ISyE 6748 Team 'Not-So-Twitterpated'

Piper Gradient Final Report

J. Scott Cardinal and James A. Roberts

Jul 26, 2021

Our project is focused on an aspect of the Piper ∇ challenge #2 – *Topic Analysis and Text Summarization*. Rather than a summary of social media we focus on a characterization of the *styles* of communication in the contents of tweets through discourse and network analyses. Specifically, we identify modes of discourse in open and sincere discussion and debate, amplify aggressively contentious communications, and that generate or foment points of friction between diverse yet legitimate pluralities of opinion.

Our analysis is based on using specific parts of speech (i.e., pronouns) as psychometric indicators of discourse style. In particular, the identification of linguistic markers for *clusivity* and *affinity* in each tweet. The combinations of these two markers indicate a user's self-identification in relation to social communities (i.e. "us" versus "them"). The linguistic markers are used to compute discourse style similarities between individual tweets, which construct a social network graph where each tweet is a node in a weighted acyclic graph connected by discourse style. High-modularity communities are identified and compared to features not used in construction of the graph network. Graph discourse communities are compared to results of both a political typology classification and sentiment classification of each tweet. Our hypothesis is that discourse-type communities reflect the nature of the speech acts, but may not necessarily correspond to individual political types. We are looking for within-community tendencies towards *othering*, strength of belief, conviction or affiliation, and general sentiment.

The motivation for the Piper ∇ concept of operations (CONOPS) is to assess and provide technologies that address information silos and the increasing polarization of public debate. The Piper CONOPS also explicitly requires that enabling technologies *ensure* objective analysis as its defining priority. By addressing styles of social media discourse directly, rather than nominal Tweet content, our method of modeling and summarization identify the behavioral sources for those issues of concern. The form of discourse analysis presented here can be used for identifying polemical and divisive speech acts occurring within social media content. It provides an additional method of understanding how communities are engaging with the specified topic. This can allow users of Piper ∇ to identify not just the dominant topics and opinions of the community, but the extent to which they are engaging in either collaborative or antagonistic discussions.

Background

This analysis uses collections of social media "tweets" on the platform Twitter that reference the current information and debate surrounding the Biden administration's proposed American Jobs Plan (a.k.a., "Infrastructure Bill"). These data were acquired by Piper through the `tweepy` API for the purposes of feasibility and ethics research related to the ∇ project. The data are publicly posted messages, queried from the Twitter API, for keywords and hashtags referencing the Bill.

The public debate surrounding the Bill has become something of a proxy for the larger political differences, with public discussions following party-line rhetoric commonly expressed in the various media outlets (e.g., Congressional Research Service 2021; Ribeiro et al. 2018; Kratzke 2017). In short, traditionally political discourse has *become* the mode of public discourse.

Our initial subjective exploration of a limited sampled from the data left a number of impressions:

1. There did not appear to be many instances of substantive discussion of the contents of the plan itself.
2. Much of the content traffic consisted of re-tweets of a relatively limited number of news items.
3. A substantial portion of tweets seemed to consist of strongly negative characterizations of opposing views.
4. Authors with self-identified political affiliations appeared to tend towards stronger language regarding opposing views.
5. Subjective classification of a sample of tweets for political affiliations along the Pew Research typology often required reliance on tweet author self-descriptions or linked content rather than explicit statements within the posts.

Approaches to social media policy analysis focus on identification of demographics and political affiliations to delineate the domains of opinion regarding topical subject matter (Wojcik et al. 2019; Bakshy et al. 2015; Adamic et al. 2005). Topic and sentiment analysis can help characterize the diversity of public opinions (Pak et al. 2010; Tang et al. 2019; Zotova 2019), while natural language processing tools can assist in summarization of textual content itself (Krejzl 2018; Allahyari et al. 2017). Other methods can be employed to help assess social media content for inauthentic actors, misinformation, or other problematic content (e.g., Ruths 2019; Allcott et al. 2019; Congressional Research Service 2021; Riedel et al. 2017; Shao et al. 2017; Karande et al. 2021; Sterelny 2007). These forms of data aggregation are useful descriptors for the overall landscape of the public debate, but do not specifically target the identification of potential sources of antagonism within the underlying beliefs of the various communities.

Under normal circumstances, it is not only expected but healthy to have a diversity of opinions expressed within public spheres of communication regarding various public policy initiatives. Free and fair debate, even impassioned and zealous arguments on merits, is not the target of our analysis. Where public discourse becomes problematic and inherently divisive (i.e., no viable compromise position or feasibility for modification of views) is where the positions communicated rely primarily on negative characterization of the opposition instead of rhetorical or empirical validity of the argument itself.

Methodology

Our general methodology is to combine anthropological discourse analysis with natural language processing and social network analysis to address the challenge domain. Initially we had identified “sentiment towards others” as the target for analysis characterizing problematic social media discussions. As such, our focus was on text classification, topic modeling, and stance analysis by identifying stance toward a subset of topics (i.e., named entities and noun phrases). Evaluation with the transformer-based tools for natural language processing suggested a more

direct approach by leveraging psychometric attributes of the performed speech act contained in the tweets .

This new approach, grounded in functional linguistics (Íñigo-Mora 2004; Tausczik et al. 2010), allowed us to assess key aspects of the style of communication directly – irrespective of any specific topic or political group – thereby avoiding a substantial area of ethical risk inherent in the ∇ CONOPS. By evaluating speech acts independently of speech content, potential for political biases in analysis are significantly reduced.

Anthropological Linguistics

When we express an opinion in matters of politics, religion, or other matters of preference and belief we do so along lines of strength of conviction and investment. The pronouns used in these statements are especially revealing of the strength of one's conviction on that matter. Pronouns tell us where people focus their attention. Among the most telling are the simple first person (singular and plural) "I," "me," "we," "us," and third person (singular and plural) "you," "they," "them," "those." From these pronouns we can establish whether or not an individual is counting themselves as an committed part of the group – "I" or "me" indicating affiliation – or are they indicating a less strong association with the use of "we" or "us" (Tausczik and Pennebaker 2010; Pennebaker 2011). By use of "we" or "us," the individual retains a certain quantity of autonomy while expressing association. An individual can agree with certain high level associations and affiliations, but they can choose to reject lower level actions.

These terms are also a matter of *clusivity*, a linguistic term used to describe whether or not an individual is counting themselves as part of a the group – inclusive vs exclusive. Exclusive indicating that the individual is regarding themselves as separate (Filimonova 2005; Moskal 2018). This information reveals the distinction between self (and groups/populations one belongs to) and those that self does not belong to and believes self to be distinct from. Generally referred to as *othering*, these pronouns of distinction ("us" vs. "them" – "they," "them," "those") typically indicate a discourse of comparison, judgment, or critical evaluation of someone else (Dervin 2015). A final measure of an individual's mode of discussion is the presence of a few key indefinite pronouns - "all," "each," "every," or "none." These terms describe all participants of a group, and generally signal the presence of a logical fallacy. The use of invalid or otherwise faulty reasoning tends to indicate extreme positions on a subject.

Linguistic Association of Selected Pronouns are detailed in Appendix 4.

Natural Language Processing (NLP)

Natural Language Processing (NLP) is a broad class of machine learning methods that provide a computational framework for analysis on unstructured text data (Joshi 1991; Jones 1994). NLP focuses on the relationship between lexicon (i.e., the individual words), syntax (i.e., the rules of arranging words), and corpus (i.e., a collection of texts). A more recent approach,Natural Language Understanding (NLU), attempts to capture the semantics or meaning of lexical elements in their usage contexts across multiple corpora (Saba 2007). Current methods for both are heavily reliant on neural network approaches to machine learning.

Our analysis incorporates some aspects of both NLP and NLU, using transfer learning

frameworks implemented in the `transformers` and `spaCy` python libraries, for both text classification and extracting parts of speech (PoS) directly from the tweet texts with minimal processing. Transfer learning frameworks use a back-end language model (LM) of high-dimensional text embedding of usage that capture word (or token) contexts that have been pre-trained on large corpora. Transfer learning applies a newly trained "head" neuron layer to apply the pre-trained embedding to the current data, thereby "transferring" the pre-trained embedding to the new texts.

A survey of NLP options revealed that the python **transformers** library by Hugging Face is widely considered to be state-of-the-art (implementing recent advances such as Google's BERT in 2018 and Facebook's development of RoBERTa) and has implementations (based on RoBERTa) that are pre-trained on and fine-tuned for Twitter specifically. Unfortunately, there are no publicly available language models or pre-trained pipelines for part-of speech classification for Twitter. For that task we used the spaCy python library, which also uses a RoBERTa-based transformer model (albeit not one specifically tuned for Twitter).

Social Network and Graph Analysis

Social network analysis and graph analysis are well established methods for addressing semantically linked or embedded data. Numerous social and physical phenomena exhibit network properties, and graph theory is an efficient computational heuristic for analyzing community structure within such networks. Community detection within network graphs operates solely by attributes defined by the vertex and edge relations within the graph, independently of the underlying (often high-dimensional) data from derived associations. This allows for very robust and efficient clustering solutions to complex or non-linear sets of relationships. Rather than computing distance measures in high-dimensional space, graph algorithms find optimal solutions by traversing the edges within the graph.

For our analysis, we construct a discourse network derived solely from the counts of certain targeted linguistic features. This allows community detection without direct assessment of the specific content of each tweet, allowing us to avoid induced biases related to the political affiliation or sentiment of the Twitter users.

Data Preparation

A relatively small amount of data preparation was required for our analysis. Transformer-based models do not necessitate many of the common text pre-processing required by previous NLP methods (e.g., lemmatization, stemming, stop-word removal, etcetera). Most of our data prep entailed replacing null or 'NaN' values with empty strings, conversion to a standard character encoding, and removal of line-breaks.

The only modifications to content required was masking of user tags and links (both required by our chosen pre-trained language model) and removal of the common re-tweet prefix "RT @User." Additional preparation included extracting mentions, hashtags, and links into separate `pandas` data frame columns in case they might be used in later steps of analysis. We also added a boolean column to indicate the row's re-tweet status and another to extract the tweet's previous source.

Feature Construction

A substantial amount of our efforts were devoted to extracting the features of interest for our analysis. There were three specific features not already isolated within the data as provided by the tweepy API:

1. linguistic features contained within the unstructured main text of the tweet,
2. the general sentiment expressed by the tweet, and
3. the category of political affiliation for the tweet along the external typology provided by Piper.

Each of these features of interest is something that needs to be extracted from within the unstructured text or to be imputed by a classification model. The data provided contained a subset that had been manually labeled for political typology, but for comparison of discourse communities and political affiliation, the remainder of the tweets need to have political affiliation imputed. No labelled sentiment classification was provided, so this was imputed by model classification to tweets as well. Although the linguistic features are contained within the main tweet text itself, part of speech tokens of interest required extraction by NLP model parsing.

Text Classification - Political Typology

Piper typology was provided for 1,950 of 16,638 tweets in the large data set. To extend the typology to all observations we trained a classification model using these 1,950 examples, with two thirds of the labeled data (1,300 observations) used for training and 650 observations reserved for evaluation of the results.

Using the `twitter-roberta-base` model, the process recommended by Cardiff NLP recommends scrubbing the text to substitute generic "@user" and "http" fo specific users and links. (No attempt was made to retrieve and evaluate the content of any links). We compared the results of using just the cleansed tweet text as a predictor, and combining it with the cleansed "user_description" where this exists (most often but not always present). The latter (combined user description and tweet text was found to have significantly higher accuracy (82% vs 57%) when evaluated on the held-out validation data. Since the Piper typology is also approximately ordinal, we also examined MAE, and found that for tweet text alone it was 0.70, while for the combined user description and text it was just 0.36. This low MAE indicates that in the case in which the typology is not accurately predicted, it is most often only one category removed. (The full truth table is shown in Appendix 5).

Runtime on a Colab GPU was <2 minutes for the large dataset.

Sentiment Analysis and NLP Model Selection

Tweets incorporate many non-standard linguistic elements: hashtags, URL links, @user mentions, emojis; casual or non-standard grammar and spellings "text-speak" (e.g., "imho," "smh," "lol," "iir," etc.); and phrases expressed solely as acronyms. These non-standard elements argue for a Twitter-specific language model. A Hugging Face transformer model based on Facebook's RoBERTA was selected: the Cardiff University models `twitter-roberta-base` and `twitter-roberta-base-sentiment` were selected. These models were pre-trained on ~58M tweets, achieving high benchmark scores, as described and evaluated in the TweetEval

benchmark (Findings of EMNLP 2020).

The performance of the Cardiff Twitter RoBERTa models was validated during our sentiment analysis, discussed below. We evaluated 18 tweets as positive or negative (an additional two were undetermined), and checked the performance of sentiment analysis via Textblob, vader, the default model used by Hugging Face transformers for sentiment analysis (`distilbert-base-uncased-finetuned-sst-2-english`, not fine-tuned for Twitter) and the Cardiff Twitter sentiment model. The results, except for Cardiff's RoBERTa, were disappointing; RoBERTa was clearly the best (though admittedly this is a small sample).

The ability of the RoBERTa model to properly assess tweet sentiment supported its performance claims and gave us confidence to choose this model for Gradient.

Runtime on a Colab GPU was <2 minutes for the large dataset. Examples and details of performance are shown in Appendix 3.

Part-of-Speech and Discourse Analysis

To perform the type of discourse analysis described above, we needed to extract the specific parts of speech - in this case pronouns - from the body text of the tweets. Simple regular expression search on the text would not suffice in this case, since it may not reliably capture only the target pronouns rather than any occurrences of the string. Furthermore, such regular expression search would not generalize to other languages for future applications. For our purposes, even targeting a relatively straightforward set of linguistic features, using NLP for extracting and scoring provided significant confidence over regular expressions.
Our target features consisted of five distinct pronoun forms:

- Inclusive Affiliation - 1st person singular
- Inclusive Association -1st person plural
- Exclusive Affiliation - 2nd person (singular or plural)
- Exclusive Association - 3rd person plural
- Absolutive Terms - Indefinite pronouns (All, None, Every, Each)

Using the spaCy NLP library, it was simply a matter of defining a small number of rules to classify four particular combinations of part of speech (pronoun) and morphology (1st, 2nd, or third person singular or plural) and a word list for the absolutive terms. Stemming and lemmatization are not necessary for transformer pipelines, so any form or spelling of these tokens will be tagged appropriately by the NLP pipeline's parser. The rule-based matching implemented in spaCy only requires that the target rules be added to the NLP pipeline as python dictionaries. For example, the rule for matching 1st person plural pronouns ("inclusive association") is added to the pipeline as:

```
incl_assoc = [{'POS': 'PRON', 'MORPH': {'IS_SUPERSET': 'Number=Plur','Person=1']}}]
morph_matcher.add('incl_assoc', patterns=[incl_assoc])
```

This allows any occurrence of a term matching that part of speech and morphology to be tagged as "incl_assoc" by the parser. The full text of each tweet is fed through this pipeline and each occurrence of one of these rules flagged by the parser, allowing the token to be extracted and the

count score for each occurrence of each rule updated for the tweet. Scoring for each of the four dimensions for the clusivity and affinity combinations and one for the absolutive terms are tabulated and added to the data frame. Scoring along these dimensions is, again for our purposes, a simple word count. Previous work by Penebrook's team (Tausczik et al. 2010; Pennebaker et al. 2015; Dudău et al. 2021) has shown that this form of scoring provides adequate psychometrics when applied to pronoun usage in NLP.

In addition to the five discourse scores, we extracted the tokens and recorded them as a list in the dataframe for the ability to inspect and compare. We also extracted and stored noun phrases and named entities for possible future use or integration with other Piper teams. The nouns and named entities within the tweets also identify the referents of the pronouns, which may be of use in more refined future analyses such as coreference identification and stance analysis.

Discourse Type Graph

Using the five dimensions of discourse type extracted during the feature construction step, we built a weighted graph network of the tweets solely from those linguistic markers. Sentiment and political classification were excluded, and any specifically identifying text within the tweets such as users or mentions were masked. Graph construction consisted of building a similarity matrix to determine node adjacency, weighting that adjacency, and passing the resulting *n* by *n* sparse matrix to the networkX python library. The API for networkX converts the adjacency matrix into *n* nodes connected by edges weighted by the adjacency of the two nodes. This returns a graph object whose edges represent the measure of similarity between tweets across all five discourse dimensions simultaneously. Our goal is to find and delineate discrete sub-graphs within this network, which would represent different combinations of those five linguistic markers.

Network Construction

Our assumption, an extension on Penebrook et al, is that scores across these five linguistic features constitute a compositional profile of the tweet's discourse style. Given the compositional assumption, and that the scores are measured as discrete count variables, we chose a signed Pearson's product-moment correlation coefficient distance (1-r) across the five variables as the basic similarity measure. This required that any tweets scoring zero across *all* five features (n=7,385) be excluded to avoid degenerating calculations on zero variance. Logically, this should not affect the community detection of the remaining tweets, which would have at least one score > 0, since this group would form its own separate clique within the network. This left 9,253 tweets for the following network analysis stages.

In order to avoid spurious weak adjacencies from low coefficient similarities, we applied a thresholding function adopted from co-expression network analysis (see Horvath et al. 2008; Langfelder et al. 2008). The adjacency matrix is calculated by a sigmoid power function $Adj = \rho^{\beta}$ across the similarity matrix, where ρ is 1 minus the coefficient of correlation *r* and β a chosen power threshold for adjacency. Tests recommended values of β= 6 for power-law distributed similarities, as typically found when similarity is reflecting multiple feature co-expressions.

An additional step of forming a topological overlap matrix (TOM, Yip et al. 2005; Yip et al. 2007; Li et al. 2007) from the adjacency matrix was also applied. Topological overlap is somewhat

analogous to gradient boosting or inverse-distance weighting (or Tobler's Law for those familiar with geography) as applied to network graphs, and its use for co-expression analysis is well supported. The premise is that nearer adjacencies are inherently stronger effects, and measures the estimated overlap of nodes within a network geometry based on the initial adjacency. Two nodes in a graph are considered to have a high topological overlap if each node is connected to approximately the same group of other nodes. This leads to a simple calculation for this weight. The resulting matrix is then passed to networkX to form the discourse graph, and an acyclic graph network is formed with edges weighted by the topological overlap between nodes.

Community Detection

For detecting communities (see Newman 2006; Clauset et al. 2004) within the discourse network, we evaluated four different methods appropriate for community detection in *weighted* and *undirected* acyclic graphs. Each one searches for a local optima to decide the number of clusters using a variety of heuristics. All of the methods are optimized to detect highly modular (i.e., coherent) sub-communities within a larger network.

The Louvain (Blondel et. al. 2008) and the later Leiden (Traag et al. 2019) methods both use a two-stage iterative algorithm to remove nodes from randomly initialized communities, then calculate the resulting gain or loss in modularity. The Leiden method adds an additional aggregation stage to ensure well-connected communities. The "Chinese Whispers" fuzzy clustering algorithm (Biemann 2006) is a randomization-based algorithm designed specifically for NLP-related graph problems. The InfoMap method of random walks (Rosvall et al. 2008) is another randomization-based approach, using random walks along graph edges between nodes, that applies an information theory heuristic to community detection.

All methods except Infomap (c=4) detected eight or nine communities and converged to a solution quickly, taking only two minutes even running locally on very modest hardware. We ran the ensemble of methods through a grid-search optimization pipeline provided by the CDlib library (**cite**) for each method's tuning parameters to ensure optimal configurations.

Community Evaluation

For evaluating and comparing the performance of the community detection algorithms, we used two common and well-supported metrics: modularity and performance. Modularity is a measure of sub-graph or network community structure. The modularity score is based on the divergence between the actual edge connectivity between nodes and the expected number of edges in a randomly connected graph. Performance is a measure of the coherence of a partitioning over each sub-graph or network community. To determine the performance score, the ratio between the sum of the number of intra-community edges and inter-community non-edges to the total number of potential edges in a graph is calculated. The optimal number of communities occurs at the intersection between peak modularity and diminishing gain in performance, and the method providing the highest score on each measure was selected.

The "Chinese Whispers" method was determined to be the best solution at nine communities, with a modularity score of 0.7856 and performance of 0.9792. The Louvain method was a close second at eight communities (mod. 0.7830, perf. 0.975). Infomap performed the worst of the four

methods, detecting only four communities at a modularity of 0.6877 and performance 0.8527.

Summarization

Having identified clusters of tweets with similar discourse styles, what features do they share? In the first instance, they will share features based on discourse, because that formed the basis of the community detection. Examining those shared features will help us understand how the various styles of discourse are differentiated. If we also examine other features, we hope to learn whether any of these others are also correlated with or characteristic of these styles of discourse.

Examining the mean values (q.v Appendix 2), we see that clusters 1 and 5 represent styles of associative and affiliative inclusion, while no others exhibit these modes. -1 is the cluster of those tweets that use none of the linguistic features used to determine discourse style. Cluster 8 invokes only absolutist terms, and the others all exhibit use of exclusive discourse in various combinations with other styles. Significantly, there are many examples of clusters of discourse style that are largely heterogeneous with respect to Piper political typology. To the extent that differentiation exists, we see clusters 4, 6, and 9 dominated (more than 50%) by Democrat (-1) and cluster 7 skewing Progressive (-2). Only cluster 8 skewed right (mean 0.15, but median of Centrist (0)). Clusters -1 and 1 are also well represented by the right, but these clusters are very heterogeneous, with significant representation across the spectrum. (refer to appendix 2 for additional detail).

Ethics Report

Concurrent with analytics methods to address the Gradient CONOPS, the ethical implications of both methods and the broader context of the project were considered. Both team members for this project shared very similar motivations, so the underlying risk of cognitive and confirmation bias was very real. Particularly so when coupled with the subjective nature of the domain. Operational users and administrators of Gradient would likely be similarly susceptible, so consideration of this challenge and identification of any safeguards or mitigating processes is very important. Increased exposure to information and subject matter expertise is likely to be no protection against such bias. The best defence would seem to be to ensure exposure to participants from diverse backgrounds and perspectives.

Technical bias from the frameworks adopted and from our own design choices is also a challenge. We have described some of those threaded through the discussion on our methods, identifying the design choices made to mitigate some of them. AI Ethics guidelines and summaries such as those from IEEE, Intel.gov, and Carnegie Mellon SEI, as well as design principles such as those espoused by MLOps, have been helpful in considering how to identify and mitigate against such ethical concerns.

The data itself is problematic, as is the overall domain of analysis. Not only is the field of NLU fundamentally concerned with the subjective (interpretation of meaning and sentiment from human languages, in all their variety), the data from Twitter is a particularly peculiar sub-genre. The samples provided for training included a high proportion of non-unique information (i.e.,

re-tweets), political categorization of only ~12% of tweets, and imbalances including under-representation of some of those political categories.

Data Biases

There are several sources of bias inherent in the sample data available for this study. Since this analysis is experimental the effects of these biases are minimal, but the issues would need to be addressed prior to wider testing or implementation. The data biases found during our project were:

- large number of re-tweets to original or unique content,
- significant number of duplicate tweets,
- imbalanced samples across political types,
- presence of corporate or sponsored content, and
- capture of tweets unrelated to the American Jobs Plan.

The data is skewed heavily towards re-tweeted content, often duplicated numerous times throughout the data set. Duplications amplify the influence of the re-tweeted content, and limit the amount of unique content on which to base a model.

It is not balanced across the eight Piper political typologies. Of the 1,950 labelled Tweets, Democratic (label "-1") alone accounts for 41% of the data, and overall, left-leaning tweets accounted for 69% of labelled data, versus 28% representing right-leaning views. This may or may not be typical of the user and/or tweet volume on Twitter, but is clearly not representative of the overall distribution of political leanings evidenced by the voting public, so any interpretation of the data should not be taken as representative of public opinion. Likewise, this method can capture discrete opinions from a user which may or may not be as strong as opinions on different topics.

Indeed, research suggests that Twitter users are not demographically diverse in other ways: Pew research suggests that among US Twitter users, besides leaning Democratic, are more likely to be young (ages 18-29 and 30-49 over-represented by more than 33%), college graduates (42% vs 31% of US adults) and higher income (>$75k, 41% vs 35%), and the 10% who tweet most often (accounting for 80% of tweets!) focus more on politics and are in fact mostly - 65% - women. More astonishingly, according to marketing agency Omnicore, 80% are "affluent millennials!" Again, this serves to highlight that size of communities and volume of tweets should not be extrapolated to the community at large.

There is a large volume of sponsored or corporately curated content, exacerbated by re-tweets, that is undifferentiated from community content, and which better represents media discussion than public discussion. We have not deliberately identified this data before performing our analysis, but we have noted anecdotally that it tends to be classified as "centrist" in the Piper typology. We hope the models of Challenge 1 may prove effective in differentiating this content. While the results of typological classification appear to be good, they suffered from under-representation of some classes, and may have been boosted by the classification of retweets for which a labelled example existed. Training on larger datasets with typology labelled would allow better modelling for auto-classification.

The sample topic provided was taken from tweets returned as a result of a search query

“Infrastructure Bill” and returned predominantly (if not completely) tweets in English, which are likely to over-represent the views of native and fluent English speakers and under-represent those (generally minorities) who may have an interest in the topic are actively engaged, but who use languages (such as Spanish or Mandarin) in their discourse. Thus the data may not only under-represent the US body politic in general, but also the Twitter discussion surrounding the Infrastructure bill. While this might not represent either a significant volume or a skewing of views in this case, one might imagine the results could be significant if, for example, the search term had been “immigration reform”!

For Gradient to move forward, we would need to consider how to collect and analyze responses that are topically equivalent across other major languages. Can dimensions be constructed independent of language? Is this desirable? And even if so, should we preserve language as a dimension, e.g. for community detection? Identifying pockets of reasonable and unreasonable discourse in Twitter cannot be assumed (absent evidence otherwise) to be a valid proxy for the broader population. However, the results can illuminate our understanding of the conversation when framed in its proper context.

Output Assessment

The majority of methods are unsupervised or semi-supervised, so there are few available diagnostic metrics to measure the accuracy of each step let alone the entire model. The combinations of NLP and transfer learning are relatively new technologies. In general, Natural Language Understanding (NLU) is not fully evaluated, and the performance (accuracy, absence of unwanted bias, compute speed) is capped by technological developments.

No objective diagnostic or benchmark metrics of final output exist. Evaluating summarization is, by nature, a largely subjective and qualitative evaluation. Our output is a summarization of communities clustered by similarity in political persuasion and in the nature of their discourse (othering or “clusivity,” affiliation, conviction, sentiment). Validation is dependent on a combination of qualitative assessment and clustering quality metrics such as coherence and modularity on centrality measures.

These can provide guidance on fine-tuning model parameters, but not the final output in any meaningful sense. Ultimately, evaluation of meaningful (i.e., human-readable) partitioning and summarization require some degree of subjective intuition and insight on the domain. Review of each cluster’s tweet text (and possibly user descriptions) will be required to determine whether clusters exhibit consistency of political viewpoint and discourse type. Internal quality control, by human assessment, of tweets associated with clusters is unlikely to scale well for ∇’s broader operations!

A computational/AI based check could be implemented which would help safeguard against bias or inaccuracies by evaluating model consistency. Given clusters and associated tweets, a text classification model could be trained on completed summarization to predict which cluster they represent. If the model is generating clusters that are, as we hope, internally consistent along the design dimensions, we might expect them to also be similar in ways that are identifiable by text classification. The performance of this model on test data could be used to evaluate whether the clusters are consistent; if successful, such a method would gain credibility and trust as past results are known and benchmarks can be established.

The challenges in evaluating the subjective output and possible risk of inaccuracies going undetected is mitigated somewhat if Gradient is only used as a tool for understanding the landscape, and not for automated decisions or intervention.

Human Interaction and Bias

Since qualitative and subjective assessment is required by the nature of the tool, human-in-the-loop is necessary. Risk of cognitive and/or confirmation bias is high given the nature of the analysis domain. Users and operators of Gradient would be presented with, and responsible for, custody of social and political views both in alignment and contradiction with their own. Being presented with an analytical assessment of the full political and social landscape may in fact serve to reinforce existing human biases if they happen to align, and conversely, people are liable to find fault when perhaps valid results conflict with their pre-conceptions.

Sampling from multiple observations by multiple people, ideally with relevant but different areas of expertise, before making decisions, is the best way to counter the introduction of bias by humans. Even if the initial model is valid and accurate, unwanted bias may be introduced by model drift. As socio-political attitudes and norms change, populations shift, topics emerge or subside, and language and expression evolve – the tendencies to retweet or hyperlink specific content increases or decreases. Baseline data and labels used to train models and evaluate output may lose relevance or accuracy, so no static model training is viable.

Ultimately, the subject domain of the ∇ project is natural human social behavior. Therefore, human accountability needs to be embedded in the ∇ CONOPS regardless of the technical sophistication involved. A Project Manager for Piper Gradient should focus on ensuring models are exposed to not their judgment, but to the judgment of as diverse a group of people as possible and practical; indeed sacrifices may have to be made to facilitate the ongoing feedback. Customer surveys and complaint processes should be in place to actively seek the feedback of the community using Gradient to identify weaknesses and opportunities to improve.

Communication of bias and potential impacts

The limitations, performance and evaluation metrics, design rationale and justifications, and both the nature and basis of all output should be communicated to both users and administrators. Understanding these risks, administrators need to not only do their best to mitigate them, but also need to communicate these shortcomings and risks to users. At minimum, administrators would need to provide information on:

- sampling frequency and period
- last labeling update and proportions of manual to model-based labels
- performance validation metrics
- update frequency of performance evaluations
- re-training frequency
- performance of intra-operation sub-models
- disclosure of the sources of pre-trained or 3rd-party models
- vetting process for models and 3rd party contributions

These should be presented and/or available in clear and concise language, understandable regardless of the technical background and proficiency of the target audiences.

Existing biases not yet remedied (e.g., English-only model, under-representation of some political typologies) can be documented and provided with Gradient training, and users taught to consider these risks when reviewing and potentially making decisions based on the Gradient output. As Carnegie Mellon SEI suggests, “(ensure) biases are known and explicitly stated”.

Both users and administrators need to be fully aware that no model can accurately or adequately account for the ambiguities of the real world. All models (mathematical or otherwise) are inherently reductionist to some degree, and therefore embed the simplifying assumptions on which the model is based.

Third-Party Technology Risks

Third party data - particularly, pre-trained libraries - may not have been trained on documents that are as representative of the documents we wish to analyze as we believe.

In general, the difference between the corpus and the target population (documents, tweets, etc) in term of the library - in language (lexicon), words/symbols/punctuation, topics, expression, sentiment - the further the embeddings may be away from those that would best characterize the population we wish to analyze. This can result in model output that inadequately represents the content of the tweets. We saw evidence of this variability, and the poor accuracy that can result, when we were comparing language models for sentiment analysis.

Likewise, 3rd party algorithms will have been designed to perform tasks that may resemble that which we wish to perform, but that have subtle distinctions - different objectives, motivations, or underlying biases. These unknown design decisions that allow the algorithms to meet the maker’s purposes could significantly affect its application to our task in undesirable ways, or might be misused by us due to misunderstanding the maker's intended scope of application.

A notable shortcoming, currently, is a lack of publicly available and well-vetted language models for Part-of-Speech tagging and parsing based on large Twitter corpora. Although the general web-trained language model available for `spaCy` performed very well on our limited sample data, it is unknown how that would generalize. Similarly, models and tasks trained for Twitter may not perform as well on data from other social or news media platforms.

The remainder of the 3rd party libraries we are using (e.g., `numpy`, `pandas`, `networkX`, `spaCy`, etc.) may potentially contain imperfections in their algorithmic implementations, but do not represent ethical hazards in any true sense. The remaining 3rd party is the Google Collaboratory platform, which may entail some privacy concerns as a cloud-based platform.

We mitigate some of the risks from use of 3rd party data and algorithms by selecting publicly available solutions that are, as well as we can discern:

- Based on or designed for data as close to ours (Twitter) as possible
- Intended to achieve the task we want
- Documented to achieve good results
- To the extent possible, well known to the NLP community

Furthermore, we assess the performance on the task versus outcomes that are judged

independently of the AI processing (i.e. versus outcomes assessed by us or the Piper team without reference to the AI outcomes). For example, we reserved some manually labelled tweets form the Piper-provided data for model testing. We also randomly sampled from the data in order to manually assess sentiment for comparison with the machine-classified results.

Explainability and Interpretability

Some model components – clusivity, affinity, conviction – are interpretable and explainable given a level of training in the associated linguistic theories. The scoring of these dimensions (word counts) is not particularly complex for a given observation. Clustering of results is also likely to be somewhat "explainable," in that the concept that clusters are based on similarity is relatively easily understood, and that (if the model is working well) the similarities are intuitively evident. Requirements for Explainability and Interpretability are lower than in some other applications, where judgments are being made that affect people's lives or finances directly. We also contend that classifying by type of discourse is likely to prove less contentious than summarizing (and possibly misrepresenting) content - though this assertion may prove to be incorrect!

Our 3rd party NLP models are not highly customized, are widely used and trusted, and are not inherently any more or less opaque than they are in other applications. For 3rd party dependencies (i.e., the language model) to move into operation, each must be as thoroughly vetted for reliability and accuracy as possible. To the extent that NLP models are not particularly transparent, we have some issues, but these are relatively minor, and if improved methods of interpreting and explaining these common NLP models are developed, these can be applied to the Gradient model.

Given the objectives, motives, and nature of the Gradient project it is imperative that all levels of the algorithmic methods are as transparently explainable and interpretable as technically possible, however, reliance on transfer learning on existing language models is the least transparent of our project's methods. Therefore, the user's of Gradient should be instructed in the workings of the model and the degree to which various elements can be understood. Appendix 1 presents a summary that could be used in training users, and seek to guide them so they know which components are from an AI, and whether and how they may be interpreted.

We have intentionally opted not to utilize closed-source or proprietary tools, platforms, or libraries in our enabling technologies as a design decision.

Operational Processes

Evaluate the model outputs in the context of the relative representation of the Twittersphere to the whole population. Conclusions may vary or carry different weight if we understand this context and any shifts therein, and awareness of any shifts in demographics may lead to insight into potential for bias. As mentioned previously, periodic generation of fresh labeled data, to validate performance of sub-models can be used to detect model drift which could result in unwanted bias. Bias might not be detected directly, because the lack of reference labels for the output makes it difficult to do so, but the model outputs - communities identified, their size and composition, and their discourse in response to various topics - can be tracked over time, and

changes can be identified.

These changes may be a reflection of underlying reality, but they may also be an indication of model bias - some change in demographics or attitude or communication style causing the model to no longer perform so well - thus, changes should be monitored and some set - either the most egregious, or random, or some combination of both, subjected to more strenuous review. If bias is confirmed, root cause analysis should be performed, corrections designed and implemented, and lessons learned documented.

Even at large scales (perhaps even more so) models need some regular process of "ground-truthing" output. Models, especially those relying on transfer learning, need frequent and regular re-training and fine-tuning to adequately capture dynamical phenomena. Any behavioral or language modeling is trying to capture the current state of what is, inherently and endemically, a complex adaptive system. Given the nature of the project's subject, any identification by the model of potentially anomalous or undesirable content should be assessed manually by human oversight before any intervention, mitigation, or action is taken.

Conclusions

Word choice and statement framing are strong psychometric indicators, particularly in the often polarizing spheres of political debate. Rather than summarize the textual content of tweets, our approach is identification of discourse types *across* political categories. We believe that identifying the functional parts of speech provides more information towards the CONOPS for Gradient than extractive or generative text summarization. The primary goal is identifying problematic forms of discourse, suggested when there is a particular alignment along othering, affinity, conviction, and sentiment dimensions.

Pronoun usage and sentiment provided the strongest indicators along these dimensions, and discourse of concern occurs where the strength of commitment to a statement precludes compromise or debate. Particularly problematic are absolutive statements combined with strong affinity, conviction, and othering. By examining the nature of the speech acts directly, these problem areas may be identified independently of more bias-laden dimensions such as political affiliation or specific statement content. It is the *nature* of the discussion and the terms we use to describe and define ourselves and others that lead to the sort of incivility that is increasingly a part of our public interactions. In finding that discourse types are only loosely coupled to political typology, we are encouraged that this indicates people of ration and reason exist across the political spectrum, and that productive public debate may in fact be possible on social media platforms. Further work will need to be done (and may have been by other teams) to determine whether these discussions are taking place across party lines.

Future research, with more refined linguistic dimensions and co-reference stance analysis, could confirm whether certain types of discourse are indeed positively associated with mature debate and sincere exchange of ideas or if certain individuals are strongly disposed to communicate in particular styles. It remains unknown if the discourse styles we found tend to be exhibited erratically, with users shifting styles as the mood takes them. Our results so far have generated more questions than they have answers!

References Cited

Adamic, L., & Nata. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog.

Allahyari, M. et al. 2017. Text Summarization Techniques: A Brief Survey. International Journal of Advanced Computer Science and Applications 8(10). Available at: http://thesai.org/Publications/ViewPaper?Volume=8&Issue=10&Code=ijacsa&SerialNo=52.

Allcott, H., Gentzkow, M., & Yu, C. 2019. Trends in the diffusion of misinformation on social media. Research and Politics 6(2).

Bakshy, E., Messing, S., & Adamic, L.A. 2015. Exposure to ideologically diverse news and opinion on Facebook. Science 348(6239): 1130–1132. Available at: https://arxiv.org/abs/9809069v1.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Findings of the association for computational linguistics: EMNLP 2020, 1644–1650. Stroudsburg, PA, USA: Association for Computational Linguistics Available at: https://www.aclweb.org/anthology/2020.findings-emnlp.148.

Bicchieri, C., & Dimant, E. 2019. Nudging with care: The risks and benefits of social information. Public Choice 401(0123456789): 1–29.

Biemann, C. 2006. Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In D. Radev & R. Mihalcea (eds) Proceedings of the first workshop on graph based methods for natural language processing, 73–80. Association for Computational Linguistics

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10): P10008. Available at: https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008.

Boldi, P., & Vigna, S. 2014. Axioms for centrality. Internet Mathematics 10(3-4): 222–262. Available at: https://arxiv.org/abs/1308.2140.

Chatila, R., & Havens, J.C. 2019. The IEEE global initiative on ethics of autonomous and intelligent systems. Intelligent Systems, Control and Automation: Science and Engineering 95: 11–16.

Clauset, A., Newman, M.E.J., & Moore, C. 2004. Finding community structure in very large networks. Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics 70(6): 6. Available at: https://arxiv.org/abs/0408187.

Congressional Research Service. 2021. Social Media: Misinformation and Content Moderation Issues for Congress. : 1–32.

Dervin, F. 2015. Discourses of othering. : 1–9.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (Mlm). Available at: https://arxiv.org/abs/1810.04805.

Dudău, D.P., & Sava, F.A. 2021. Performing Multilingual Analysis With Linguistic Inquiry and Word Count 2015 (LIWC2015). An Equivalence Study of Four Languages. Frontiers in Psychology 12(July):

1–18. Available at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.570568/full.

Filimonova, E. 2005. Clusivity : Typology and case studies of inclusive-exclusive distinction. Amsterdam Philadelphia, Pa: J. Benjamins Pub.

Horvath, S., & Dong, J. 2008. Geometric interpretation of gene coexpression network analysis. PLoS Computational Biology 4(8): 24–26.

Íñigo-Mora, I. 2004. On the use of the personal pronoun we in communities. Journal of Language and Politics 3(1): 27–52. Available at: http://www.jbe-platform.com/content/journals/10.1075/jlp.3.1.05ini.

INTEL.gov. Artificial intelligence ethics framework for the intelligence community. Available at: https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community.

Jones, K.S. 1994. Natural Language Processing: A Historical Review. In Current issues in computational linguistics: In honour of don walker, 3–16. Dordrecht: Springer Netherlands

Joshi, A.J. 1991. Natural language processing. Science 253(5025): 1242–1249. Available at: http://www.jstor.org/stable/2879169.

Karande, H., Walambe, R., Benjamin, V., Kotecha, K., & Raghu, T. 2021. Stance detection with BERT embeddings for credibility analysis of information on social media. PeerJ Computer Science 7: e467. Available at: https://peerj.com/articles/cs-467.

Kocaman, V., & Talby, D. 2021. Spark NLP: Natural language understanding at scale. Software Impacts: 100058. Available at: https://www.sciencedirect.com/science/article/pii/S2665963821000063.

Kratzke, N. 2017. The #BTW17 twitter dataset–recorded tweets of the federal election campaigns of 2017 for the 19th German Bundestag. Data 2(4).

Krejzl, P. 2018. Stance detection and summarization in social networks PhD Study Report Stance detection and summarization in social networks. PhD Study Report. University of West Bohemia in Pilsen. Available at: http://hdl.handle.net/11025/35991.

Langfelder, P., & Horvath, S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics 9(1): 559. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559.

Li, A., & Horvath, S. 2007. Network neighborhood analysis with the multi-node topological overlap measure. Bioinformatics 23(2): 222–231.

MLOps. MLOps principles. Available at: https://ml-ops.org/content/mlops-principles.

Moskal, B. 2018. Excluding exclusively the exclusive: Suppletion patterns in clusivity. Glossa: a journal of general linguistics 3(1): 130.

Mosleh, M., Martel, C., Eckles, D., & Rand, D. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In Proceedings of the 2021 CHI conference on human factors in computing systems, 1–13. New York, NY, USA: ACM

Newman, M.E.J. 2006. Modularity and community structure in networks. Proceedings of the National

Academy of Sciences 103(23): 8577–8582. Available at: https://arxiv.org/abs/0602124.

Newman, M.E.J., & Girvan, M. 2003. Mixing Patterns and Community Structure in Networks. In R. Pastor-Satorras, M. Rubi, & A. Diaz-Guilera (eds) Statistical mechanics of complex networks, 66–87. Springer Available at: https://arxiv.org/abs/0210146.

Omnicore. Twitter by the numbers: Stats, demographics & fun facts. Available at: https://www.omnicoreagency.com/twitter-statistics/.

Pak, A., & Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010: 1320–1326.

Pennebaker, J. 2011. The secret life of pronouns : What our words say about us. New York: Bloomsbury Press.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. 2015. The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin: 1–22. Available at: http://www.liwc.net/LIWC2007LanguageManual.pdf.

Ribeiro, M.H., Calais, P.H., Santos, Y.A., Almeida, V.A.F., & Meira, W. 2018. Characterizing and detecting hateful users on twitter. 12th International AAAI Conference on Web and Social Media, ICWSM 2018: 676–679. Available at: https://arxiv.org/abs/1803.08977.

Riedel, B., Augenstein, I., Spithourakis, G.P., & Riedel, S. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. : 1–6. Available at: http://arxiv.org/abs/1707.03264.

Rosvall, M., & Bergstrom, C.T. 2008. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America 105(4): 1118–1123. Available at: https://arxiv.org/abs/0707.0609.

Ruths, D. 2019. The misinformation machine. Science 363(6425): 348–348. Available at: https://www.sciencemag.org/lookup/doi/10.1126/science.aaw1315.

Saba, W.S. 2007. Language, logic and ontology: Uncovering the structure of commonsense knowledge. International Journal of Human Computer Studies 65(7): 610–623. Available at: https://arxiv.org/abs/0610067.

Saba, W.S. 2020. Language and its Commonsense : Where Formal Semantics Went Wrong, and Where it Can (and Should) Go. Journal of Knowledge Structures & Systems 1(November): 40–62. Available at: https://philpapers.org/rec/SABLAI.

Shao, C. et al. 2017. The spread of low-credibility content by social bots. Nature Communications (2018). Available at: http://arxiv.org/abs/1707.07592.

Singh, A., Blanco, E., & Jin, W. 2019. Incorporating Emoji Descriptions Improves Tweet Classification. In Proceedings of the 2019 conference of the north, 2096–2101. Stroudsburg, PA, USA: Association for Computational Linguistics Available at: http://aclweb.org/anthology/N19-1214.

Sterelny, K. 2007. Social intelligence, human intelligence and niche construction. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 362(1480): 719–30. Available at: http://rstb.royalsocietypublishing.org/content/362/1480/719.short

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2346527&tool=pmcentrez&rendertype=abstract.

Tang, M., Jin, J., Liu, Y., Li, C., & Zhang, W. 2019. Integrating Topic, Sentiment, and Syntax for Modeling Online Reviews: A Topic Model Approach. Journal of Computing and Information Science in Engineering 19(1): 137–144.

Tausczik, Y.R., & Pennebaker, J.W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology 29(1): 24–54.

Traag, V.A., Waltman, L., & Eck, N.J. van. 2019. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports 9(1): 1–12. Available at: https://arxiv.org/abs/1810.08473.

Vaswani, A. et al. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems 2017-Decem(Nips): 5999–6009. Available at: https://arxiv.org/abs/1706.03762.

Wallach, H.M., Murray, I., Salakhutdinov, R., & Mimno, D. 2009. Evaluation methods for topic models. In Proceedings of the 26th annual international conference on machine learning - ICML '09, 1–8. Montreal, Quebec, Canada: ACM Press Available at: http://portal.acm.org/citation.cfm?doid=1553374.1553515.

Wojcik, S., & Hughes, A. 2019. Sizing Up Twitter Users. : 22. Available at: https://www.pewinternet.org/wp-content/uploads/sites/9/2019/04/twitter_opinions_4_18_final_clean.pdf.

Yip, A.M., & Horvath, S. 2005. The Generalized Topological Overlap Matrix For Detecting Modules in Gene Networks. : 1–19.

Yip, A.M., & Horvath, S. 2007. Gene network interconnectedness and the generalized topological overlap measure. BMC bioinformatics 8: 22. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1797055&tool=pmcentrez&rendertype=abstract.

Zotova, E. 2019. Automatic Stance Detection on Political Discourse in Twitter. Master's Thesis. University of the Basque Country UP- V/EHU.

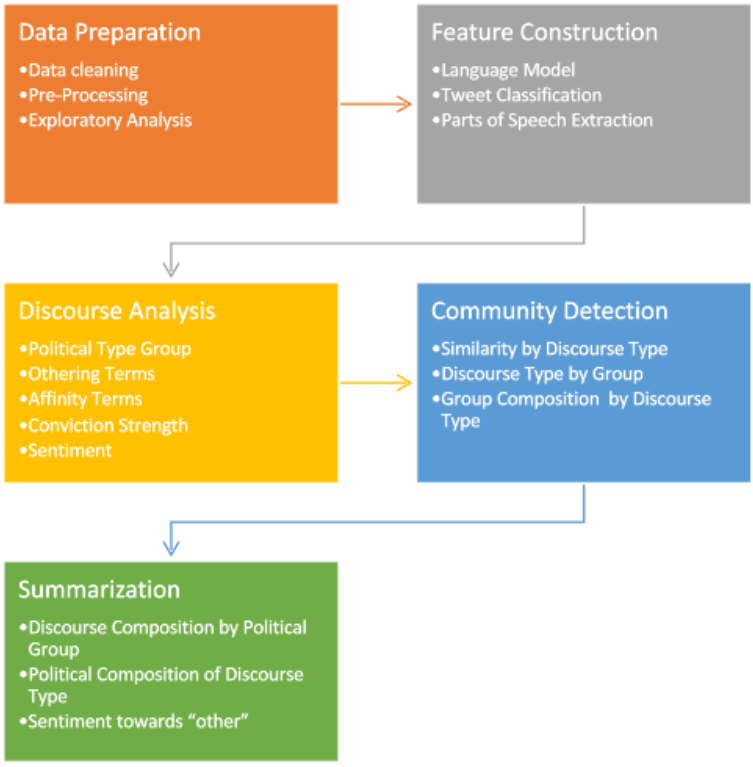
Appendices

1. Explainability and Interpretability: User Training

The following material could be used to familiarize users with the working of our Gradient model, in the interests of helping them know what elements come from AI, and the degree to which those elements are "black boxes" and are Explainable or Interpretable (E/I). This helps address the issue "How are outputs marked to clearly show that they came from an AI?"

Piper Gradient analyzes text from tweets on subjects of interest and attempts to understand how different modes of discourse are being used, and by which communities. [Refer to training on discourse modes if need be].
Machine Learning, or AI, is used in several steps of the process. For example, political typologies are inferred from the tweets (and their associated user descriptions) using a model that

infers meaning from tweets using advanced NLP methods which are deeply complex (i.e. deep learning) and trained of tens of millions of examples, and try to associate typological labels to these tweets based on examples classified by humans. The process of both solidarity in the new tweets to the training tweets can be explained by a study of deep learning methods, but the detailed process by which the AI any two tweets are identified as maximally similar is obscure. However, if it has done a good job, the results should seem intuitively "right".
Similar processes are performed for analyzing sentiment as positive or negative.
More intuitively obvious methods are used to score discourse styles based on the presence of certain keywords deemed to be important indicators in the underlying linguistic theories.
Community detection (collections with similar styles of discourse) is done by using assigning similarities of the scores for each of several styles as connections between tweets, and the number of connections (and absence of others) helps algorithmically determine which tweets are most closely and related while distinct from others - a relatively transparent process known as network analysis.
Summarization is a relatively straightforward process entailing how the detected clusters have features that overlap or differ.



Element	Description	E/I	Explanation
Language Model / Tweet	Deep learning models seeking to understand natural	Very low	Models scoring methods are based on thousands or

Classification	language and predict typology from data labelled by humans		millions of calculations
Sentiment Analysis	Deep learning models seeking to understand natural language and assess sentiment	Very low	Models scoring methods are based on thousands or millions of calculations
Othering (Clusivity) Affinity Conviction Absolutism	Based on linguistic theories and presence of specific keywords	High	User should be able to assess scores for specific observations and arrive at same result
Community Detection	Network (graph) analysis	Moderately high	Results will be less immediately obvious to users but similarities can be seen, particularly after summarization
Summarization	Heatmap highlighting similarities/differences	High	Collectively the differences can be overwhelming at first but teased out by element-wise comparison, doable by humans with reasonable repeatability

2. Clusters and Piper Typology

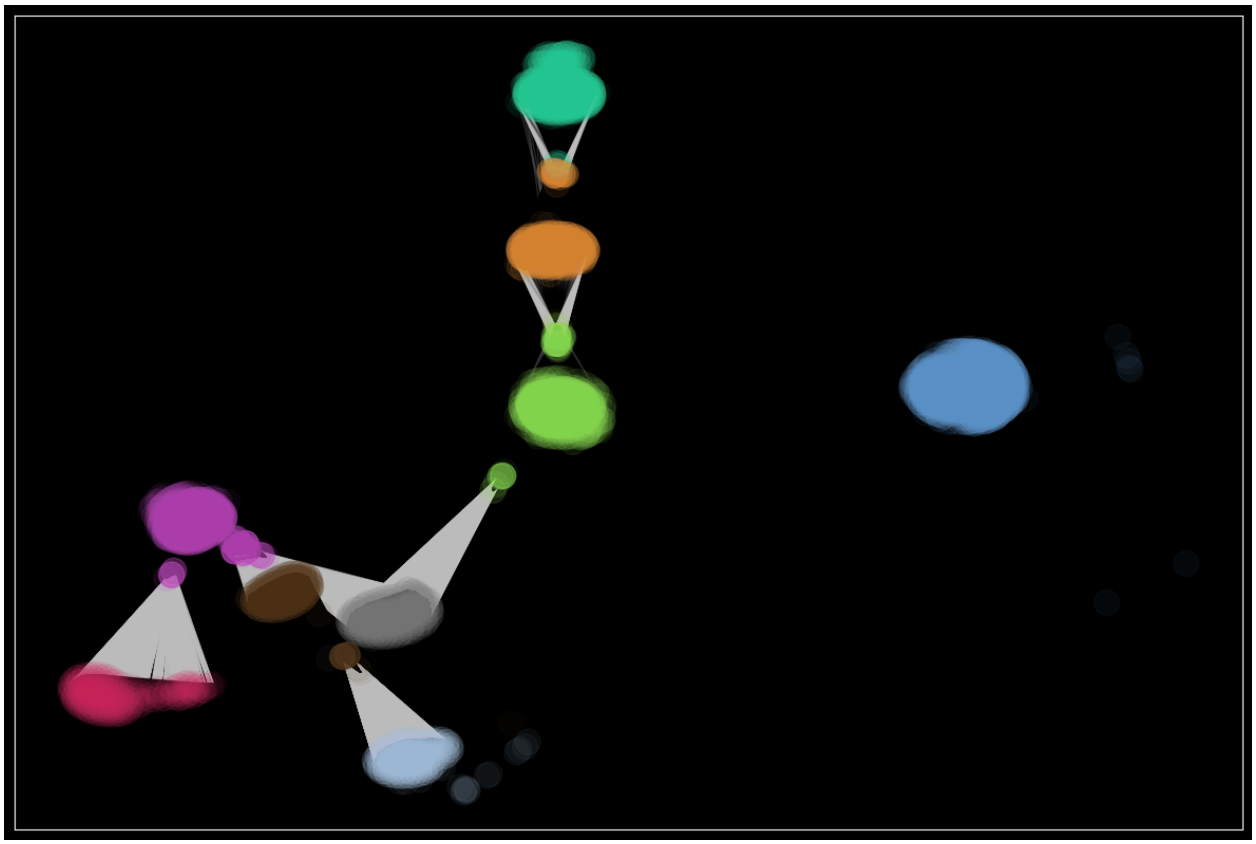


Figure: Discourse graph network showing identified communities and their connections.

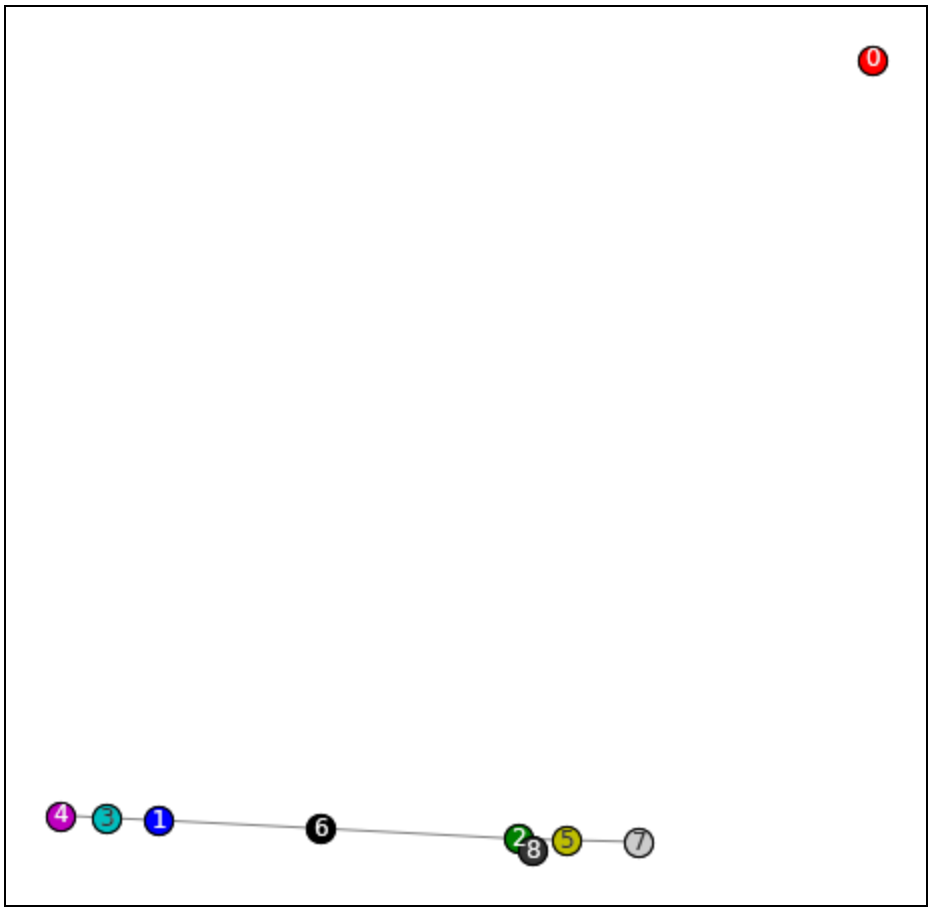


Figure: Relationship between detected communities. Note that the community labeled “0” is a separate and isolated sub-graph. This community is identified as “1” in subsequent summary, and is the largest community of the network.

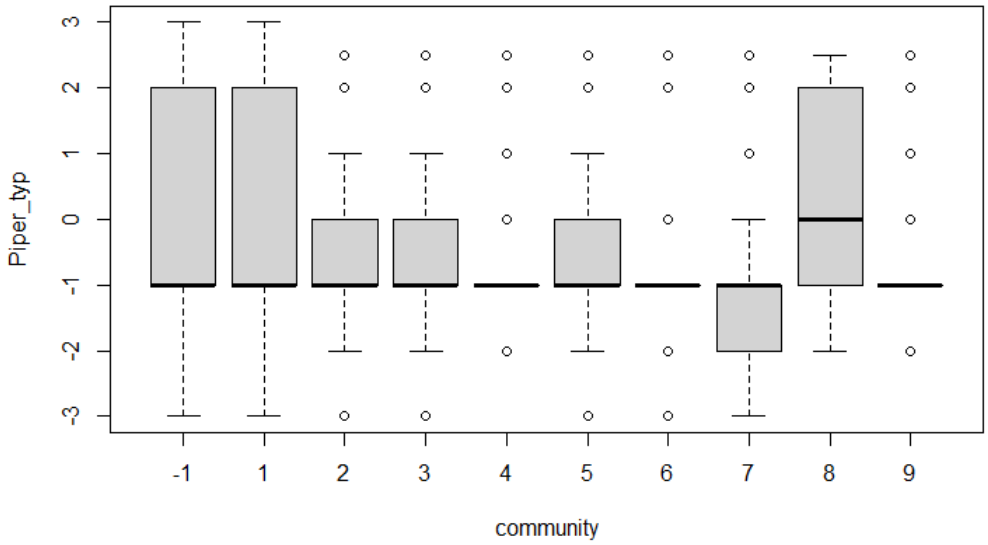
Heatmap of the derived clusters and their distinguishing features:

		Community										
	Features	-1	5	1	8	6	4	9	7	3	2	Overall
Discourse	incl_affil	0.03	1.43	0.14	0.01	0.00	1.11	1.03	0.00	0.05	0.12	0.21
	incl_assoc	0.04	0.12	1.56	0.04	0.00	0.06	0.00	0.00	0.08	0.11	0.30
	excl_affil	0.02	0.02	0.11	0.01	0.01	1.02	0.00	1.07	0.04	2.01	0.28
	excl_assoc	0.03	0.04	0.12	0.00	1.02	0.00	1.03	1.07	2.01	0.09	0.28
	abs_terms	0.02	0.12	0.12	1.10	1.11	0.01	0.00	0.00	0.17	0.13	0.12
	Piper_typ	0.06	-0.27	-0.17	0.15	-0.94	-0.65	-0.87	-1.00	-0.27	-0.21	-0.16
	cdf_neg	0.36	0.38	0.33	0.33	0.57	0.51	0.14	0.65	0.53	0.47	0.40
	cdf_neu	0.46	0.42	0.35	0.45	0.35	0.43	0.39	0.31	0.36	0.37	0.40
	cdf_pos	0.18	0.19	0.32	0.22	0.09	0.06	0.47	0.04	0.11	0.16	0.19
	is_reply	0.12	0.26	0.19	0.28	0.09	0.08	0.08	0.07	0.24	0.43	0.17
	is_retweet	0.70	0.55	0.66	0.53	0.88	0.88	0.86	0.88	0.59	0.35	0.67
	unique ratio	0.35	0.56	0.41	0.59	0.14	0.13	0.19	0.15	0.51	0.75	0.39
	favorite_count	3.0	2.6	3.0	2.1	0.9	3.0	5.6	3.8	3.0	1.4	2.8
	Tweet total	7737	921	2694	368	539	939	308	395	1357	1380	16638
	Unique tweets	2691	513	1097	217	75	125	59	60	688	1039	6564
	Unique users	4285	644	1624	286	397	656	258	332	895	832	

The truth table for clusters (on the left) and Piper's typology is:

	Typology							
	-3	-2	-1	0	1	2	2.5	3
-1	22	864	3054	1319	102	1521	841	14
1	14	432	1012	518	25	573	118	2
2	3	171	576	317	14	257	42	0
3	6	204	549	300	17	232	49	0
4	0	58	657	143	6	71	4	0
5	2	127	378	227	14	127	46	0
6	2	70	414	30	0	18	5	0
7	2	166	139	49	1	30	8	0
8	0	61	93	78	10	111	15	0
9	0	63	188	34	2	17	4	0

A boxplot of these distributions helps visualize them:



Analysis of Variance (AOV) was run on the additional (non-discourse) features not used in clustering to identify whether means of these were hypothetically the same; rejection of this hypothesis (low p-value) implies the means are different for at least one cluster. The results show extremely low p-values, supporting the hypothesis that these features can be used to identify some clusters, or that they have differing characteristics.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Piper_typ	1	3218	3218	451.73	< 2e-16 ***
cdf_neg	1	1396	1396	196.00	< 2e-16 ***
cdf_pos	1	294	294	41.33	1.32e-10 ***
is_reply	1	532	532	74.67	< 2e-16 ***
retweet_count	1	804	804	112.90	< 2e-16 ***
Residuals	1	6632	118495	7	

2(a). Representative tweets

These tweets were the most significant in each cluster, as determined by pagerank. Note that the user and links are shown here as they are before being replaced with generic “@user” and “http” tokens for the RoBERTa transformers.

Cluster	Tweet
-1	RT @TinResistAgain: @tedcruz The GOP Is Voting Against Its Base Republicans are making a risky bet by opposing Biden's infrastructure plan. https://t.co/1JNiSql38k
1	RT @TheLastWord: .@SenMarkey tells @Lawrence that President Joe Biden is "right on message" tying jobs to climate change: "The kind of climate bill

	that Joe Biden is proposing is going to be the greatest, blue-collar job creation engine in our country in two generations." https://t.co/MDRzoWuYHd https://t.co/2p7pk1CbcY
2	@POTUS @VP you both know that some GOP leaders are rebelling against you bc they lost the election in 2020. They know that you both want more ppl to be vaccinated and you both are promoting your infrastructure bill @cnn @abc @SpeakerPelosi @SenSchumer @BarackObama @MichelleObama
3	@Zomblerone @StopaNotFoot I just don't care if they ask kids about their opinions on a topic of political discourse that they are the center of. If the Fox guy asked him about Biden's infrastructure bill and deficit spending, then yea, that'd be weird as fuck lmao
4	@WeirderIsGood @POTUS @DonaldJTrumpJr It's the same ole tweets day in and day out, everyday, get your shots, jobs are coming, pass my pork filled infrastructure bill, pass my shitty jobs bill, and on and on and on and on and on and on.....
5	RT @onlyfansofOPP: @oren_cass I really don't know how daycare is not infrastructure. I understand that people who don't have children (🙋🏻) don't wanna fit the bill, but to everyone I know with children daycare is a tippy top priority and expense. Go write your books, but real world problems need solutions.
6	RT @SenatorCardin: Clean water is not a partisan issue. Every resident in every community has a right to expect that the water coming from their tap is safe and affordable to drink and dispose of. Clean water produces better public health outcomes and good jobs. https://t.co/hHttf5QCY1
7	RT @CorporalBen: America, do you realize that under Biden's "Infrastructure Bill" Anyone who works on Any of the projects outlined in this bill will be working for the Federals? The goal is to have everyone dependent on the Government for their livelihood; either work or be on Entitlements.
8	@morningmika The majority of Americans approve of Biden's infrastructure bill from all parties. That's the only bipartisanship that should count. Stop with the bipartisan questions, the people have spoken

9	RT @SolidRedPeon: Senate passes $35B water bill, preview of 'infrastructure' spending fight https://t.co/lxGyW2GHB5 It's not my responsibility to replace Flints rusty pipes. It's theirs. The Fed has no such Constitutional authority to charge the bill to me yet there's the GOP saying they do.

3. Sentiment Performance and Examples of Error

Comparison of sentiment classification performance.

Model	Agreed +/-	Neutral	Opposite
Textblob	8	5	5
vader	8	4	6
distilbert	12	0	6
RoBERTa	18	0	0

Examples in which non-RoBERTa classifiers erred included – Vader classifies

> Biden pushes another preposterous lie about his “infrastructure” bill, saying it will create 16,000,000 jobs

and

> Joe Biden falsely claims his “infrastructure” bill will create 16 million jobs

as slightly positive (Textblob also classified the former as neutral polarity). Textblob, vader and distilbert all classified

> Republicans are against: - The COVID Relief Bill - $15 minimum wage - Billionaires paying taxes - Infrastructure - Voting rights - Gun reform - Elderly care - Health care - Child care - Equality But they’re just fine with all of the insurrection and sex trafficking

as positive.

4. Linguistic Association of Selected Pronouns

Part of Speech	Pronouns	Affinity	Clusivity	Description
1st person, singular	I, Me	Affiliation	Exclusive	Self-preferential statement of conviction, active
1st person, plural	We, Us	Association	Inclusive	Collective conviction, may exclude self, passive
3rd person, plural	Them, They, Those	Association	Inclusive	Collective assignment of other’s belief, may exclude audience, passive
2nd person (sing./plur)	You, Your, Yours	Affiliation	Exclusive	Individual assignment of other’s belief, inclusive of audience, active

Indef.	All, None, Every, Each	Affiliation	Dependent	Absolutive statement towards referent, potential logical fallacy, polemical

5. Piper Typology classification accuracy

Truth table for Piper Typology classification:

			Predicted					
Actual	**-3**	**-2**	**-1**	**0**	**1**	**2**	**2.5**	**3**
-3	0	0	4	2	1	3	3	0
-2	0	94	18	6	0	1	1	0
-1	0	2	233	19	1	3	6	0
0	0	2	7	39	1	5	0	0
1	0	0	1	0	16	0	0	0
2	0	1	8	4	0	43	2	0
2.5	0	0	4	2	0	5	107	0
3	0	0	2	1	0	0	3	0

```
Piper Typology Prediction Error
MAE: 0.3630769230769231
MSE: 1.0284615384615385
RMSE: 1.014130927672329
```