work consistently.

coord.pairwise\_scores(corpus\_2, [('russia-Game8', 'germany-Game8')],

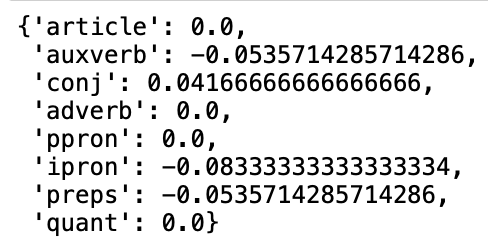
target\_thresh=0, speaker\_thresh=0)

***Could I email Christian back on this?***

Overall Game Coordination vs coordination within the game

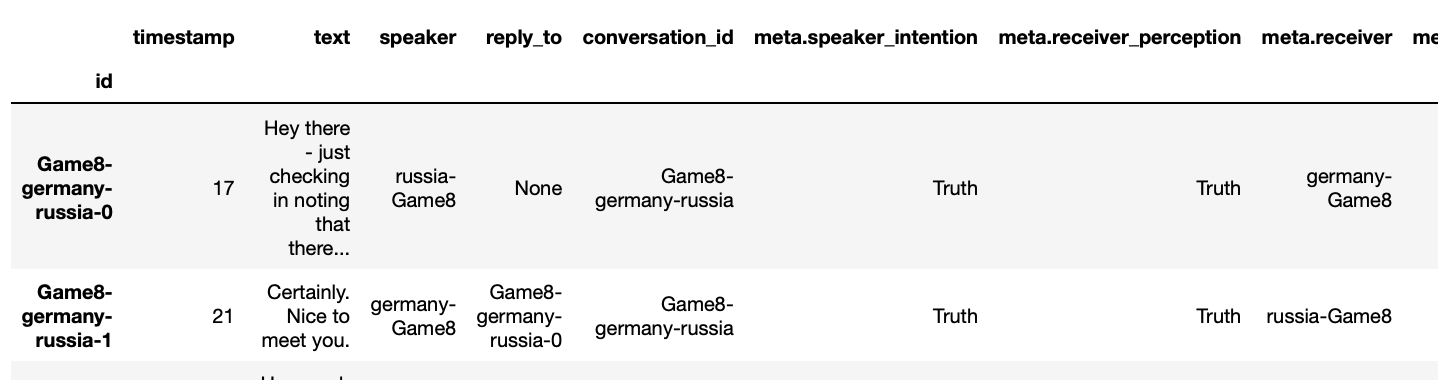
Problem with using the coordination score is that it is an overall coordination score throughout the game. However, within the game, a person may coordinate with someone to work together, then betray the person.

Since Christian’s overall coordination score is made up of some algorithm using the following LIWC values, we might want to compute the values for each utterance of the conversation, and perform a cosine similarity between target person and friends at each utterance. That would show when the scores converge (coordinating) and then diverge (betray and don’t need his help already)



Overall data structure

This data is cool because it is annotated with whether the text is originally a lie or not from both speaker and receiver point of view.



**4 Sept 2020**

How people express affect when they are trying to collaborate with each other

Kokil - Can get the raw data from Christian if we ask

Discord diplomacy:

<https://disboard.org/servers/tag/diplomacy>

<https://disboard.org/server/712324842065625108> => Look at scraping from Discord

<https://github.com/DenisPeskov/2020_acl_diplomacy/tree/master/data>

ACL 2020 paper - <https://sites.google.com/view/qanta/projects/diplomacy>