# Persuasive Strategies across Conversational Contexts Development of a natural language processing tool to detect benign and malicious persuasion in online dialogue

November 2022

Robert Rhys Thomas
University of Bath
rt953@bath.ac.uk

Kate Muir
Bath Spa University
k.muir@bathspa.ac.uk

Ekaterina Kochmar
University of Bath
ek762@bath.ac.uk

Adam Joinson
University of Bath
aj266@bath.ac.uk

Faye Walker
Bath Spa University
f.walker2@bathspa.ac.uk

# Contents

**Abstract**

In this project, we have developed a natural-language processing tool to support the automation of persuasive dialogue detection within large corpora of text. The corpus has been generated using our framework for persuasion strategies which consists of conversational moves which a persuader will use to achieve their overarching persuasive aim. Seven dominant strategies are identified: Rapport and liking, , Negotiating, Appeals to emotion, Appeals to logic, Appeals to authority, Social proof and Overt persuasion, manipulation, and pressure. The model was converted into a codebook, which was tested and refined, before our model was used to conduct a series of studies in which the codebook was marked up across by a team of annotators. We tested whether the model can accurately identify conversations in which attempted persuasion is known to be present, using the Persuasion for Good corpus (Wang et al. 2019), and created a classifier for predicting the probabilities of a given persuasion move being present in any given utterance. We present the initial results of these studies and discuss the implications of using the classifier to identify where both benign and malicious forms of persuasion are taking place in online conversations.

# 1 Introduction

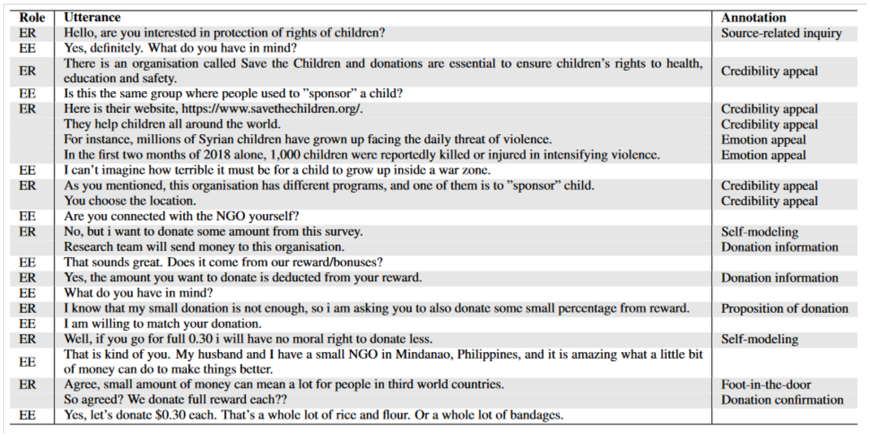
The following report details the processes and decisions made during the Persuasion Strategies project. The main purpose of the project was to develop a universal framework for persuasion that could be applied in any domain as well as the development of a model and dataset for applying the framework. The framework was developed prior to beginning of my involvement however during the course of the project it was further refined and distilled to allow for the training of a machine learning based model. With the framework on hand a dataset was built with the help of PhD students from varying backgrounds, once the data was collected and processed a machine learning based model was trained to be able to classify the persuasion strategies presented in the framework. The trained model is able to predict persuasion strategies and has been shown to be effective when evaluating on the dataset. The process for building the dataset included a number of steps such as training the annotators, measuring agreement and distilling the annotation criteria. A number of metrics and processes were followed to ensure the quality of the data could be as reliable as possible. With the dataset complete numerous models were tested and evaluated to try and get the highest classification accuracy for each tag in the framework. The first step however was the identification of a suitable dataset to apply the framework, the following section details the process for finding the initial data.

# 2 Building The Dataset

Before applying the framework it was necessary to identify a dataset within the domain of persuasion to be a good candidate for experimentation. There are a number of datasets openly available that contain examples of persuasive language. We looked at a number of datasets, notably, Change My View. Change My View dataset comes from a scraping of data from the social media site Reddit. A popular subreddit, which is a specific forum, named /r/changemyview exists to allow users to invite other users to present a counter argument to a users' stance on a particular matter. Users that succesfully persuade the original poster's position on a topic are rewarded with the delta symbol. Others included the Perverted Justice dataset, which consists of transcripts from online communications involving potential groomers and vulnerable users, and a dataset of a small number of transcripts of hostage negotiations from a Naval Academy. After consideration we used the dataset Persuasion For Good [2]. Persuasion For Good aimed to train a machine learning based model to identify any correlation between persuasion strategies and personality type, for example if users were more susceptible to persuasion strategies depending on their personality type.

## 2.1 Persuasion For Good

The researchers of P4Good set up a task online using Amazon's platform Mechanical Turk. The task was performed by users who had signed up to the platform, users are able to earn money for tasks completed. The task involved completing a personality test and then being paired with another user who would be attempting to persuade the user to donate to the charity Save The Children. The resulting model would work in conjunction with a dialogue system to allow an agent to persuade users online to donate to charity, an application that could definitely not be used for anything nefarious.



| Role | Utterance | Annotation |
|---|---|---|
| ER | Hello, are you interested in protection of rights of children? | Source-related inquiry |
| EE | Yes, definitely. What do you have in mind? | |
| ER | There is an organisation called Save the Children and donations are essential to ensure children's rights to health, education and safety. | Credibility appeal |
| EE | Is this the same group where people used to "sponsor" a child? | |
| ER | Here is their website, https://www.savethechildren.org/. | Credibility appeal |
| | They help children all around the world. | Credibility appeal |
| | For instance, millions of Syrian children have grown up facing the daily threat of violence. | Emotion appeal |
| | In the first two months of 2018 alone, 1,000 children were reportedly killed or injured in intensifying violence. | Emotion appeal |
| EE | I can't imagine how terrible it must be for a child to grow up inside a war zone. | |
| ER | As you mentioned, this organisation has different programs, and one of them is to "sponsor" child. | Credibility appeal |
| | You choose the location. | Credibility appeal |
| EE | Are you connected with the NGO yourself? | |
| ER | No, but i want to donate some amount from this survey. | Self-modeling |
| | Research team will send money to this organisation. | Donation information |
| EE | That sounds great. Does it come from our reward/bonuses? | |
| ER | Yes, the amount you want to donate is deducted from your reward. | Donation information |
| EE | What do you have in mind? | |
| ER | I know that my small donation is not enough, so i am asking you to also donate some small percentage from reward. | Proposition of donation |
| EE | I am willing to match your donation. | |
| ER | Well, if you go for full 0.30 i will have no moral right to donate less. | Self-modeling |
| EE | That is kind of you. My husband and I have a small NGO in Mindanao, Philippines, and it is amazing what a little bit of money can do to make things better. | |
| ER | Agree, small amount of money can mean a lot for people in third world countries. | Foot-in-the-door |
| | So agreed? We donate full reward each?? | Donation confirmation |
| EE | Yes, let's donate $0.30 each. That's a whole lot of rice and flour. Or a whole lot of bandages. | |

Figure 2: Transcript collected from the task, ER and EE refer to persuader and persuadee respectively

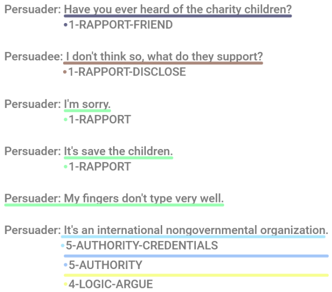
The researchers also included their own persuasion strategies in the corpus. This was also helpful as it allowed us to experiment with different classification techniques whilst building the dataset. The dataset included 300 annotated transcripts and a further 716 without annotations for persuasion strategies. We are now able to markup the transcripts with our own framework for persuasion strategies. To do this we needed a platform for annotation, annotators and a distilled framework.

## 2.2 Initial Data Markup

Once the dataset was decided upon the next steps were to install an app that allowed multiple annotators to annotate the same data. The app had to be open source with the ability to import/export data easily. We looked at a number of apps but settled on Doccano [1].

### 2.2.1 Doccano

Doccano is an open source data annotation app that can support multi-label and multi-class classification. Fr the dataset we needed to be able to annotate utterances in transcripts and be able to access a number of labels as well as the ability to choose more than one. The application was easily hosted on the webapp deployment service Heroku.



Persuader: Have you ever heard of the charity children?
•1-RAPPORT-FRIEND

Persuadee: I don't think so, what do they support?
•1-RAPPORT-DISCLOSE

Persuader: I'm sorry.
•1-RAPPORT

Persuader: It's save the children.
•1-RAPPORT

Persuader: My fingers don't type very well.

Persuader: It's an international nongovernmental organization.
•5-AUTHORITY-CREDENTIALS
•5-AUTHORITY
•4-LOGIC-ARGUE

Figure 3: Example of multiclass and multilabel annotations on Doccano

## 2.3 Small Annotation Task

We uploaded five transcripts to the app and took to annotating the data ourselves with the aim of measure the agreement between the project team. We would also be able to see the level of complexity in the task and identify tags that are most commonly present. The results to this task can be found in the previous project reports. In general we found that a number of tags were much more present than others as well as some disagreement. Based on this we condensed certain tags and totally removed the third layer of subtag. From the results we also agreed that it would be necessary to provide training to the annotators to help them understand the differences between the tags that were most disagreed upon between the project team.

### 2.3.1 Measuring Agreement

To get an understanding of the quality of the annotated data it is important to get a measure of the inter-annotator agreement. The level of agreement can give an indication of the complexity of the task, a low agreement may imply that the framework is difficult to follow, the data is difficult to understand or even potentially that the annotators have not understood the task. There are various metrics that can help calculate the agreement, in this case we have used Cohen's kappa and accuracy.

The kappa score gives is a reliable metric as it an indication of the percentage of agreed examples, taking into account the chance agreement. This to say when a pair of annotators are agreeing purely by chance, the kappa score indicates that user-pairs are agreeing consistently.

$$k = \frac{P_0 - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e}$$

Figure 4: Cohen's kappa coefficient

As well as metric scores we can also observe the confusion matrices. In the small and intermediate task a *"gold standard"* tag is automatically generated. The gold tag is the tag chosen by the majority of annotators, when generating the final dataset this will be used to generate the labels. For each annotator the agreement formulae are applied and a confusion matrix is generated comparing the annotator to the gold standard.

### 2.3.2 Agreement amongst Project Group

The following agreement was observed amongst the project members

| | Full span Full Tag | Full Span Partial Tag | Partial Span Full Tag | Partial Span Partial Tag | Full Span No Tag | Partial Span No Tag |
|---|---|---|---|---|---|---|
| user_1 -> user_2 | 0.588 | 0.69 | 0.699 | 0.735 | 0.575 | 0.602 |
| user_1 -> user_3 | 0.682 | 0.727 | 0.761 | 0.773 | 0.591 | 0.693 |
| user_1 -> user_4 | 0.675 | 0.738 | 0.778 | 0.786 | 0.635 | 0.675 |
| user_2 -> user_3 | 0.605 | 0.693 | 0.702 | 0.763 | 0.57 | 0.667 |
| user_3 -> user_4 | 0.6 | 0.736 | 0.757 | 0.779 | 0.586 | 0.621 |
| user_3 -> user_4 | 0.575 | 0.6 | 0.642 | 0.65 | 0.55 | 0.608 |
| mean | 0.621 | 0.697 | 0.723 | 0.748 | 0.584 | 0.644 |
| standard deviation | 0.042 | 0.048 | 0.047 | 0.047 | 0.026 | 0.035 |

Table 1: Span accuracy across user pairs

The purpose of the above table is to show the agreement amongst annotators in terms of the spans they chose as well as the tags. The metric is the percentage of matching spans over total spans and the columns can be explained as follows:

- Full span full tag: These are the examples whereby the user pairs have chosen the exact span of text and used the same full tag, which is both the top level and bottom
- Full span partial tag: as above however only the top level tag needs to match
- Partial span full tag: The users have annotated overlapping but not fully matching spans but have chosen the full tag
- Partial span partial tag: As above but only the top level tags are identical

- Full span no tag: The users have highlited the same span but do not agree on the tag
- Partial span no tag: users choose overlapping spans of text but don't agree on the tag

The users are showing an agreement into what constitutes persuasive language however the agreement increases when the level of strictness is relaxed. This lead to the decision that all users were required to annotate every span of text which in this case was considered to be a full sentence. The agreement amongst tags can be observed further when considering the kappa scores.

| | Top Level Kappa | Top Level Observed | Full Tag Kappa | Full Tag Observed |
|---|---|---|---|---|
| user_1 -> user_2 | 0.447 | 0.603 | 0.204 | 0.244 |
| user_1 -> user_3 | 0.465 | 0.63 | 0.454 | 0.5 |
| user_1 -> user_4 | 0.476 | 0.607 | 0.404 | 0.459 |
| user_2 -> user_3 | 0.391 | 0.58 | 0.213 | 0.26 |
| user_2 -> user_4 | 0.504 | 0.645 | 0.242 | 0.274 |
| user_3 -> user_4 | 0.366 | 0.556 | 0.354 | 0.417 |

Table 2: Kappa Scores and Observed Agreement amongst Project Group

The table shows relatively low agreement for the full tag which could be in indication of the difficulty of the task, however the kappa score increases when considering only the top-level tag.

### 2.3.3 Distillation of Tags

The persuasion strategy framework contains seven top level tags with each top-level tag containing numerous subtags bringing the total number of available tags for annotators to be 41. We made the decision that it may be necessary to distil the tags to have fewer available to annotators. For a supervised machine learning model to achieve generalisation it needs to be presented with a large amount of labeled data, in a multiclass classification task the model also, for a greater level of accuracy, be presented with a dataset where each label is represented equally. This gets harder to achieve depending on how many tags are present and the difficulty of the task.

### 2.3.4 Changes to Framework

With the agreement results from the small task we made a number of changes to the framework. This included the removal of the lowest level of tag, this aimed to limit the total number of tags available to annotators, if it were possible to have much more data and annotators it could be possible to include them but it would not be possible in this small task. We also decided to add an *OTHER* subta to each of the top level tags. The final tagset is as follows.

| Tags 1-12 | Tags 13-24 | Tags 25-36 | 36-41 |
|---|---|---|---|
| **1-RAPPORT-OTHER** | 2-NEGOTIATE-SCARCITY | 4-LOGIC-CONCEDE | 7-PRESSURE-MANIPULATE |
| 1-RAPPORT-COMPLIMENT | **3-EMOTION-OTHER** | 4-LOGIC-COUNTER | 7-PRESSURE-PRESIST |
| 1-RAPPORT-DISCLOSE | 3-EMOTION-ALTRUISTIC | 4-LOGIC-WITHDRAW | 7-PRESSURE-THREAT |
| 1-RAPPORT-FRIEND | 3-EMOTION-ALTERCASTING | **5-AUTHORITY-OTHER** | 7-PRESSURE-FORCE |
| 1-RAPPORT-SIMILAR | 3-EMOTION-ANGER | 5-AUTHORITY-CREDENTIALS | **8-NO-PERSUASION** |
| **2-NEGOTIATE-OTHER** | 3-EMOTION-EMPATHY | 5-AUTHORITY-SUPERIOR | |
| 2-NEGOTIATE-BARGAIN | 3-EMOTION-GUILT | **6-SOCIAL-OTHER** | |
| 2-NEGOTIATE-OPINION | 3-EMOTION-HUMOUR | 6-SOCIAL-GROUP | |
| 2-NEGOTIATE-RECIP | 3-EMOTION-PROMISE | 6-SOCIAL-NORMS | |
| 2-NEGOTIATE-REMIND | 3-EMOTION-STORY | 6-SOCIAL-PEER | |
| 2-NEGOTIATE-REQUEST | **4-LOGIC-OTHER** | **7-PRESSURE-OTHER** | |
| 2-NEGOTIATE-REWARD | 4-LOGIC-ARGUE | 7-PRESSURE-INTIMIDATE | |

Table 3: Final Tagset

Users taking part in the annotation task are able to login online to a shared space where they can access information about each tag as well as examples, it is also possible in the Doccano app to quickly overlay instructions and examples. With the small annotation task and tagset refined we provided the annotators an intermediate task to complete. This would allow us to see the results of the distillation and training to see if the agreement between annotators improves.

## 2.4 External Annotators

After initial impressions of applying the framework and using the annotation software we then hired PhD students to markup the data. To get them started we gave them the same five transcripts that had already been annotated by the group to see how the results compared.

| User | Transcript 1 | Transcript 2 | Transcript 3 | Transcript 4 | Transcript 5 |
|---|---|---|---|---|---|
| Tina->Emily | 0.5476 | 0.5796 | 0.5841 | 0.3996 | 0.2667 |
| Tina->Hannah | 0.5176 | 0.5836 | 0.6213 | 0.2874 | 0.25 |
| Tina->Jessie | 0.4586 | 0.4257 | 0.6188 | 0.3574 | 0.125 |
| Tina->Pablo | 0.387 | 0.1759 | 0.5521 | 0.2642 | 0.4943 |
| Tina->Rob_P | 0.4586 | 0.4616 | 0.5165 | 0.375 | 0.1619 |
| Tina->Shashank | 0.5486 | 0.5195 | 0.6109 | 0.5025 | 0.1951 |
| Tina->Tobias | 0.4206 | 0.4984 | 0.4192 | 0.3119 | 0.1809 |
| Tina->Jessica | 0.1689 | 0.2105 | 0.1399 | -0.046 | 0.172 |
| Emily->Hannah | 0.6185 | 0.5016 | 0.6294 | 0.7065 | 0.45 |
| Emily->Jessie | 0.5863 | 0.3868 | 0.691 | 0.5755 | 0.5769 |
| Emily->Pablo | 0.4292 | 0.3278 | 0.5303 | 0.2847 | 0.6118 |
| Emily->Rob_P | 0.5847 | 0.496 | 0.5199 | 0.7271 | 0.34 |
| Emily->Shashank | 0.6622 | 0.5325 | 0.8252 | 0.586 | 0.4884 |
| Emily->Tobias | 0.5086 | 0.5 | 0.5268 | 0.508 | 0.1287 |
| Emily->Jessica | 0.2222 | 0.1747 | 0.2029 | 0.0546 | 0.2584 |
| Hannah->Jessie | 0.481 | 0.2182 | 0.6288 | 0.473 | 0.4211 |
| Hannah->Pablo | 0.4179 | 0.4031 | 0.5617 | 0.2136 | 0.443 |
| Hannah->Rob_P | 0.4906 | 0.3126 | 0.6123 | 0.5136 | 0.2979 |
| Hannah->Shashank | 0.5894 | 0.4632 | 0.6907 | 0.6351 | 0.1951 |
| Hannah->Tobias | 0.4445 | 0.5011 | 0.4292 | 0.459 | 0.0388 |
| Hannah->Jessica | 0.1901 | 0.3355 | 0.0595 | -0.0482 | 0.3373 |
| Jessie->Pablo | 0.3347 | 0.1628 | 0.5656 | 0.1638 | 0.4634 |
| Jessie->Rob_P | 0.5128 | 0.421 | 0.58 | 0.4752 | 0.3465 |
| Jessie->Shashank | 0.5313 | 0.4104 | 0.7175 | 0.4845 | 0.2235 |
| Jessie->Tobias | 0.447 | 0.3281 | 0.5209 | 0.4106 | 0.2376 |
| Jessie->Jessica | 0.1449 | 0.1412 | 0.1042 | -0.0604 | 0.1348 |
| Pablo->Rob_P | 0.3543 | 0.219 | 0.4923 | 0.2549 | 0.3333 |
| Pablo->Shashank | 0.4267 | 0.3119 | 0.5914 | 0.3418 | 0.3293 |
| Pablo->Tobias | 0.3776 | 0.4516 | 0.4379 | 0.2741 | 0.12 |
| Pablo->Jessica | 0.0496 | 0.0851 | 0.092 | -0.2272 | 0.125 |
| Rob_P->Shashank | 0.4562 | 0.4386 | 0.5766 | 0.4405 | 0.01 |
| Rob_P->Tobias | 0.4827 | 0.4638 | 0.4563 | 0.4463 | 0.3832 |
| Rob_P->Jessica | 0.182 | 0.074 | 0.1826 | 0.0937 | -0.0421 |
| Shashank->Tobias | 0.4894 | 0.4086 | 0.5479 | 0.5063 | 0.1809 |
| Shashank->Jessica | 0.1955 | 0.1715 | 0.1906 | 0.0027 | 0.2326 |
| Tobias->Jessica | 0.2215 | 0.3779 | 0.0206 | 0.0288 | 0.0294 |
| Mean | 0.415 | 0.3631 | 0.4735 | 0.3271 | 0.265 |

Table 4: Pairwise Agreement Kappa Scores

The above table shows the pair-wise agreement between annotators, as well can see there is better agreement amongst some annotators and others as well as one transcripts that all annotators struggle to agree on.

### 2.4.1 Annotator Training

After analysing the annotations from the first experiment the annotators were given a brief training session. The purpose was to identify common disagreements between annotators and try to gain an understanding whether the disagreement was due to mistakes, differences in opinion or ambiguity in the framework. For each top-level annotation a brief number of examples were presented for a discussion with the annotators [1].

The training session showed that there were conversations to be had about the definitions of certain tags and that some of the language in the transcripts did belong to multiple labels. Furthermore the task did show that some annotators did not fully know how to use the app and that the labels provided were somewhat confusing. Each user was also given a report that compared their annotations to the *gold standard* so that they could get an idea of how they compared to the majority of the taggers.

| User | Transcript 1 | Transcript 2 | Transcript 3 | Transcript 4 | Transcript 5 | Average |
|---|---|---|---|---|---|---|
| Tina | 0.737749151 | 0.838771593 | 0.75 | 0.612154696 | 0.4 | 0.667735088 |
| Emily | 0.9 | 0.788279773 | 0.917525773 | 0.957974138 | 1 | 0.912755937 |
| Hannah | 0.822267447 | 0.793357934 | 0.793814433 | 0.831896552 | 1 | 0.848267273 |
| Jessie | 0.71088769 | 0.532554257 | 0.875862069 | 0.652818991 | 0.75 | 0.704424601 |
| Pablo | 0.558196924 | 0.458699473 | 0.673469388 | 0.342261905 | 0.78125 | 0.562775538 |
| Rob_P | 0.68438805 | 0.586715867 | 0.690322581 | 0.76122449 | 0.4 | 0.624530198 |
| Shashank | 0.831083985 | 0.797101449 | 0.957597173 | 0.78987069 | 0.588235294 | 0.792777718 |
| Tobias | 0.671074657 | 0.723320158 | 0.689320388 | 0.658269441 | 0.054054054 | 0.55920774 |
| Jessica | 0.347923103 | 0.422138837 | 0.255172414 | 0.055363322 | 0.6 | 0.336119535 |

Table 5: User vs Aggregate Kappa Scores

## 2.5 Intermediate Data Markup

After analysing the results from the small task the annotators were then provided the refined tagset as well as further instructions. They were as follows:

[1] https://docs.google.com/presentation/d/1p1wP9EfV3gapPLEMedAyOGibfnNEXJE0qqdzaj1HCVI/edit?usp=sharing

- Annotate every sentence
- If no Persuasion is present provide *8-NO-ANNOTATION* tag
- If appropriate use more than one tag
- Annotate only the Persuader's utterances

### 2.5.1 Intermediate Agreement Results

After the training meeting the annotators were given a further 20 transcripts. The aim of this test was to see if the annotators agreement had improved as well as to see if their interpretation of the framework was slightly more in sync. The table 6 below shows the kappa scores between annotators averaged over a selection of the transcripts with the average in the final column.

Transcript 18 caused a lot of disagreement which has brought the score down for the final four transcripts. Considering the data points have been expanded the lowered kappa scores are to be expected. It will be possible to test again after we have more annotated data.

| User | Transcript 1-5 | Transcript 6-10 | Transcript 11-15 | Transcript 16-20 | Average |
|---|---|---|---|---|---|
| Tina | 0.417297 | 0.163952 | 0.586983 | 0.360095705 | 0.382081687 |
| Emily | 0.70086 | 0.675172 | 0.61899 | 0.493291065 | 0.622078296 |
| Hannah | 0.537582 | 0.514502 | 0.509816 | 0.439970113 | 0.50046762 |
| Jessie | 0.729113 | 0.528785 | 0.572547 | 0.526860041 | 0.58932605 |
| Pablo | 0.517722 | 0.45508 | 0.551383 | 0.31849077 | 0.460669038 |
| Rob_P | 0.649483 | 0.630101 | 0.684888 | 0.545271778 | 0.62743572 |
| Shashank | 0.308761 | 0.386922 | 0.50424 | 0.203403112 | 0.35083146 |
| Tobias | 0.611906 | 0.473552 | 0.589186 | 0.411435417 | 0.521519737 |
| Jessica | 0.501304 | 0.580601 | 0.493023 | 0.487420105 | 0.515587057 |

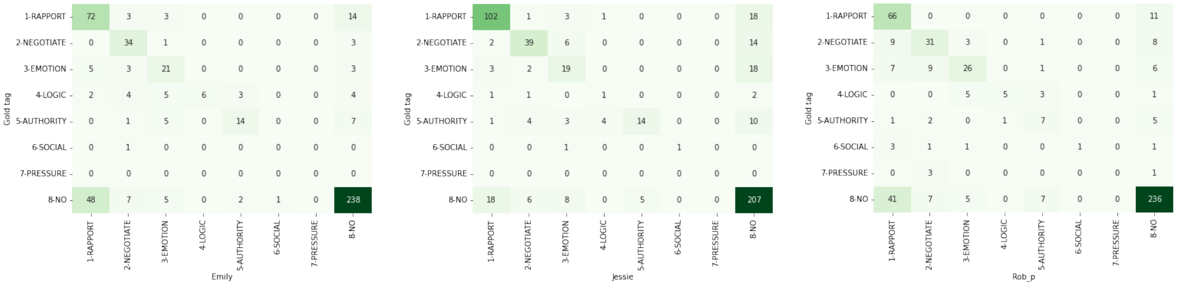
Table 6: User vs Aggregate Tag Intermediate Markup

Again we are comparing each annotator to the majority tag which is considered the *gold standard*. We can see in general users are adequately in agreement, a kappa score of 0.6 and above is reliable. Furthermore we can look at the confusion matrices that compare each user against the majority tag.

### 2.5.2 Tag Confusion Matrices

The below confusion matrices 5 shows examples from the three users who were in most agreement with the majority and the three users who had the lowest agreement.

One of the largest disagreements appears to be between *1-RAPPORT* and *NO-PERSUASION*. This is understandable as questions were raised during the training on how to identify friendly and sociable language with a concerted attempt at building rapport with the intent to persuade. Overall most of the disagreement is when the majority has chosen one tag and the others have used no persuasion. This could possibly be due to the code that adds *NO PERSUASION* to those that haven't annotated a span.



| Gold tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|
| 1-RAPPORT | 72 | 3 | 3 | 0 | 0 | 0 | 0 | 14 |
| 2-NEGOTIATE | 0 | 34 | 1 | 0 | 0 | 0 | 0 | 3 |
| 3-EMOTION | 5 | 3 | 21 | 0 | 0 | 0 | 0 | 3 |
| 4-LOGIC | 2 | 4 | 5 | 6 | 3 | 0 | 0 | 4 |
| 5-AUTHORITY | 0 | 1 | 5 | 0 | 14 | 0 | 0 | 7 |
| 6-SOCIAL | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7-PRESSURE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8-NO | 48 | 7 | 5 | 0 | 2 | 1 | 0 | 238 |

| Gold tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|
| 1-RAPPORT | 102 | 1 | 3 | 1 | 0 | 0 | 0 | 18 |
| 2-NEGOTIATE | 2 | 39 | 6 | 0 | 0 | 0 | 0 | 14 |
| 3-EMOTION | 3 | 2 | 19 | 0 | 0 | 0 | 0 | 18 |
| 4-LOGIC | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 2 |
| 5-AUTHORITY | 1 | 4 | 3 | 4 | 14 | 0 | 0 | 10 |
| 6-SOCIAL | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 7-PRESSURE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8-NO | 18 | 6 | 8 | 0 | 5 | 0 | 0 | 207 |

| Gold tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|
| 1-RAPPORT | 66 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 2-NEGOTIATE | 9 | 31 | 3 | 0 | 1 | 0 | 0 | 8 |
| 3-EMOTION | 7 | 9 | 26 | 0 | 1 | 0 | 0 | 6 |
| 4-LOGIC | 0 | 0 | 5 | 5 | 3 | 0 | 0 | 1 |
| 5-AUTHORITY | 1 | 2 | 0 | 1 | 7 | 0 | 0 | 5 |
| 6-SOCIAL | 3 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| 7-PRESSURE | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 |
| 8-NO | 41 | 7 | 5 | 0 | 7 | 0 | 0 | 236 |

Figure 5: Users in most agreement with majority

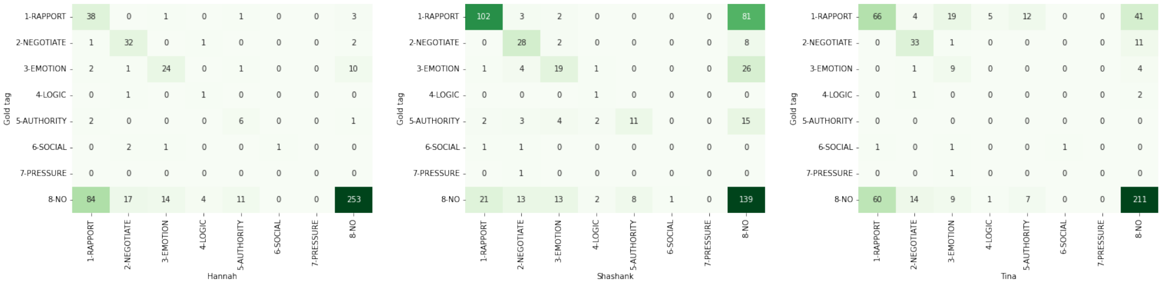


Figure 6: Users in least agreement with majority

## 2.6 Examples

Below are a number of examples for each top-level tag to give an idea of disagreements. Each table contains the utterances along with the *gold tag* with each remaining column being a top label tag. The number in each cell represents how many annotators chose that tag for the utterances. As annotators were advised they can choose more than one tag not all rows add up to nine which is the total number of annotators.

### 2.6.1 Rapport

Rapport has been the largest represented tag out of the framework. This could largely be due to the fact that the task that the persuader is performing relies on their ability to form a trusting relationship with the persuadee. The language that has been tagged as "RAPPORT" has mostly been friendly in nature, when building rapport the persuader is seemingly trying to relate to the persaudee. There is some overlap with negotiation when a question is involved. There is a tendency for annotators to tag as negotiation if there is a question, due to this we allowed annotators to use multiple tag for the same span of text.

| Text | Gold Tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|---|
| I'll do the same. | 1-RAPPORT | 5 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| That's so generous! It's easy to donate when you feel attached to the situation, but helping someone you don't know takes a really special kind of person | 1-RAPPORT | 9 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| tell me about the good ones? | 1-RAPPORT | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| do you ever donated to charities? | 1-RAPPORT | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| Good! I am well. I have a really important question to ask of you | 1-RAPPORT | 6 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| I know I was surprised by those numbers. | 1-RAPPORT | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 4 |

### 2.6.2 Negotiate

The features of negotiation that the annotators are picking up on tend to be phrases that include questions. There tends to be explicit requests for the donations and bargaining. Negotiation has overlapped with Rapport, as well can see below the confidence is not as strong, we can see that other users have chosen a number of other tags.

| Text | Gold Tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|---|
| Just to make sure, you would like to donate $0.10 US? Is that right? | 2-NEGOTIATE | 1 | 5 | 0 | 0 | 0 | 0 | 0 | 3 |
| hats okay, a donation as little as 77 cents can house a refugee family for a day. | 2-NEGOTIATE | 0 | 5 | 3 | 2 | 0 | 0 | 0 | 0 |
| Then definitely donate the full two. | 2-NEGOTIATE | 0 | 5 | 1 | 1 | 0 | 0 | 0 | 2 |
| How much do you like to donate to the charity now? Your donation will be directly deducted from your task payment. You can choose any amount from $0 to all your payment ($2). | 2-NEGOTIATE | 0 | 5 | 0 | 1 | 0 | 0 | 0 | 3 |
| I can guess, but isn't there any amount you could donate like a little bit of money that you use for a snack or candy? Even a little amount helps. | 2-NEGOTIATE | 0 | 5 | 4 | 1 | 0 | 0 | 0 | 0 |
| I was hoping you would be interested in donating a small amount of .10 to help a child that will never have the same opportunities as us? | 2-NEGOTIATE | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 3 |

### 2.6.3 Emotion

The most common form of emotional persuasion tends to be guilt based. The examples below do show crossover in other labels but the common theme tends to be trying to create empathy for children. Negotiation is a common overlap, this could be that the persuader is trying to suggest even a small amount of money will help but also referring to the positive emotional response one gets from being charitable. The annotators are able to use more than one label and that is reflected in the results, if there are more than 9 labels this means that one or more user has used more than one label.

| Text | Gold Tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|---|
| Even a few cents would help provide food or clean water for a child | 3-EMOTION | 0 | 4 | 7 | 0 | 0 | 0 | 0 | 1 |
| Just a small amount, pennies, would make a huge difference to these kids | 3-EMOTION | 0 | 3 | 5 | 1 | 0 | 0 | 0 | 1 |
| And your small donation can go a long way, something as small as the price of a candy bar can be the difference in these kids lives | 3-EMOTION | 0 | 3 | 5 | 0 | 0 | 0 | 0 | 3 |
| Thank you, that would be great. Even with our little helps, the children around the world may have a better chance | 3-EMOTION | 2 | 2 | 6 | 1 | 0 | 0 | 0 | 1 |
| Yeah. It had a link to their website so I'm poking around. It sounds like it makes sure that somebody is sticking up for kids and helps them out where it can. It must be simultaneously really hard and really fulfilling to work for them | 3-EMOTION | 2 | 2 | 3 | 0 | 0 | 0 | 0 | 2 |
| I think about the amount of money I spend on buying cup of coffee these days, I could definitely go without one in order to help some kids! | 3-EMOTION | 1 | 2 | 5 | 2 | 0 | 0 | 0 | 1 |

### 2.6.4 Logic

There is not as much confidence in the *LOGIC* tag as there is with the first three however there is certainty that *SOCIAL* and *PRESSURE* are not present. Logical persuasion tends to be present when the persuader is using logical statements which in this case are factual information about the charity which can also be assosiated with *5-AUTHORITY*. It is this overlap between logical statements, information about the authority figure and the emotive language used to express the information that appears to be causing a disagreement.

| Text | Gold Tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|---|
| The research team conducting this conversation task will collect all donations and send them to Save the Children, so you don't need a credit card. | 4-LOGIC | 2 | 1 | 0 | 3 | 1 | 0 | 0 | 3 |
| I am hear to talk to you about an amazing charity called Save the Children, that uses donated money to help provide essentials of life to children in developing countries. | 4-LOGIC | 1 | 1 | 1 | 3 | 1 | 0 | 0 | 2 |
| In 2017 86% of all fund raise went directly to the children in need, and of the 86%, 34% was directly put into providing food and nutrition. | 4-LOGIC | 1 | 0 | 0 | 4 | 3 | 0 | 0 | 1 |
| Education, Emergency Relief, HIV/Aids treatment, and helping the children stay in a safe environment | 4-LOGIC | 1 | 0 | 2 | 4 | 1 | 0 | 0 | 2 |
| 2019 will be 100 years, I believe. | 4-LOGIC | 1 | 0 | 0 | 3 | 3 | 0 | 0 | 2 |
| yes but children are the future. | 4-LOGIC | 0 | 0 | 1 | 7 | 0 | 0 | 0 | 1 |

### 2.6.5 Authority

Utterances that received the majority tag of *5-AUTHORITY* tend to references to the charity Save The Children's legitimacy and trustworthiness. Annotators generally recognise these statements but there is some overlap with *1-RAPPORT* this could possibly still be due to the conversational nature of the utterances, the annotators have been suggested to be aware of overusing *1-RAPPORT*. There are few annotators choosing *8-NO-PERSUASION* this could possibly be because they deem the utterance to be too conversational.

| Text | Gold Tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|---|
| Please refer the link - URL s/about-us/awards-and-rankings and you will find its a legitimate organisation | 5-AUTHORITY | 1 | 0 | 0 | 0 | 7 | 0 | 0 | 1 |
| And let me assure you, this is a professional and trustworthy fund. | 5-AUTHORITY | 1 | 1 | 0 | 0 | 5 | 0 | 0 | 4 |
| Save the Children is an international non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries. | 5-AUTHORITY | 1 | 0 | 0 | 1 | 5 | 0 | 0 | 2 |
| Yes, it\'s a great cause. And the charity is highly rated with many positive rewards. Including an "A" rating from the American Institute of Philanthropy. | 5-AUTHORITY | 1 | 0 | 0 | 1 | 5 | 0 | 0 | 2 |
| Sure! First it's a reputable organization and trustworthy. Independent charity watchdog groups give them high ratings. | 5-AUTHORITY | 1 | 0 | 0 | 0 | 5 | 0 | 0 | 3 |
| They're one of the more reputable charities luckily. | 5-AUTHORITY | 0 | 1 | 0 | 0 | 4 | 0 | 0 | 4 |

There were not many instances of *6-SOCIAL* to draw conclusions from, we can also see that the annotators were not confident that this was the majority tag. There was one example where *6-SOCIAL* had a soft majority but it was in equal standing to *8-NO-PERSUASION*.

### 2.6.6 Social

| Text | Gold Tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|---|
| I imagine a lot of people end up doing these hits. It probably adds up, assuming everyone pitches in a little | 6-SOCIAL | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 5 |
| Maybe we can each encourage someone else in our lives to make a donation, and spread the impact. | 2-NEGOTIATE | 1 | 4 | 1 | 0 | 0 | 3 | 0 | 1 |
| I'm going to doate again. | 2-NEGOTIATE | 2 | 3 | 1 | 0 | 0 | 3 | 0 | 1 |
| You are part of the solution and we all have a moral responsibility to help the children around the world. | | | | | | | | | |
| I hope you would reconsider. | 3-EMOTION | 1 | 0 | 4 | 0 | 0 | 3 | 1 | 1 |
| It is! Imagine if we convinced our friends and family to donate a small amount as well, people like you and | | | | | | | | | |
| I could actually make a difference in the world. | 3-EMOTION | 3 | 1 | 4 | 0 | 0 | 2 | 0 | 1 |
| I'm giving twenty cents. | 8-NO | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 6 |

### 2.6.7 Pressure

There were very few examples of "PRESSURE" in the annotations. Though the examples below do show signs that the persuader is trying to pressure the persuadee. The issue may be that pressure and emotional guilt can overlap somewhat, it is quite hard to identify threatening or malicious pressure being used to persuade the user.

| Text | Gold Tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|---|
| But for this to work, I really need you to donate both dollars. | 2-NEGOTIATE | 0 | 6 | 0 | 0 | 0 | 0 | 2 | 1 |
| You are part of the solution and we all have a moral responsibility to help the children around the world. I hope you would reconsider. | 3-EMOTION | 1 | 0 | 4 | 0 | 0 | 3 | 1 | 1 |
| Hello! I need you to make a donation for us to work this out. | 2-NEGOTIATE | 1 | 8 | 0 | 0 | 0 | 0 | 1 | 0 |
| We have to make sure we get a bonus by donating the $2. | 2-NEGOTIATE | 0 | 7 | 0 | 0 | 0 | 0 | 1 | 2 |
| It would be great if you can contribute right now | 8-NO | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 7 |
| Thanks! Then definitely donate the full two. You'll be really happy you did, trust me | 8-NO | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 7 |
| why is that? | 8-NO | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 5 |

### 2.6.8 No Persuasion

The training session found that a lot of annotators were tagging phatic communication as rapport building. For example, greetings and thanking. The annotators were given advice that functional language like greetings could not necessarily be considered rapport based persuasion and to consider the context when choosing the tag. Consequently there has been a greater agreement in where to use "NO-PERSUASION". We have also added a case whereby an annotator hasn't

| Text | Gold Tag | 1-RAPPORT | 2-NEGOTIATE | 3-EMOTION | 4-LOGIC | 5-AUTHORITY | 6-SOCIAL | 7-PRESSURE | 8-NO |
|---|---|---|---|---|---|---|---|---|---|
| It is! | 8-NO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| hello, | 8-NO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| yes i saw it | 8-NO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| Hello! | 8-NO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| he charity is called Save the Children. | 8-NO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| I think so, it had been 10 turns. | 8-NO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| Hello is someone there? | 8-NO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| Hello! | 8-NO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |

### 2.6.9 Training Conclusion

The experiments showed that it was necessary to provide annotators with feedback to ensure that they all had a consistent understanding of the framework. The task has proven to be relatively difficult due to the number of tags and the nature of the dataset as teh transcripts are domain specific and contain largely conversational utterances. The purpose of annotator training and monitoring is to ensure consistency during the main markup of the remaining transcripts. The annotators responded well to the feedback and we were able to gave interesting discussions regarding the framework and how to define each persuasion strategy. It may be necessary to provide further training to annotators before commencing annotating with more focus on defining persuasion strategies and their differences.

## 2.7 Main Markup

The remaining transcripts were split into 18 batches, each batch was assigned two annotators. The main markup process took roughly two months with the majority of the annotations complete in five weeks. The annotators required little support during the mark up phase and reported no issues with the task. We had to upgrade the server storage for the webapp however due to the amount of data being stored in the database. As we could only assign two annotators to each batch it was not possible to have to generate a *golden* tag to be able to break any disagreement. We shall be looking at the final tag distribution amongst the annotators to get an idea the balance of the data.

| | User_1 | User_2 | User_3 | User_4 | User_5 | User_6 | User_7 | User_8 | User_9 |
|---|---|---|---|---|---|---|---|---|---|
| 1-RAPPORT-COMPLIMENT | 141 | 168 | 71 | 299 | 149 | 377 | 124 | 94 | 291 |
| 1-RAPPORT-DISCLOSE | 313 | 180 | 341 | 187 | 170 | 366 | 192 | 1787 | 313 |
| 1-RAPPORT-FRIEND | 741 | 219 | 577 | 1178 | 186 | 455 | 174 | 15 | 526 |
| 1-RAPPORT-OTHER | 239 | 175 | 288 | 78 | 145 | 1059 | 477 | 1087 | 335 |
| 1-RAPPORT-SIMILAR | 46 | 79 | 14 | 60 | 155 | 66 | 85 | 168 | 164 |
| 2-NEGOTIATE-BARGAIN | 9 | 121 | 39 | 39 | 349 | 25 | 211 | 210 | 286 |
| 2-NEGOTIATE-OPINION | 28 | 33 | 68 | 132 | 239 | 174 | 91 | 58 | 2 |
| 2-NEGOTIATE-OTHER | 66 | 155 | 242 | 187 | 36 | 39 | 3 | 138 | 30 |
| 2-NEGOTIATE-RECIP | 1 | 97 | 51 | 32 | 1 | 52 | 85 | 0 | 49 |
| 2-NEGOTIATE-REMIND | 13 | 132 | 24 | 24 | 7 | 13 | 51 | 64 | 80 |
| 2-NEGOTIATE-REQUEST | 293 | 311 | 254 | 48 | 178 | 269 | 200 | 261 | 208 |
| 2-NEGOTIATE-REWARD | 0 | 11 | 50 | 20 | 2 | 5 | 30 | 12 | 233 |
| 2-NEGOTIATE-SCARCITY | 20 | 1 | 0 | 10 | 18 | 0 | 6 | 0 | 17 |
| 3-EMOTION-ALTERCASTING | 0 | 13 | 12 | 5 | 50 | 71 | 61 | 0 | 1 |
| 3-EMOTION-ALTRUISTIC | 98 | 292 | 148 | 321 | 78 | 680 | 72 | 0 | 3 |
| 3-EMOTION-ANGER | 8 | 8 | 5 | 1 | 5 | 11 | 3 | 4 | 1 |
| 3-EMOTION-EMPATHY | 95 | 189 | 49 | 170 | 62 | 202 | 137 | 83 | 226 |
| 3-EMOTION-GUILT | 59 | 62 | 85 | 46 | 434 | 64 | 82 | 46 | 90 |
| 3-EMOTION-HUMOUR | 12 | 34 | 14 | 0 | 23 | 17 | 9 | 0 | 3 |
| 3-EMOTION-OTHER | 149 | 63 | 26 | 10 | 9 | 57 | 7 | 7 | 6 |
| 3-EMOTION-PROMISE | 50 | 29 | 23 | 13 | 21 | 18 | 45 | 24 | 0 |
| 3-EMOTION-STORY | 120 | 70 | 36 | 14 | 119 | 44 | 72 | 100 | 140 |
| 4-LOGIC-ARGUE | 3 | 186 | 286 | 38 | 24 | 104 | 38 | 28 | 78 |
| 4-LOGIC-CONCEDE | 18 | 60 | 197 | 1 | 17 | 229 | 32 | 1 | 233 |
| 4-LOGIC-COUNTER | 10 | 47 | 50 | 13 | 62 | 19 | 20 | 0 | 59 |
| 4-LOGIC-OTHER | 743 | 257 | 5 | 35 | 43 | 70 | 689 | 35 | 92 |
| 4-LOGIC-WITHDRAW | 0 | 1 | 0 | 1 | 2 | 0 | 4 | 0 | 0 |
| 5-AUTHORITY-CREDENTIALS | 247 | 187 | 112 | 15 | 351 | 12 | 378 | 0 | 200 |
| 5-AUTHORITY-OTHER | 3 | 11 | 10 | 188 | 1 | 28 | 1 | 1 | 161 |
| 5-AUTHORITY-SUPERIOR | 79 | 0 | 106 | 0 | 8 | 14 | | 2 | 209 |
| 6-SOCIAL-GROUP | 4 | 0 | 0 | 2 | 8 | 6 | 1 | 0 | 27 |
| 6-SOCIAL-NORMS | 0 | 12 | 0 | 4 | 81 | 9 | 14 | 1 | 1 |
| 6-SOCIAL-OTHER | 1 | 14 | 3 | 9 | 0 | 27 | 69 | 0 | 0 |
| 6-SOCIAL-PEER | 1 | 1 | 6 | 6 | 8 | 1 | 2 | 1 | 2 |
| 7-PRESSURE-FORCE | 0 | 2 | 0 | 3 | 8 | 10 | 0 | 0 | 0 |
| 7-PRESSURE-INTIMIDATE | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 12 | 0 |
| 7-PRESSURE-MANIPULATE | 0 | 2 | 5 | 3 | 17 | 0 | 0 | 7 | 0 |
| 7-PRESSURE-OTHER | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 7-PRESSURE-PRESIST | 0 | 4 | 2 | 1 | 1 | 7 | 1 | 20 | 2 |
| 7-PRESSURE-THREAT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 7-PRESSURE-OTHER | 1 | 0 | 1 | 6 | 5 | 0 | 0 | 6 | 0 |
| 8-NO-PERSUASION | 214 | 1247 | 36 | 251 | 0 | 8 | 436 | 232 | 219 |
| Total | 3825 | 4473 | 3238 | 3450 | 3072 | 4608 | 3903 | 4505 | 4288 |

Table 7: Full Tag Distribution

Table 7 shows that there is a large amount of tags concentrated in the first three top level tags. Based on this, for the classification task we shall only use the top level tags as this will give us the best chance of achieving decent adequate classification results, the data would be far too sparse if we used the bottom level tags as well. There is hardly any presence of *7-PRESSURE* however this is expected as the purpose of the task to persuade users to donate to charity, the subtags of *7-PRESSURE* include tactics such as manipulation, threatening and intimidation. Were a user to use these malicious methods of persuasion it is very likely they would be rejected from the study. We are perhaps getting an understanding here that the domain of the transcripts is too specified towards benign persuasion, to be able to model more malicious persuasion we would need wider data sources that encapsulate the malicious persuasion.

| | Emily | Hannah | Pablo | Shashank | Rob_P | Tobias | Jessica | Tina | Jessie |
|---|---|---|---|---|---|---|---|---|---|
| 1-RAPPORT | 1480 | 821 | 1291 | 1802 | 805 | 2323 | 1052 | 3151 | 1629 |
| 2-NEGOTIATE | 430 | 861 | 728 | 492 | 830 | 577 | 677 | 743 | 905 |
| 3-EMOTION | 591 | 760 | 398 | 580 | 801 | 1164 | 488 | 264 | 470 |
| 4-LOGIC | 774 | 551 | 538 | 88 | 148 | 422 | 783 | 64 | 462 |
| 5-AUTHORITY | 329 | 198 | 228 | 203 | 360 | 54 | 379 | 3 | 570 |
| 6-SOCIAL | 6 | 27 | 9 | 21 | 97 | 43 | 86 | 2 | 30 |
| 7-PRESSURE | 1 | 8 | 10 | 13 | 31 | 17 | 2 | 46 | 3 |
| 8-NO-PERSUASION | 214 | 1247 | 36 | 251 | 0 | 8 | 436 | 232 | 219 |

Table 8: Top Level Tag Distribution

The distribution in the above table8 again shows an imbalance weighted towards tags 1-4 with little to no representation of *6-SOCIAL* and *7-PRESSURE*. With a data imbalance like this is is in issue down the line when training a multiclass classifier.

### 2.7.1 Agreement Results

As there were two users per batch we can only observe the agreement between the annotators assigned to each batch. It is not possible to compare the individual users to the *gold standard*. Future work may involve inviting a third annotator to break the deadlock. We can however generate an aggregated agreement metric over the entire dataset. To calculate this each users' tags were converted into multihot vectors. The kappa and accuracy scores were generated between each user's vector and added to a running tally, the final value was averaged.

- kappa : 0.3689
- observed agreement : 0.8699

Whilst a score below 0.6 is not completely adequate we have managed to collect a large amount of data whereby the users agree on the persuasion strategies. This can still be used to train the model. The observed agreement is quite high as the calculation was made on multi-hot vectors, there was a lot of agreement into which persuasion strategies were *not* present and this has helped push the calculation up. The following table also goes into detail into pair-wise agreement.

| | User_1 | User_2 | User_3 | User_4 | User_5 | User_6 | User_7 | User_8 | User_9 |
|---|---|---|---|---|---|---|---|---|---|
| User_1 | | 0.39 | 0.48 | 0.49 | | | | 0.47 | |
| User_2 | 0.87 | | | 0.34 | 0.31 | | | 0.33 | |
| User_3 | 0.89 | | | | | | 0.35 | 0.22 | 0.25 |
| User_4 | 0.89 | 0.86 | | | 0.44 | 0.33 | | | |
| User_5 | | 0.85 | | 0.88 | | 0.31 | 0.3 | | |
| User_6 | | | | 0.86 | 0.85 | | 0.28 | | 0.33 |
| User_7 | | | 0.87 | | 0.85 | 0.85 | | | 0.48 |
| User_8 | 0.89 | 0.86 | 0.84 | | | | | | 0.46 |
| User_9 | | | 0.83 | | | 0.85 | 0.88 | 0.88 | |

Table 9: Kappa and Observed Agreement Between User Pairs

The lower left corner of the table contains the accuracy scores across each batch and the top right shows the kappa scores. We can see that some users are in better agreement than others with the lowest kappa score being 0.22 between users 8 and 3. This can be an indication of which annotators are better suited to continue if more data comes available as well as which data may be disregarded due to low agreement and therefore low quality.

### 2.7.2 Full Tag Confusion Matrix

Finally a confusion matrix 7 can show which tags are confused and again we can see it is the first five tags that have the most presence and the disagreement is most visible. The *NO-TAG* is a placeholder tag whereby a user has chosen a multitag and the other has not. It can be disregarded in this graph.

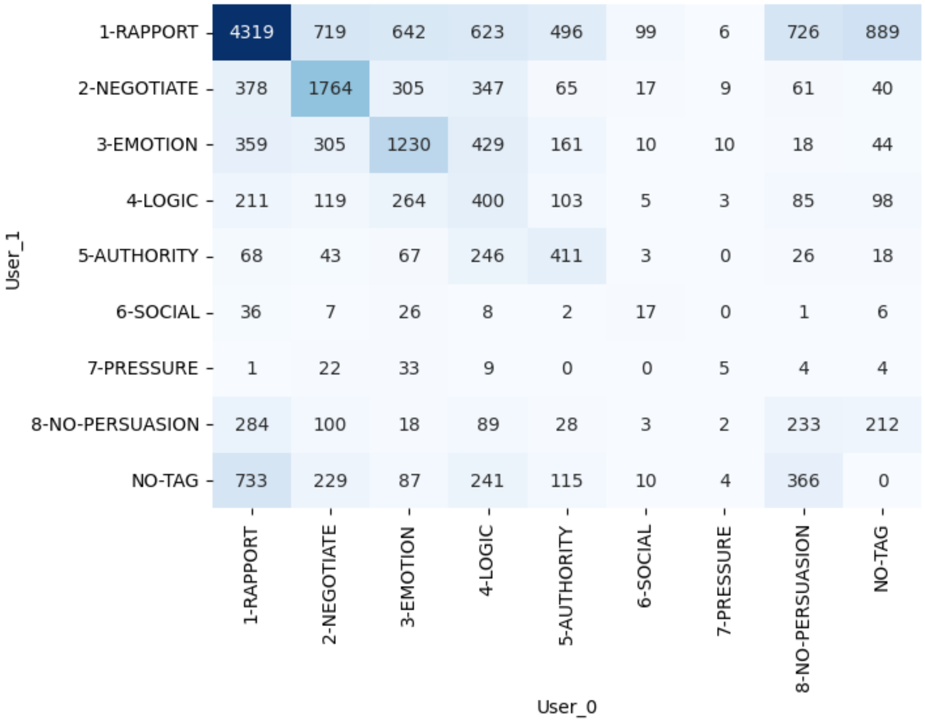


Figure 7: Random User vs User Confusion Matrix

### 2.8 Persuasion Strategy Annotation Conclusion

The annotation process is is largely the most labour and cost intensive part of the project. The annotators did not always keep to deadlines and the agreement has not been as adequately high. However it was important to be able to get feedback from the annotators and learn about their impressions of the data and the persuasion framework. The inter-annotator agreement is slightly less than adequate. The kappa score, to be of high confidence and quality, is recommended to be over 0.8, sadly in this case we managed a max of 0.49 and a low of 0.22. To improve the inter-annotator agreement it may be necessary to give further training to the annotators as well as provide them with further examples. Now that the data has been collected it must now be processed into an exportable format.

## 3 Persuasion Strategies Corpus

Once the process of collecting the data was complete the annotations were processed into a format that can be easily imported. There were a number of considerations when exporting the data, such as:

- How do we represent the utterances?
- How do we represent the labels
- What format do we choose?
- Where do we store the data?
- Split Training/Testing

The format of the data will be a jsonlist file. This is the same format that Doccano exports to and is easily interpretable in a text editor as well as with Python. A jsonlist is a newline separated file with each line containing a json formatted string containing each utterance in the dataset as well as meta data so that the utterance can be viewed in context with the entire dataset. The below figures show how each of the training and testing jsons are formatted for import.

```

        {
          'transcript_idx': '113',
          'line_idx'      : 0,
          'user'          : 1,
          'line'          : 'Persuader: Hi.  HOw are you',
          'text'          : 'Hi.',
          'multilabel'    : [ 0,  0,  0,  0,   0, 0,   0, 2,  0],
          'gold_tag'      : '8-NO-PERSUASION',
          'confidence'    : 1.0,
          'user_0'        : '8-NO-PERSUASION',
          'user_1'        : '8-NO-PERSUASION'
        }

```

Figure 8: Json representation of training data

As the training jsons contain the annotations of two annotators there are only two user tags. We have also provided only the top level tag as the distribution of bottom level tags were too few to train a comprehensive model.

```

        {
          'transcript_idx': 1,
          'line_idx'      : 18,
          'user'          : 1,
          'line'          : 'Persuader: Do you currently donate to
                             your charity?',
          'text'          : 'Do you currently donate to your charity?'
  ,
          'multilabel'    : [6, 0, 1, 0, 0, 1, 0, 0, 5],
          'gold_tag'      : '1-RAPPORT',
          'confidence'    : 0.46,
          'user_0'        : 'NO-TAG',
          'user_1'        : '1-RAPPORT',
          'user_2'        : '1-RAPPORT',
          'user_3'        : '1-RAPPORT',
          'user_4'        : '1-RAPPORT',
          'user_5'        : '6-SOCIAL',
          'user_6'        : 'NO-TAG',
          'user_7'        : '1-RAPPORT',
          'user_8'        : 'NO-TAG',
          'user_9'        : '3-EMOTION',
          'user_10'       : '1-RAPPORT',
          'user_11'       : 'NO-TAG',
          'user_12'       : 'NO-TAG'
        }

```

Figure 9: Json representation of testing data

The testing data comes from the initial and intermediate data markup tasks. As there are much more annotators present for the tagging of this data the majority tag is more discernable. The multilabel tags allow for further experimentation. The following sections describe the metadata that is present in the jsons.

## 3.1 Transcripts

The raw exports from the annotation app come in the form of Jsonlist files, which are essentially jsons separated with a newline. These can be easily processed in Python into different formats. Each line contains the annotations of an entire transcript, containing both the persuader and persuadee's utterances. The transcripts come in the form of a string with each line being separated by a newline special character. The jsonlist file contains every utterance in the dataset, the transcripts can be identified by the unique ID number and line number, they can be put back together by a simple search query.

## 3.2 Persuasion Strategy Labels

The annotators were given instructions to annotate with multiple labels if they felt necessary. Because of this we can provide more than one way to represent the labels if more than one label is present. We opted to provide both textual labels and multi-hot vector representation. A multi-hot vector is a sparse vector whereby the index of the label is incremented by one if present. For example, if an annotator chose both *1-RAPPORT* and *3-EMOTION* the multilabel vector one looks like this [1,0,1,0,0,0,0,0] as the first and 3rd index are increased by one. With the testing data having more than two annotators per utterance the multihot vector will contain more information into which tag was the most confidence.

### 3.2.1 Tag Confidence

Inside the meta data for the annotations we have provided a *confidence* score. The confidence score is the percentage of the annotators that chose the golden tag. For example in a pool of 10 annotators seven chose 1-RAPPORT then this would have a 0.7 confidence. It is also possible, due to the multihot representation, to perform other calculations to find the weighting of each tag in the vector.

## 3.3 Training Data

The final tally of training data is 19176 utterances, however 7755 are mutually agreed tags. For training we have included only the utterances that have an agreed top level persuasion strategy. The disagreed utterances can be imported if necessary.

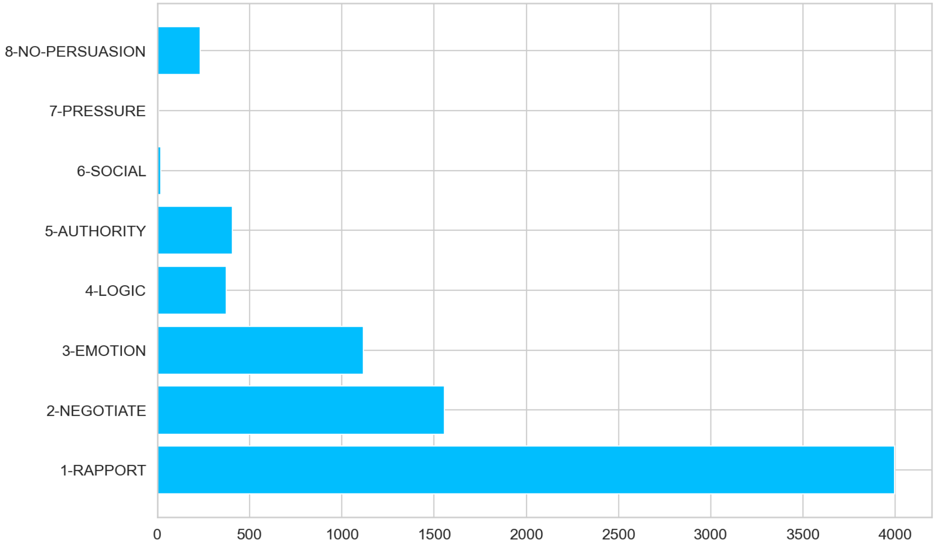


Figure 10: Distribution of Top Level Tags in Training Set

The above chart shows the tag distribution in the training data set. We can see that few of the tags are quite underrepresented in comparison to the first three tags. This can be an issue in training as the model will learn more about *1-RAPPORT* than the others and could lead to over fitting. It would be advisable to attempt to balance the dataset before training.

### 3.3.1 Testing Data

The testing data consists of 25 transcripts which equate to 892 utterances. As the testing data was collected by all of the annotators as well as the project group there are multiple tags which allows us to select the *golden* tag. We can also set a confidence level to choose how confident we want the gold standard to be. The distribution of tags is as follows.

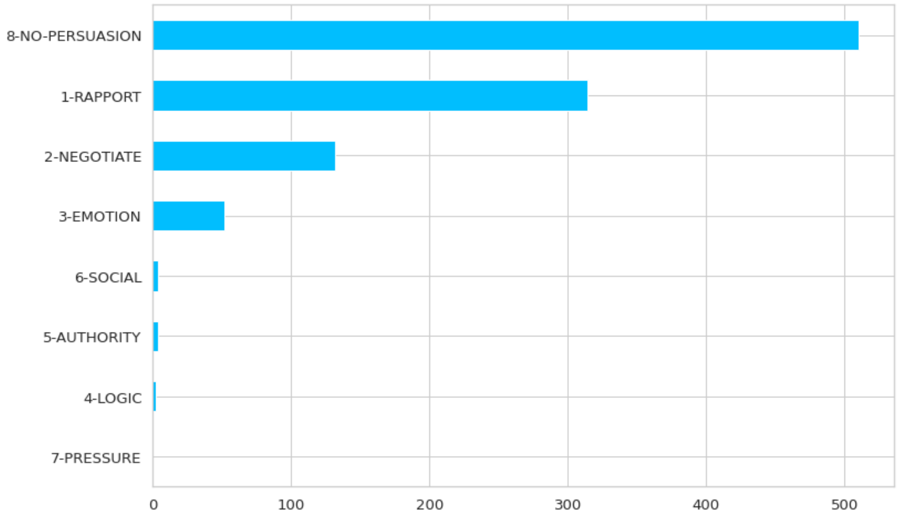


Figure 11: Distribution of Top Level Tags in Testing Set

The testing data shows a similar pattern to the training data, with little to know *7-PRESSURE* and a large distribution of the tags 1-3.

### 3.4 Data Storage

The data has been exported, as well as some notebooks, to a GitHub repository set up by the research project. All data and code can be downloaded through cloning of the repository.

## 4 Dialogue Acts Corpus

As well as the persuasion strategies corpus, we developed a corpus of dialogue acts. The annotators were tasked with re-annotating 100 transcripts marking up dialogue acts present in the utterances. The dialogue acts framework consists of numerous linguistic devices used by speakers. Rather than describing the semantic content of the utterances that contribute to a persusion strategy, the dialogue acts are themselves the techncical function of the language. As the dialogue acts are less abstract and more grammatical it is likely that the annotators would find this task less perplexing as their should be less ambiguity between tags. The full list of tags are below.

| | User_1 | User_2 | User_3 | User_4 | User_5 |
|---|---|---|---|---|---|
| Acknowledge | 21 | 53 | 49 | 71 | 76 |
| Affirmative-Answer | 51 | 58 | 28 | 46 | 30 |
| Agreement | 43 | 63 | 43 | 44 | 50 |
| Backchannel | 1 | 3 | 13 | 0 | 0 |
| Backchannel-Question | 1 | 5 | 10 | 0 | 2 |
| Conventional-Closing | 6 | 21 | 10 | 8 | 13 |
| Conventional-Opening | 43 | 50 | 45 | 39 | 48 |
| Declarative-WH-Question | 2 | 4 | 15 | 8 | 0 |
| Declarative-Yes/No-Question | 3 | 11 | 33 | 29 | 7 |
| Disagreement | 0 | 3 | 4 | 0 | 1 |
| Instruct | 1 | 34 | 2 | 56 | 4 |
| Negative-Answer | 4 | 6 | 5 | 8 | 6 |
| Open-Question | 45 | 20 | 40 | 20 | 36 |
| Other | 9 | 41 | 4 | 14 | 5 |
| Other-Answer | 55 | 98 | 120 | 78 | 146 |
| Request | 5 | 21 | 70 | 47 | 17 |
| Rhetorical-Question | 4 | 11 | 7 | 8 | 6 |
| Statement | 416 | 312 | 200 | 272 | 451 |
| Suggest | 1 | 26 | 10 | 31 | 4 |
| Tag-Question | 4 | 6 | 17 | 0 | 9 |
| Thanking | 60 | 57 | 48 | 58 | 52 |
| WH-Question | 37 | 38 | 24 | 37 | 36 |
| YesNo-Question | 114 | 105 | 40 | 73 | 109 |

Table 10: Dialogue Acts Tag Distribution

The annotators were again given resources online where they can view descriptions and examples of each tag. Due to the annotaotors other commitments we were able to secure five annotators to annotate the small batch of 100 transcripts.

## 4.1 Inter-annotator Agreement

The kappa and accuracy scores were again used to measure the agreement between the annotators. Same as with the testing set we were able to generate a *gold standard* tag from the majority tag chosen amongst annotators.

| | User_1 | User_2 | User_3 | User_4 |
|---|---|---|---|---|
| Kappa | 0.75 | 0.39 | 0.57 | 0.72 |
| Accuracy | 0.79 | 0.45 | 0.63 | 0.78 |

Table 11: Inter-annotator Dialogue Acts Agreement

Table 11 shows that in general the majority of annotators agree with the majority tag. This can suggest that the task is not as complex, apart from User_3 who seemed to have a large amount of disagreement with the gold standard. We can further look into the tags with a confusion matrix. The matrix below 12 is a comparrison between User_4 and the gold standard as this user had the highest level of agreement. The *NO-TAG* placeholder again is present but should be disregarded.

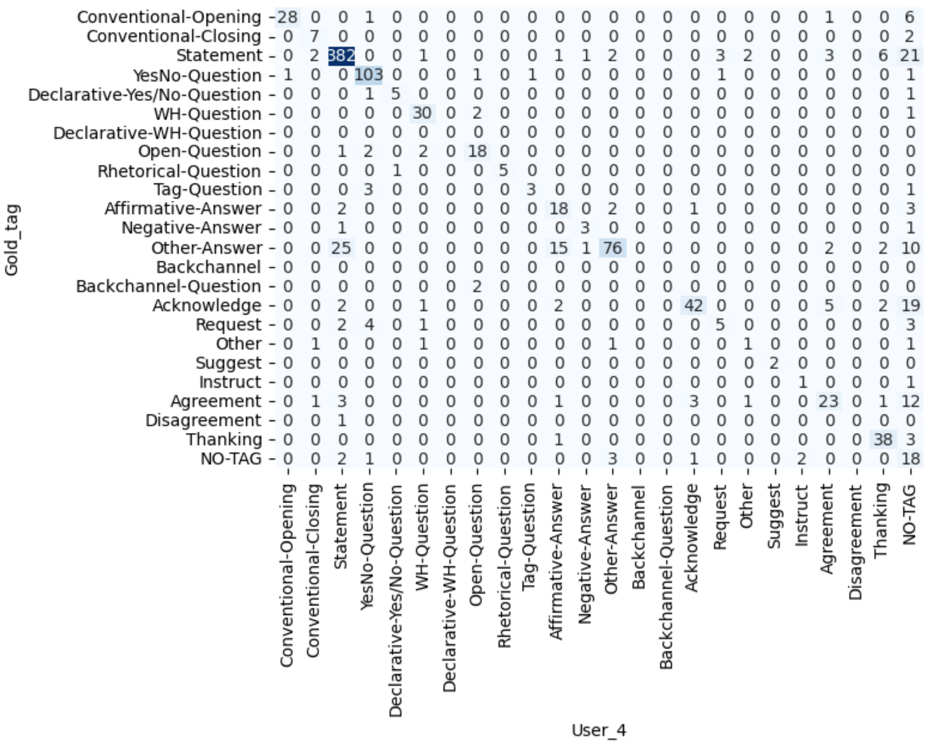


Figure 12: Dialogue Acts Confusion Matrix User_5

### 4.2 Dialogue Acts Corpus Conclusion

The data collected for dialogue acts corpus is a good jumping off point for further work. In the future if resources are available it may be necessary to bypass training annotators and put the task on an online platform such as Prolific or Mechanical Turk. Due to the task being less complex than persuasion strategies, as shown by the agreement scores, thorough annotator training may not be necessary as annotators managed to follow limited instructions and achieve high agreement. The data collected is not, however, sufficient to experiment with modeling due to small amount of transcripts. It would be interesting to revisit the dialogue acts corpus and somehow combine the features with the persuasion strategies to see if they can contribute to higher classification accuracy.

## 5 Modeling the Data

Now that the data has been annotated and exported to a suitable filetype we began experimenting with modeling the data using statistical methods and deep learning. It is important to initially be realistic about what you may be able to achieve from the data available, as the data consists of 1016 transcripts, which equates to around 11,000 utterances depending on how strict we are with the confidence level, it is unlikely that a model that can generalise will can be trained off of the low amount of training data. To train a robust model that is able to generalise across persuasion strategy labels it would be necessary to have a large amount of data with balanced distribution amongst all tags. To get an idea for a base level of accuracy a number of different models were evaluated.

### 5.1 Initial Experimentation

Initial experiments made use of more traditional methods for text classification. The first step of classification is the encoding of the text, of which there are various different methods. Different configurations of text representation and model architecture of model were used to create a relatively well performing baseline model from which to improve from.

### 5.1.1 Pre-processing

Each utterance from the training data is converted into a vector based representation to be used with different statistical models. The simplest of these being a bag of words representation. Each individual word that is present in the training data is given a unique index. A vector of zeros that is as long as the number of types in the vocabulary is generated, to encode a sentence the index of each token is incremented by one. We are then left with a vector that has encoded the counts of each type in the vocabulary that is present.

Further processing to the bag of words model is the transformation of the word counts to the term frequency–inverse document frequency. This representation adds a weighting to each word as some words can be present significantly more than others however but may not necessary contain important semantic information.

## 5.2 Shallow Learning

To begin with we looked at the more traditional and computationally less expensive methods for classification. Initially experiments were made on different configurations of machine learning model, in particular the support vector machine and multilayer perceptron. Both models are considered *shallow* networks as they can be effective on smaller datasets as well deep layers of abstraction. We have experimented with the support vector machine (or SVM) and multilayer perceptron (mlp).

### 5.2.1 Experiment 1: Support Vector Machine

The svm is a good starting point for training a model as it doesn't require as much data, generally, as deep learning models. The model aims to find boundaries between the labels using and use the raw count vectors as inputs.

The data pipeline for this experiment is as follows:

- X inputs consist of bag of words encoded utterances consisting only of unigrams
- y inputs consist of integer representing the class index of the persuasion strategy
- Only tags with single labels are considered
- A max number of 200 iterations or early stopping if the loss does not change after 2 iterations
- The model can only predict one label

### 5.2.2 Experiment 2: Multilayer Perceptron

The multilayer perceptron is a feedforward neural network with one hidden layer. The input feature, which is our bag of words representation of the utterance, is fed through numerous layers and functions to an output layer that gives us the probability distribution of that input belong to each persuasion strategy. As the model trains it is changing weights and biases for each token in our vocabulary to find the best values for each weight. As we are using a shallow network we should hopefully see results even on a low amount of data.

- X inputs consist of bag of words encoded utterances consisting only of unigrams
- y inputs consist of multi-hot vectors, allowing for multilabels
- Only tags with a confidence of 1 are considered, this allows for only data points where both annotators agree
- Onevsrest wrapper, the mlp trains 7 binary classifiers for each label
- A max number of 200 iterations or early stopping if the loss does not change after 2 iterations
- The model can predict more than one label

Deep Learning - Experiment 3:Bert

BERT is an extremely large, pretrained language model for encoding language that is extremely useful when working with small datasets. An advantage of using BERT is that we can encode the utterances using pretrained embeddings and the model does not have to learn as it trains. BERT encodes the whole utterance rather than token by token. A word's embedding may change depending on context rather than the context free representation used in the shallow learning models.The configuration is as follows:

- X inputs consist BERT encoded vectors
- y inputs consist of multi-hot vectors, allowing for multilabels
- Only tags with a confidence of 1 are considered, this allows for only data points where both annotators agree
- Once the utterances are encoded they are pased through a feedforward neural network
- A max number of 200 iterations or early stopping if the loss does not change after 2 iterations
- The model can predict more than one label

The encoding of the utterence is then passed through a feedforward network, like the MLP. Rather than a bag of words representation we have a deep encoding of our utterance that should capture much more contextual and semantic meaning than the simple bag of words.

## 5.3 Evaluation

With the above models configured and trained we can now test them on our testing dataset. Ideally we would like to train a generalised model, that is to say, it is able to classify all persuasion strategies equally without mislabeling. Our amount of data is not ideal, with supervised machine learning we would like to be able to train the networks on as much data as possible. The different models will give an idea on how the imbalance and amount of data affects the accuracy of the model.

To evaluate the models we can look a few metrics, precision, recall and f1 score. These metrics are more useful than plain accuracy. especially in multiclass classification tasks as it can often be misleading. Observing other metrics can be helpful especially when working with imbalanced data. The metrics are defined as follows:

- Precision: In this case precision can tell us, for each class, out of all the positive predictions, how many were correct. Low precision could suggest the model is predicting positives that are not. A very precise model will correctly assign positive labels but may not be good at finding them all.
- Recall: Recall tells us how many positive cases are found however it may assign negative cases as positive. High recall suggests a large amount of positive cases are identified but could include negative cases. Low recall would suggest the model is struggling to find positive classes.
- F1 Score: The f1 score aims to combine the two above. When recall and precision are both high the f1 score is high. A high f1 score can be interepreted as the model classifying correctly without making wrong predictions.

Essentially we are looking for a high balance between precision and recall as well as a high f1-score. If the model is performing well across all classes we would expect to see a high f1 score for each class.

## 5.4 Results

The table 12 below shows the results from classifying persuasion strategies on our testing set. The first model testing is the SVM.

### 5.4.1 SVM Results

When observing the f1 score we can see it is handling tags 2,3 and 5 but struggles to classify the others. There is a very high precision score and a medium recall, this might suggest that the model is predicting the class and getting it correct but at the expense of classifying other tags incorrectly as this label. If we look at the SVM's classification of *8-NO-PERSUASION* has a precision of 1, this implies it is predicting the class positively every time however the recall is next to 0, this might imply that the model is predicting *8-NO-PERSUASION* for the majority of samples that are not.

### 5.4.2 MLP Results

The mlp's results are slightly different, this is because the model is classifying using multi-hot labels as opposed to single label. The differences between the precision and recall are not as large as the previous model however the model has still no performed adequately. It must be remembered that the model is attempting to learn about persuasion strategies simply form the word counts in the utterances rather than a deep representation of the language. Combined with an imbalanced and relatively small dataset it would be difficult for the model to learn and be able to make accurate predictions across multiple classes.

### 5.4.3 BERT Results

Finally the BERT classifier's results are a bit more promising. There is generally not enough variation in the test samples to gain much insight into labels 4-6 but it seems be perform fairly well on the rest of the tags. For tags 1-6 the f1 scores are all over 0.6, this is quite promising as it implies model has learnt something about persuasion strategies and isn't just randomly guessing.

| | **SVM** | | | **Testing Examples** | **MLP** | | | **Testing Examples** | **BERT** | | | **Testing Examples** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1-Score | Single Label | Precision | Recall | F1-Score | Multilabel | Precision | Recall | F1-Score | Multilabel |
| 1-Rapport | 0.42 | 0.98 | 0.58 | 314 | 0.43 | 0.92 | 0.59 | 314 | 0.43 | 0.98 | 0.6 | 314 |
| 2-Negotiate | 0.92 | 0.54 | 0.68 | 130 | 0.75 | 0.58 | 0.65 | 132 | 0.8 | 0.73 | 0.76 | 132 |
| 3-Emotion | 0.56 | 0.83 | 0.67 | 36 | 0.5 | 0.38 | 0.43 | 52 | 0.73 | 0.62 | 0.67 | 52 |
| 4-Logic | 0.00 | 0 | 0 | 2 | 0.12 | 1 | 0.22 | 2 | 1 | 1 | 1 | 2 |
| 5-Authority | 1.00 | 0.5 | 0.67 | 4 | 0.67 | 1 | 0.8 | 4 | 1 | 1 | 1 | 4 |
| 6-Social | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 4 | 0 | 0 | 0 | 4 |
| 7-Pressure | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8-No Persuasion | 1 | 0.05 | 0.09 | 404 | 0.96 | 0.1 | 0.18 | 510 | 0.83 | 0.23 | 0.36 | 510 |

Table 12: Results For Different Model Configurations

# 6 Conclusion

The project has resulted in the completion of three main tasks. The creation of the persuasion strategies dataset, dialogue acts dataset and the training of a classifier that is able to predict persuasion strategies. The persuasion strategies dataset is a great jumping off point for adding more persuasion data to diversify the persuasion strategies present as the dataset is unbalanced and greatly needs more tags that encapsulate tags 5-7.The dialogue acts dataset as well shows that it would be possible to collect more data with further resources, perhaps even at less expense due the task being less complex. Finally the model shows that, with BERT, it is possible to train a classifier with limited data that is also imbalanced.

# References

[1] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text annotation tool for human, 2018. Software available from https://github.com/doccano/doccano.

[2] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. CoRR, abs/1906.06725, 2019.