Predicting Intergroup Speech in Reddit Football Comments

Matianyu Zang
Email: matianyu_zang@brown.edu

***Abstract*—In this DREU summer experience, I am honored to be advised by professor Jessy Junyi Li at University of Austin at Texas and mentored by her Ph.D. student Venkata S Govindarajan. This project aims to understand the task of intergroup speech prediction in social comments. We scraped Football comments from Reddits for all NFL Football teams. Each team competes on average 18 times per season. For each game, a game post, sometimes also pre-game posts and post-game post, is submitted in the Team's subreddit, where fans exchange game messages. We downloaded all comments from last two years, obtaining more than 5 million comments. As each comment belongs to a subreddit/team, we hypothesize that the subreddit team is the ingroup speech, the opponent team is the outgroup speech, and all 30 remaining teams are the third party. Then, we randomly sample comments, mask team names, and instruct both gpt4 and lab members to categorize the masked entity. Both human and the machine score an accuracy between 0.4 and 0.5, which is slightly above the chance level but far from good. We perceive the difficulty of the task and will adjust the task design, possibly providing more contexts or recruiting people with domain knowledge, in the future.**

***Index Terms*—Intergroup Speech, Language Models**

I. Introduction and Related Work

In-group and out-group analysis is a concept rooted in social psychology that is now frequently applied to various fields including linguistics, social networks, and discourse analysis. This method examines how individuals classify themselves and others into either 'in-groups', 'out-groups', or 'third-party' based on perceived membership in a social, cultural, linguistic, or ideological group. In this project, we are focusing on the in-group/outgroup analysis on the nfl football game discussion comments, aiming to dig out the cues that characterize in-group speech versus outgroup speech. The work preceding this project concentrates on how emotional cues reflect the group speech [1]. The author creates a dataset of English tweets by US Congress members. In this study, we take advantage of the NFL Football game Reddit comments to further study the intergroup bias.

II. Method

This project is in the starting phase. Our current goal is to understand the difficulty level of predicting ingroup/outgroup/third-party tag for masked entities in the social comments. We then compare human and machine's task performances and train language models that are specialized on the task. Targeting those goals, we plan the following steps.

A. Data Collection

We focus on the Football game comments on Reddit. Each of the 32 NFL Football team has their own Subreddit, where users game day thread, optionally pre-game thread and post-game thread, as a discussion forum for each game that their supporting team competes. We leverage the praw [2], a python library designed for quick Reddit scraping, to extract game posts in the past two years. On average, each team has 19 games per season.

B. Annotation Task

After collecting all the data, we sample comments from the data pool and test the difficulty level on both the language model and humans.

C. Model Training

We have not proceeded to this step, but depending on the research question we want to dive into, we may training language models that outwit human on group tagging task in the future.

III. Results and Analyses

Data Collection:

It takes us about four weeks to finish scraping all Reddit game comments in the past two years. 12 of the teams, including panthers, 49ers, jets, etc. have a common post manager, 'nfl_gdt_bot', who is in charge of all game thread posts. We thus extract all comments from its submissions. For the remaining 20 teams, there's no universal way of scraping all relevant comments, so we take advantage of the keywords, look for channels, and manually pick out relevant posts. Since we want to know how fans from competing teams interpret the same game, we pair game posts from competing teams together. The following are the numbers of posts. In addition, in our pilot experiments, we realize that sufficient amount of context is needed for the annotation task, so we filter out those comments with fewer than 5 words. All posts sum up to 1360 posts and 5,019,538 usable comments.

Since every comment belongs to a certain subreddit, we make two assumptions. First, we hypothesize that users who post comments in a thread under team A is a fan of team A. Second, the mention of team A should be an ingroup speech while its opponent in this game is a outgroup speech;

	pregame	game day	post game
all	262	1083	1015
paired	28	527	461

TABLE I
NUMBER OF (PAIRED) POSTS

all remaining NFL Football teams are third-party. Given this presupposition, we are able to obtain a bunch of gold labels for team name mentions. Table II showcases the number of in-group/out-group/third-party instances/mentions in the dataset. 'mention' is the total number of comments that include some category of masked entities, and 'instance' is the number of total occurances of some type of masked entities. To simplify the task, we first make use of only comments with one type of masked entities in it, so we have 463,742 comments at hand.

Finally, we leverage the co-reference resolution to match pronouns like they/them/their/theirs and we/us/our/ours with team names. Hence, more gold labels come naturally. To retrieve the availability of pronouns from the dataset, we count the number of pronouns. Table III demonstrates the number. Based on our current knowledge, in top comments (those that directly respond to the post), 'we' refers to the fan's team, and 'they' refers to the opponent's team by conventions. This is not necessarily true concerning the non-top comments (those that respond to some other comments) as the focus of the discussion might has been shifted to other games/teams.

Annotation Task:

In this phase of the project, the main goal is to understand the annotation task. We tested it on gpt4, one of the most powerful language models up to date. Meanwhile, we plan to recruit human annotators to work on the tasks. As a pilot testing, our lab members first experimented on the annotation task.

Once we decide the entity to mask, we devised two types of tasks: one masking the team names with [ENT] and one only highlighting the target entity.
To instruct gpt4 to annotate the entities as expected, we prepare prompts for both the masked and highlighted task following the instruction by openAI. The prompt below is for the masked task:

> We are interested in how the writer of a reddit comment feels towards/in connection with the people they're talking about. You will be asked to annotate for masked entities in a comment - examples are shown below the instructions.
> ## Instructions
> Read the text carefully. We are interested in how sports fans online talk about the team they support, or players in the team they support, versus opposing teams/players.
> You will be reading comments by sports fans before/during/after a game between their team and an opponent. Comments are usually about one of the teams, or a player from the teams, and we want to understand how fans talk about them.
> We have replaced the mentioned team name in a comment with [ENT]. Your task is to guess if the [ENT] the commenter is talking about refers to the fan's team or the opponent based on the rest of the comment.
> Sometimes it might be the case that [ENT] can refer to either the team the speaker supports or an opponent. You can choose the either option if you think [ENT] is ambiguous. But first, make your strongest guess if the reference is to their team or the opponent.
> If there are multiple masked words in a sentence, it's possible they refer to different groups. Therefore, make sure to analyze each one individually based on its context.
> As output, please copy the whole sentence and replace [ENT] with one of [IN] if it refers to fan's team, [OUT] if it refers to the opponent, or [EITHER] if either works.
> Note: please annotate all and only [ENT]s!
> Here are examples of each
> fan's team: Also I miss [ENT], [ENT] was always great on 3rd downs Annotation: Also I miss [IN], [IN] was always great on 3rd downs
> opponent: Everything going [ENT] way so far....[ENT] are fucking going to win this game arent they Annotation: Everything going [OUT] way so far....[OUT] are fucking going to win this game arent they
> either: Uh [ENT], what are you doing? Annotation: Uh [EITHER], what are you doing?
> Now annotate this comment using the format above, using only the 3 labels defined above in your answers, and following all instructions given above. Return only the answer dictionary object.
> (Provide actual comments here.)

Here we omit the prompt for highlighted task as it is almost the identical except the annotation samples. The following is an example of highlighted comment.

> fan's team: Also I miss [Patrick], [he] was always great on 3rd downs Annotation: Also I miss [IN], [IN] was always great on 3rd downs

Gpt4 is allowed to opt an answer among fan's team, opponent's team, or either for each masked/highlighted entity. It takes the gpt4 model about 10 second to finish one comment, and we run it five times to extract the average performances. Knowing that temperature impacts the language model's prediction accuracy, we change the value from 0.1 to 1.0 to seek the best temperature. However, as following chart illustrates, the temperature is not a dominant impacting factor of the task performance. We thus defaults the temperature to 1.0. In the figure, 'acc' stands for average prediction accuracy among five

	in-group	out-group	third-party	in & out	in & third	out & third	all
mention	177301	162211	124230	17606	12832	10384	1888
instances	225591	201075	190963	/	/	/	/

TABLE II
COUNTS FOR EACH TYPE OF MASKED ENTITIES

	in-group	out-group	third-party
we-like words	17943	31634	20502
they-like words	17105	23118	8992

TABLE III
COUNTS OF PRONOUNS

runs while 'score' represents the alignment score among five predictions.

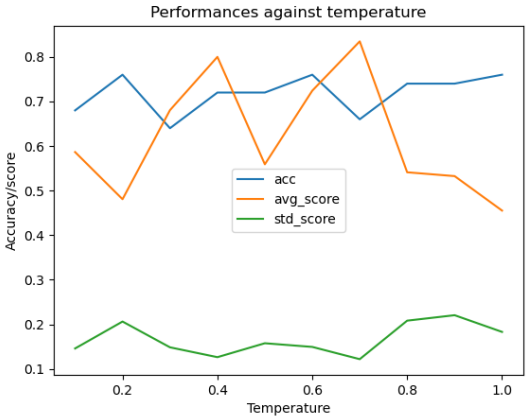

Fig. 1. Temperature v.s. Performance

In the first experiment, gpt4 achieves an accuracy of 75% and 84.38% respectively for masked and highlighted task. This experiment is biased though as the number of in-group, out-group, and third-party entities is not balanced. We then fix this data imbalance issue and sample 10 new comments from the dataset. This time the gpt4 prediction accuracy drops to 48.67% and 59.33% respectively for masked and unmasked tasks.

Our lab members also complete the same task for comparisons. To improve on the task design, we experimented multiple times, adjusting the UI designs and answer choices. The following picture depicts the latest web design for the annotation task. Participants can drag the slider to choose among five options: fan's team, likely fan's team, either, likely opponent's team, opponent's team. Alternatively, they may opt the checkbox if they think [ENT] refers to some third-party teams that are non-present in the game. While the answer design in the task is straightforward and comprehensive for annotators, coming up with a metric for accuracy computation is hard due to it. For simplication and reasonableness, we consider 'likely fan's team' equivalent with 'fan's team', and similarly for the 'likely opponent's team' option. The average accuracy among human participants is between 0.4 and 0.5, on par with the model's performance.

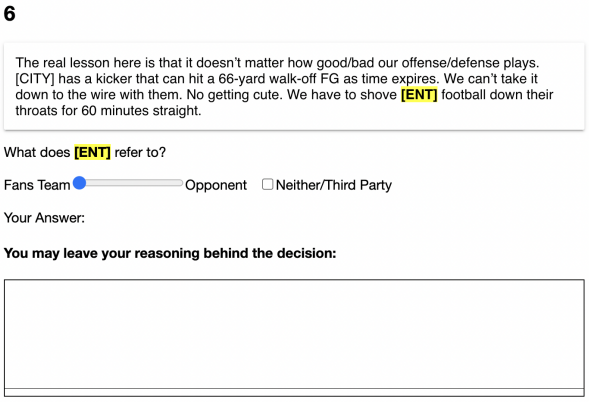

Fig. 2. An image of a galaxy

The fleiss kappa score among annotators is 0.22. This shows that the annotation task is hard for both the machine and human. The reason could be the lack of context, the absence of domain-specific knowledge, and the involvement of no-top comments. We would further modfiy the task design.

Coreference Resolution:

To incorporate more usable gold labels, we leverage the co-reference resolution to pair more pronouns with team names. We apply the model of LingMess [3]. The model takes in a raw text and outputs a list of lists. Each sublist contains multiple tuples of indices, which corresponds to the starting and ending position of tokens in the raw text. We give an example as following:

doc_key: hknbn8z
Raw texts:
The team that played today was the team I 've been expecting to pay all year . Our D - Line was getting great pressure and our secondary was hitting guys as soon as they caught the ball . This game shows us this team is not as bad as we thought it was and capable of competing with the big boys . We have had tons of opportunities to take advantage of games , most notably the Chiefs and Packers games , and when we do we can win (duh) . The game at Carolina is huge , but these last 8 games are significantly easier .
index_list:
[[17,17], [26,26], [42,42], [50,50], [63,63], [85,85],[87,87]]
[[30,30], [34,34]]
[[0,4], [39,40], [52,52]]
actual_tokens:

[['Our'], ['our'], ['us'], ['we'], ['We'], ['we'], ['we']]
[['guys'], ['they']]
[['The', 'team', 'that', 'played', 'today'], ['this', 'team'], ['it']]
####################

Take the last list for instance, [0,4] represents the tokens span from position 0 to position 4, inclusive, so it should be 'the team that played today'. Meanwhile, this refers to the same thing as other tokens in the same outer list, namely 'this team' and 'it'. To understand by what amount of the pronouns the model can pair with some team names, we randomly sample 1000 comments from the dataset and run the model on them. 481 of the comments have some coreferences detected by LingMess, among which 182 have we-like pronouns (42 out of 338 instances are paired with some team names) and 186 have they-like pronouns (187 out of 287 instances are paired with team names). In general, LingMess does not pair a decent portion of 'they'/'we' with team names, but we still plan to use this tool as we have a plentiful of comments at hand.

IV. Future Work

Some work that we plan to finish in the near future include 1) improve on the current human annotation task setup; 2) adjust the difficulty level, by providing more context information like game results or recruiting only football fan's equipped with domain knowledge, to make the annotation a reasonable task; 3) compare machine and human task performances; 4) understand how the contextual information influences the task performance; 5) train a model that specializes at telling the parties of masked/highlighted entities in social comments.

References

[1] V. S. Govindarajan, D. Beaver, K. Mahowald, and J. J. Li, "Counterfactual probing for the influence of affect and specificity on intergroup bias," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 853–12 862. [Online]. Available: https://aclanthology.org/2023.findings-acl.813

[2] [Online]. Available: https://praw.readthedocs.io/en/stable/

[3] S. Otmazgin, A. Cattan, and Y. Goldberg, "Lingmess: Linguistically informed multi expert scorers for coreference resolution," 2023.