BEND Annotator Agreement

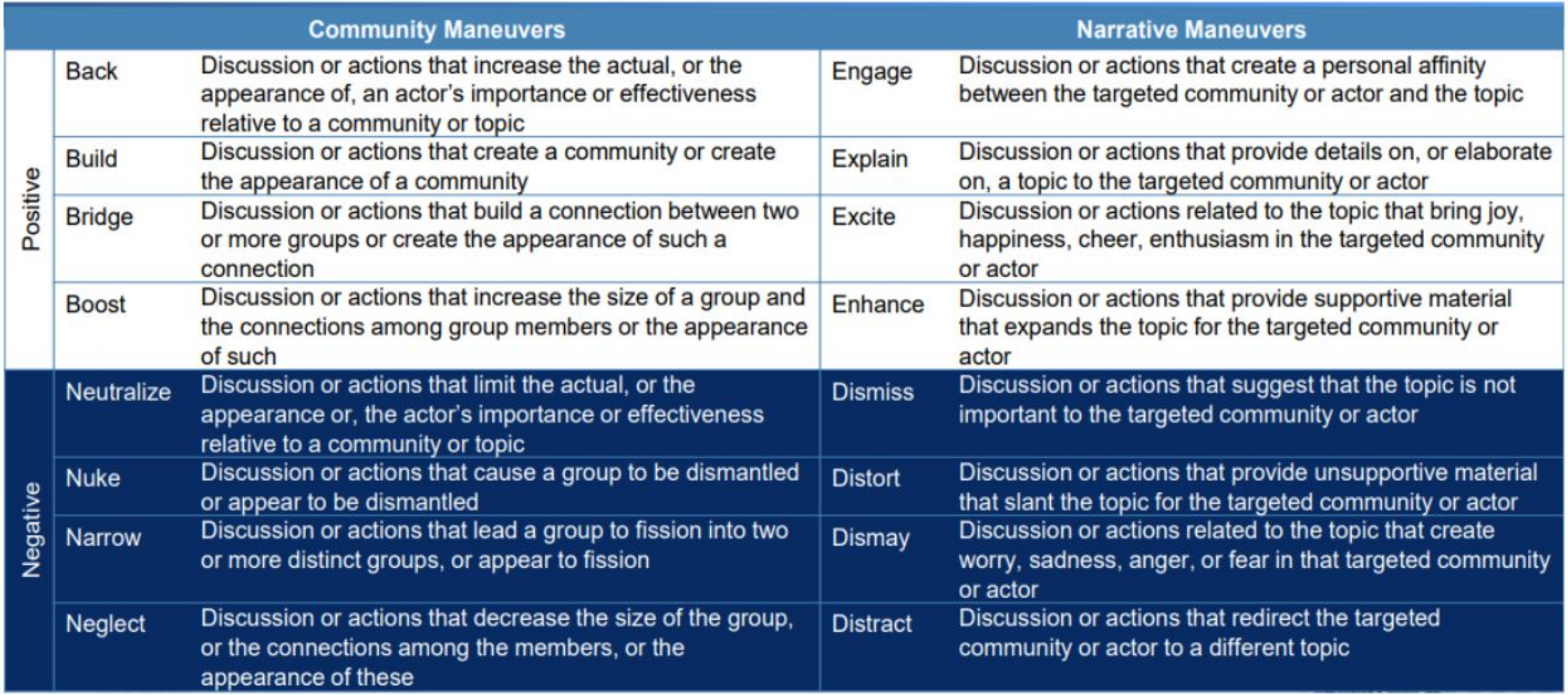
Coco Gong - Carnegie Mellon University Internship

Introduction

BEND Framework

- A set of 16 defined maneuvers that characterize influence campaigns in social networks
- Divided into two categories:
    - **Narrative**: altering the narrative of topic-oriented communities
    - **Community**: altering the positions of actors within topic-oriented communities
- Topic-oriented community: a group of agents discussing the same topic (concept or theme) in a social network

BEND Maneuver Definitions



	Community Maneuvers		Narrative Maneuvers	
Positive	Back	Discussion or actions that increase the actual, or the appearance of, an actor's importance or effectiveness relative to a community or topic	Engage	Discussion or actions that create a personal affinity between the targeted community or actor and the topic
	Build	Discussion or actions that create a community or create the appearance of a community	Explain	Discussion or actions that provide details on, or elaborate on, a topic to the targeted community or actor
	Bridge	Discussion or actions that build a connection between two or more groups or create the appearance of such a connection	Excite	Discussion or actions related to the topic that bring joy, happiness, cheer, enthusiasm in the targeted community or actor
	Boost	Discussion or actions that increase the size of a group and the connections among group members or the appearance of such	Enhance	Discussion or actions that provide supportive material that expands the topic for the targeted community or actor
Negative	Neutralize	Discussion or actions that limit the actual, or the appearance or, the actor's importance or effectiveness relative to a community or topic	Dismiss	Discussion or actions that suggest that the topic is not important to the targeted community or actor
	Nuke	Discussion or actions that cause a group to be dismantled or appear to be dismantled	Distort	Discussion or actions that provide unsupportive material that slant the topic for the targeted community or actor
	Narrow	Discussion or actions that lead a group to fission into two or more distinct groups, or appear to fission	Dismay	Discussion or actions related to the topic that create worry, sadness, anger, or fear in that targeted community or actor
	Neglect	Discussion or actions that decrease the size of the group, or the connections among the members, or the appearance of these	Distract	Discussion or actions that redirect the targeted community or actor to a different topic

BEND Annotations

Annotating (Labeling) Process

- Datasets of 100 tweets
- Two annotators labeled each tweet with one or more BEND maneuvers (or "NONE" if no maneuvers are present)

Topics

- Captain Marvel
- Black Panther
- Election of 2020
- COVID vaccine
- Russo-Ukrainian Crisis
- Attack on France

Purpose of Calculating Annotator Agreement

Guideline Research Questions

- Are humans good detectors of the maneuvers?
- Are some maneuvers more easily detected by humans than others?
- How much do the annotators agree in what they are labeling?
- What is the correlation between maneuver and agreement?

Methodology

Cohen’s Kappa

- A metric used to assess the level of agreement between two annotators (inter-rater reliability)
- Based on the confusion matrix
- Takes imbalance in class distribution and chance agreement into account
- Implemented in Python: `from sklearn.metrics import cohen_kappa_score`

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where p_0 is the overall accuracy of the model and p_e is the measure of the agreement between the model predictions and the actual class values as if happening by chance

Methodology

Calculation Process

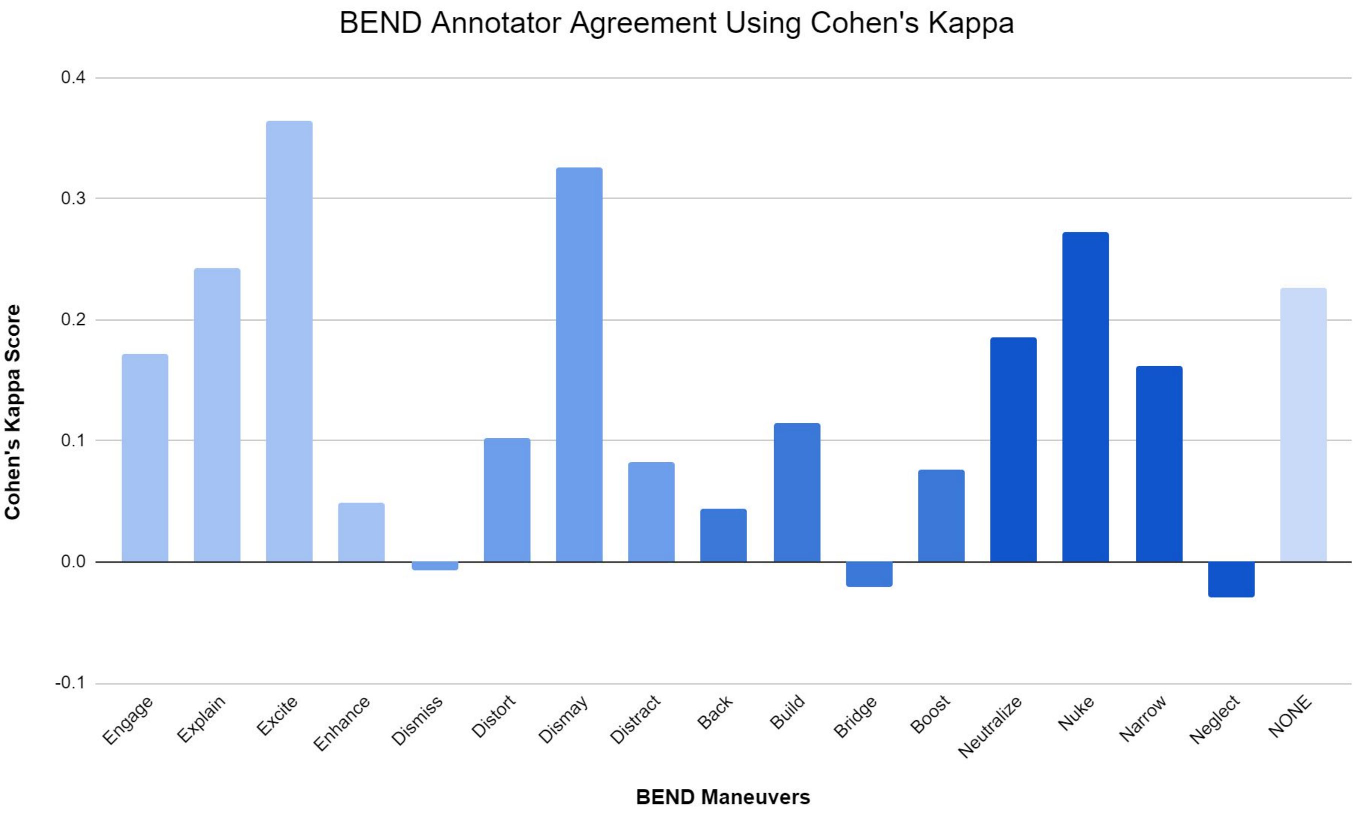
- Processing data for calculation
  - Extracting labels from CSV using Pandas
  - Binarizing labels (Cohen's Kappa does not support multi-label input)
- Calculate kappa score by maneuver
- Calculate average of kappas across dataset topics
- For overall agreement, calculate average of all kappas by maneuver

Methodology

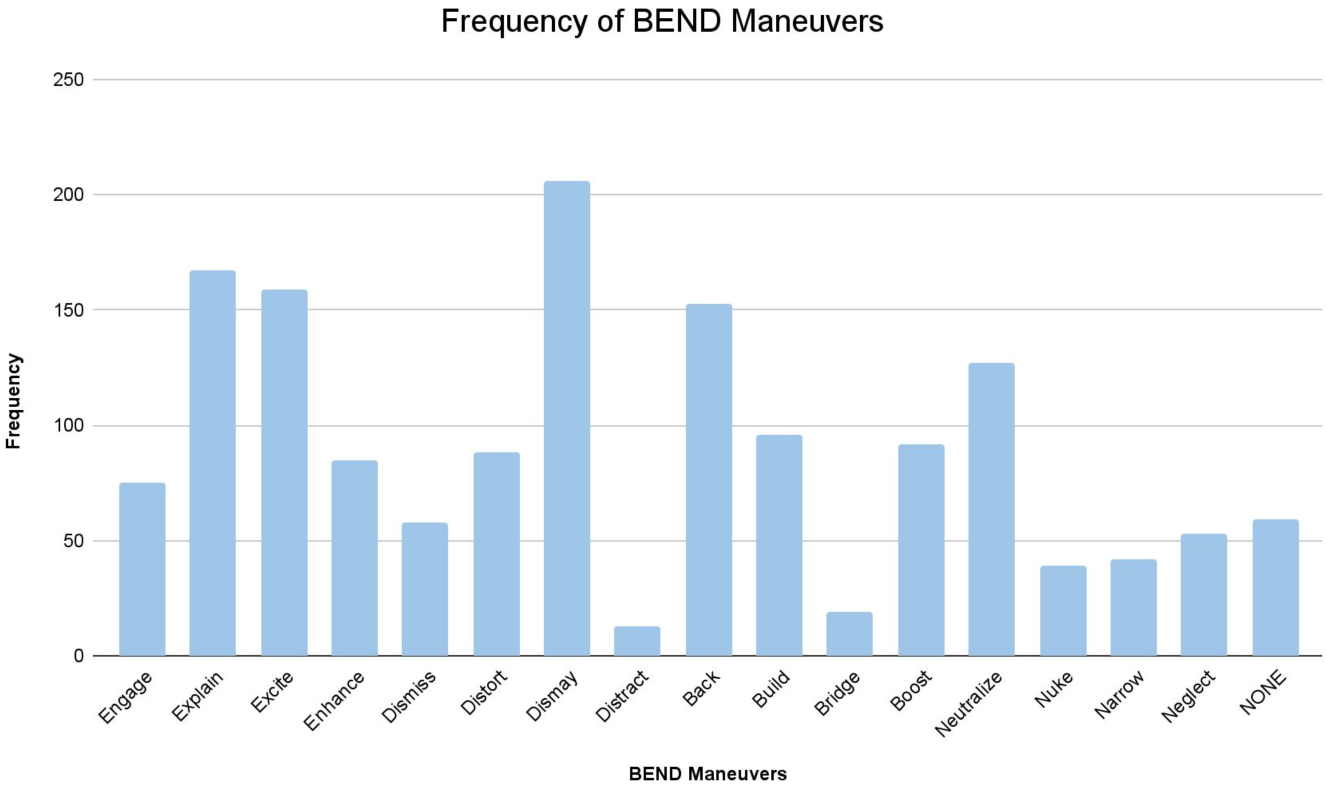
Interpreting Cohen’s Kappa

- <0: no agreement
- 0-0.20: slight agreement
- 0.21-0.40: fair agreement
- 0.41-0.60: moderate agreement
- 0.61-0.80: substantial agreement
- 0.81-1: perfect agreement

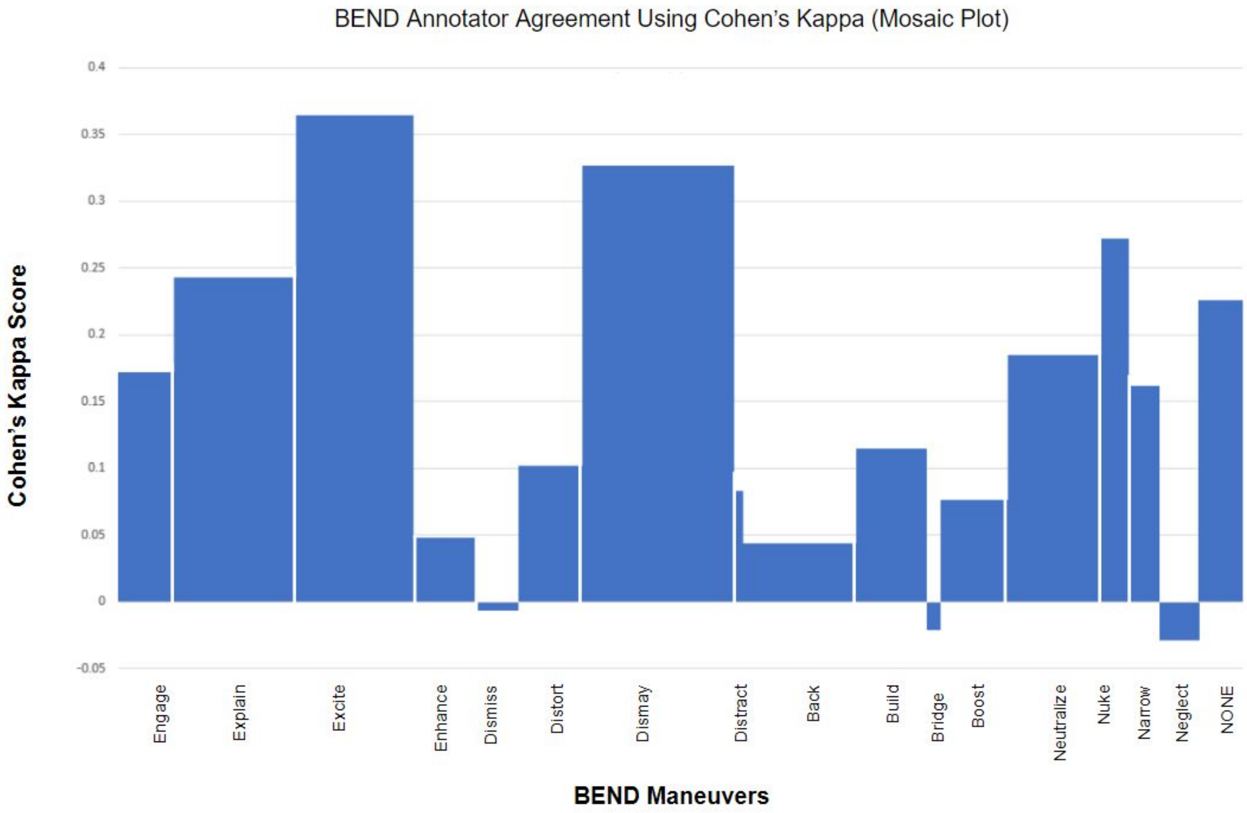
Agreement By Maneuver



Frequency of Maneuvers



Agreement By Maneuver (Mosaic Plot)



Note: The **width** of each bar in the mosaic plot is based on the **frequency** of the maneuver. This demonstrates the **consistency** of the agreement across imbalanced class distributions.

Findings & Results

Engage	0.1719889309739445	**Back**	0.04380936799871941
Explain	0.2427149255131402	**Build**	0.11491840719474185
Excite	0.36445016345346787	**Bridge**	-0.021021821907249622
Enhance	0.048521264862550484	**Boost**	0.07654837656807412
Dismiss	-0.0063405797101449375	**Neutralize**	0.1850906974837134
Distort	0.10213817841038982	**Nuke**	0.27209916683600893
Dismay	0.3263528981968355	**Narrow**	0.1621152872485969
Distract	0.08281523991710504	**Neglect**	-0.029075893616940036
		NONE	0.2262449038073281

Overall agreement between both annotators: 0.1390217360723695

Limitations

- The level of agreement may be affected by:
  - Subjective interpretation
  - Differing amounts of tweets for each maneuver
  - Small amount of datasets used in this analysis
  - Ambiguous of language in the tweet (eg. sarcasm, slang)
- Ways to improve annotator agreement
  - Establishing clearer labeling guidelines
  - Working closely with the other annotator

Other Inter-Rater Agreement Metrics

Fleiss' Kappa

- Works for multiple annotators (Cohen's kappa works for two)
- Allows annotators to measure different items (Cohen's kappa assumes the same item is measured)

Krippendorff's Alpha

- Supports multi-label input (Cohen's kappa does not)
- Works for missing values (Fleiss' Kappa does not)

Conclusion

- The two annotators agreed the most on maneuvers with **obvious indicators**
  - “Excite” and “dismay” had the highest Cohen’s kappa scores and were consistent through a high number of tweets
  - Humans may be good detectors of **emotional maneuvers**
- The two annotators disagreed on maneuvers whose effects are hard to predict
  - Bridge" and "neglect" had the lowest Cohen’s kappa scores and were also infrequent in the tweets used
  - Humans may struggle to detect **future influences** of maneuvers
- The annotators’ overlap is small (slight agreement)