CAN YOU CONVINCE ME?

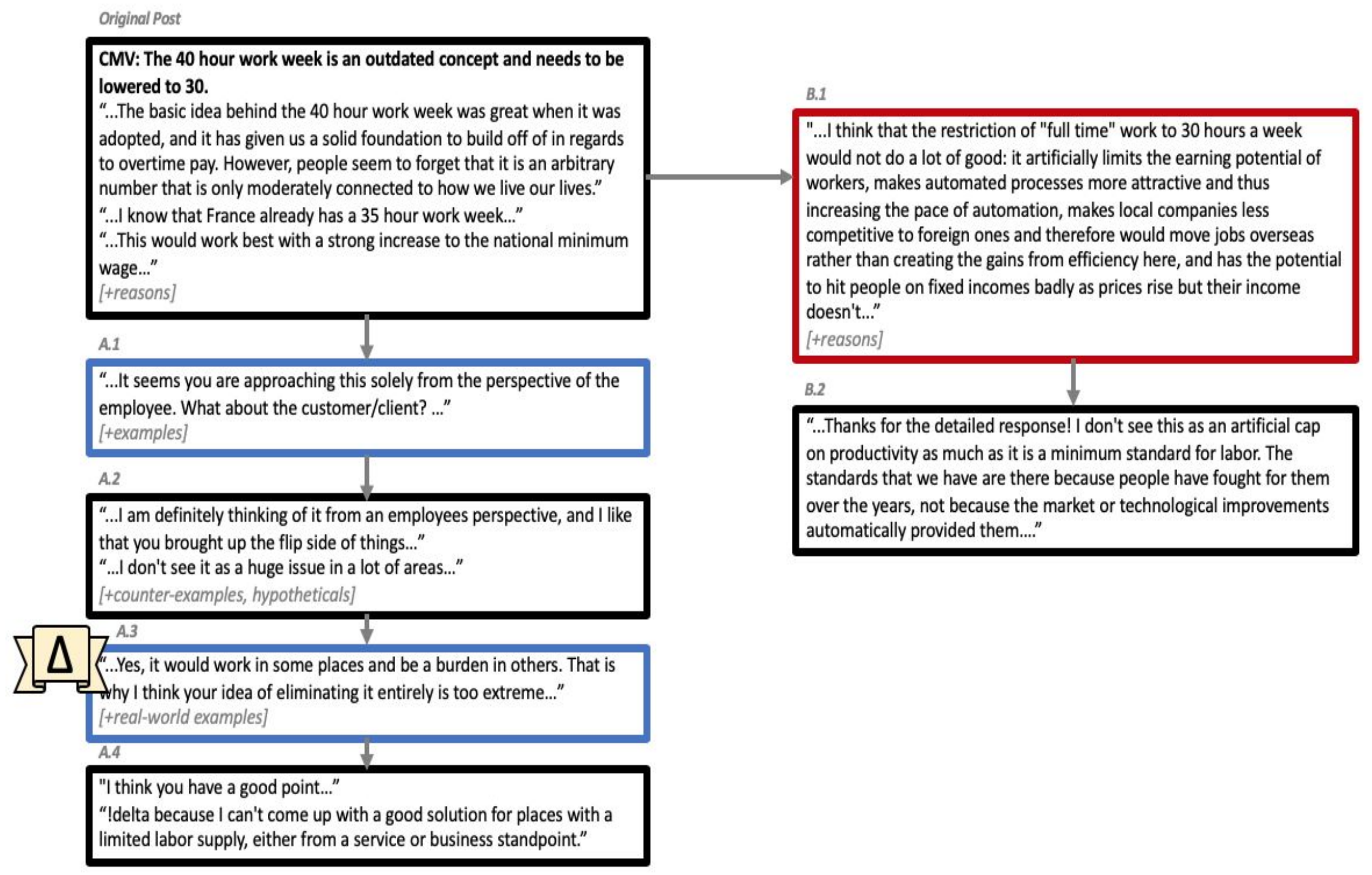
Shared internet communities and persuasion on Reddit r/ChangeMyView

Roshan Ramkeesoon, Geoffrey Li, and Deepthi Hegde | CSE 547 Final Project

INTRODUCTION

- Prior studies have shown that psycholinguistic factors make some arguments more persuasive than others [2]
- We examine whether shared communities with another person or structural context of a social network also has a measurable effect on persuasiveness using Reddit r/ChangeMyView (CMV)
- ChangeMyView is an active community on Reddit where users present their opinions and reasoning, invite others to contest them, and acknowledge when the ensuing discussions change their original views

Dynamics of r/ChangeMyView



Original Post

CMV: The 40 hour work week is an outdated concept and needs to be lowered to 30.
"...The basic idea behind the 40 hour work week was great when it was adopted, and it has given us a solid foundation to build off of in regards to overtime pay. However, people seem to forget that it is an arbitrary number that is only moderately connected to how we live our lives."
"...I know that France already has a 35 hour work week..."
"...This would work best with a strong increase to the national minimum wage..."
[+reasons]

A.1

"...It seems you are approaching this solely from the perspective of the employee. What about the customer/client? ..."
[+examples]

A.2

"...I am definitely thinking of it from an employees perspective, and I like that you brought up the flip side of things..."
"...I don't see it as a huge issue in a lot of areas..."
[+counter-examples, hypotheticals]

A.3

Δ

"...Yes, it would work in some places and be a burden in others. That is why I think your idea of eliminating it entirely is too extreme..."
[+real-world examples]

A.4

"I think you have a good point..."
"!delta because I can't come up with a good solution for places with a limited labor supply, either from a service or business standpoint."

B.1

"...I think that the restriction of "full time" work to 30 hours a week would not do a lot of good: it artificially limits the earning potential of workers, makes automated processes more attractive and thus increasing the pace of automation, makes local companies less competitive to foreign ones and therefore would move jobs overseas rather than creating the gains from efficiency here, and has the potential to hit people on fixed incomes badly as prices rise but their income doesn't..."
[+reasons]

B.2

"...Thanks for the detailed response! I don't see this as an artificial cap on productivity as much as it is a minimum standard for labor. The standards that we have are there because people have fought for them over the years, not because the market or technological improvements automatically provided them...."

Figure 1. Sample snapshot of a discussion tree on r/CMV. Colors indicate unique posters on the thread, with OP in black. Challenger in blue is awarded a Delta after some back-and-forth discussion, while the challenger in red is unsuccessful.

- From each discussion thread one positive and negative example are selected based on highest Jaccard similarity in content words used
- The balanced dataset allows us to control for the argument being made and examine other features, namely the network features of the OP-Challenger relationships
- Subreddit r/ChangeMyView is carefully moderated and has well-defined rules
- Leveraged dataset from Tan et. al. [2] for period 1/1/13 to 9/1/15:
  - 4,546 training examples and 1,047 test examples

Reddit: a Bipartite Graph

- All comments hosted on Google BigQuery
- Bipartite graph: Users → Subreddits
  - Edge if user has written in given subreddit
  - Edge weight is # comments
  - One graph for every 6 month period to allow graph to change over time
- In the time period 1/1/13 to 9/1/15:
  - ~**1.6 billion** comments
  - ~**10.5k** subreddits with > 1,000 comments
  - ~**84k** unique users
  - ~**5.6 million** edges

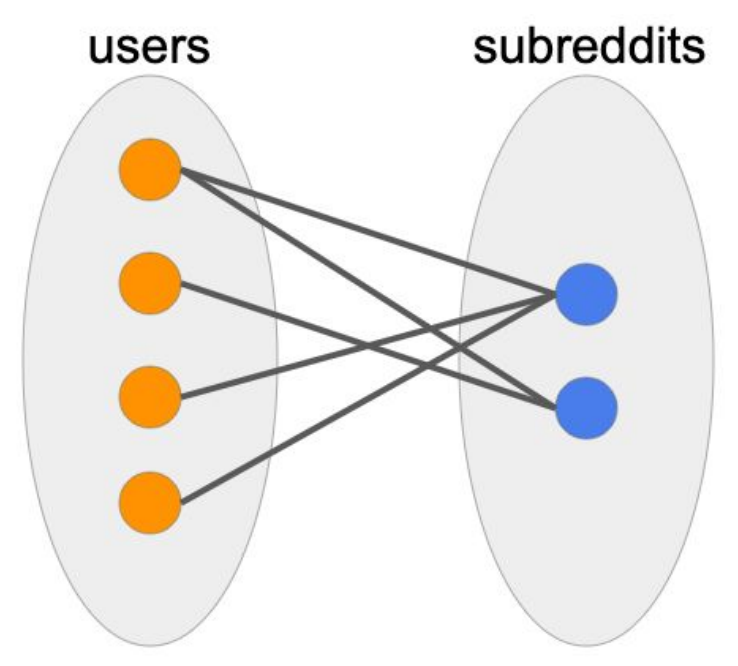


Fig. 2. Representation of bipartite graph of user interactions with subreddits.

LINGUISTIC FEATURES

- We use 5 broad categories of linguistic features -
  - Interplay between original post and counterargument
  - Word category (Articles, pronouns, links, quotations)
  - Word score (Arousal, Valence, Dominance)
  - Full argument (Flesch-Kincaid grade levels, word-entropy)
  - Markdown-formatting (bold, italics)
- We obtained 55% AUC using logistic regression
  - Our results match the results from the baseline paper [2]

USER SIMILARITY

Cosine Similarity

$$\frac{\#\ \text{shared subreddits}}{(\#\ \text{subreddits}\ user_1\ \text{is in})(\#\ \text{subreddits}\ user_2\ \text{is in})}$$

SimRank

- Second-order proximity measure: compares the neighborhoods of two nodes [1]
- Extension of PageRank - localized PageRank
- Initiate random walkers on a node of interest and measure likelihood of seeing the random walkers at other nodes in the graph after several iterations.
- Parameters:
  - random-walk size = 10
  - teleportation parameter β = 0.8

Exploration

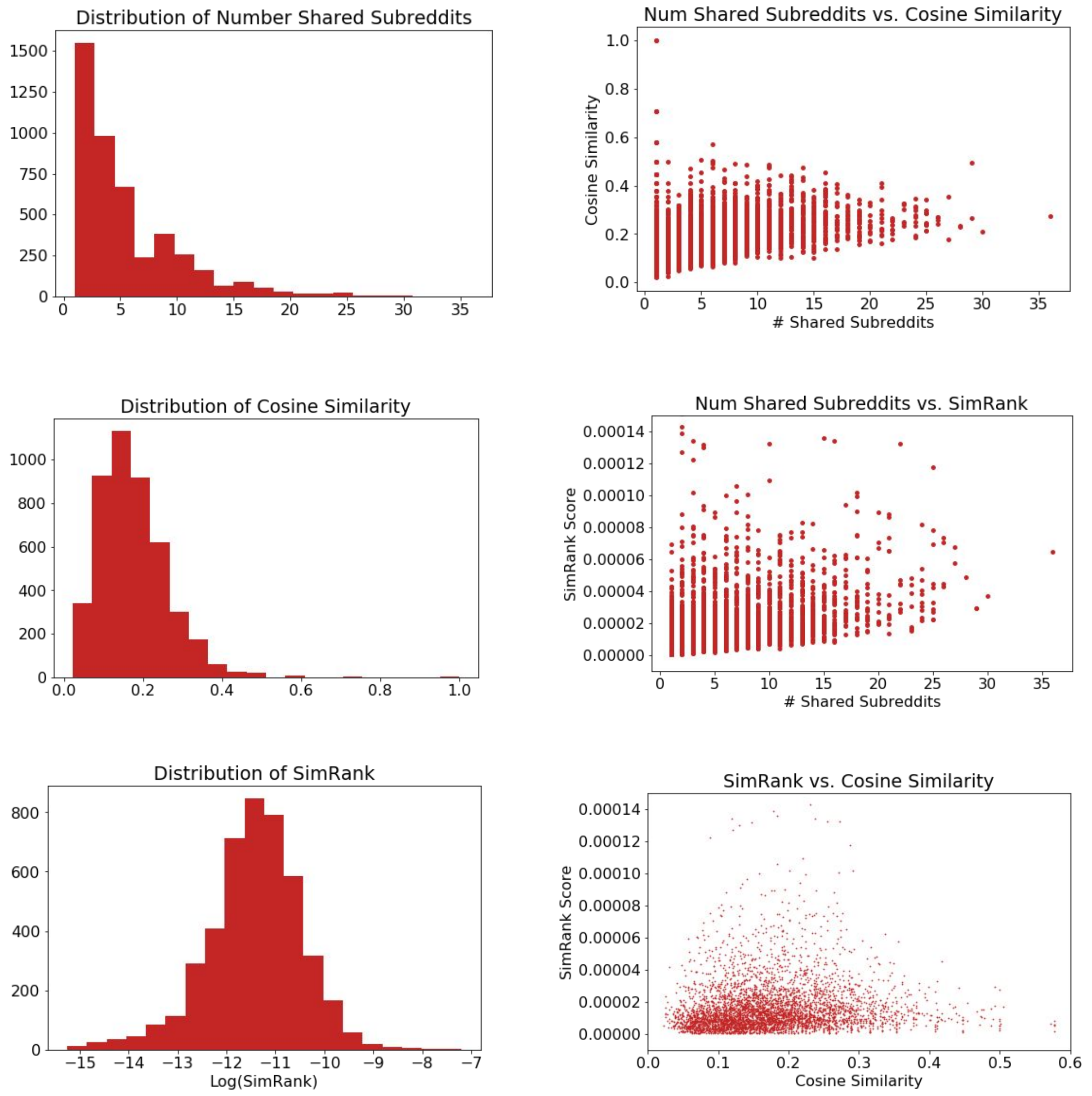


Figure 3a. Distribution of similarity measures in the Reddit bipartite graph (left). Figure 3b. Relationship of similarity scores with each other (right).

RESULTS

	t-statistic	p-value	significance
Cosine Similarity	-0.60	0.55	
SimRank	-2.04	0.09	*

Figure 4. Two sample t-test p values for similarity measures.

- Cosine Similarity was not statistically significantly associated with Delta awarded
- SimRank was weakly associated with Delta awarded (α = 0.10)
- Second order methods which consider graph structure may help predict Delta better than first order methods like Cosine Similarity

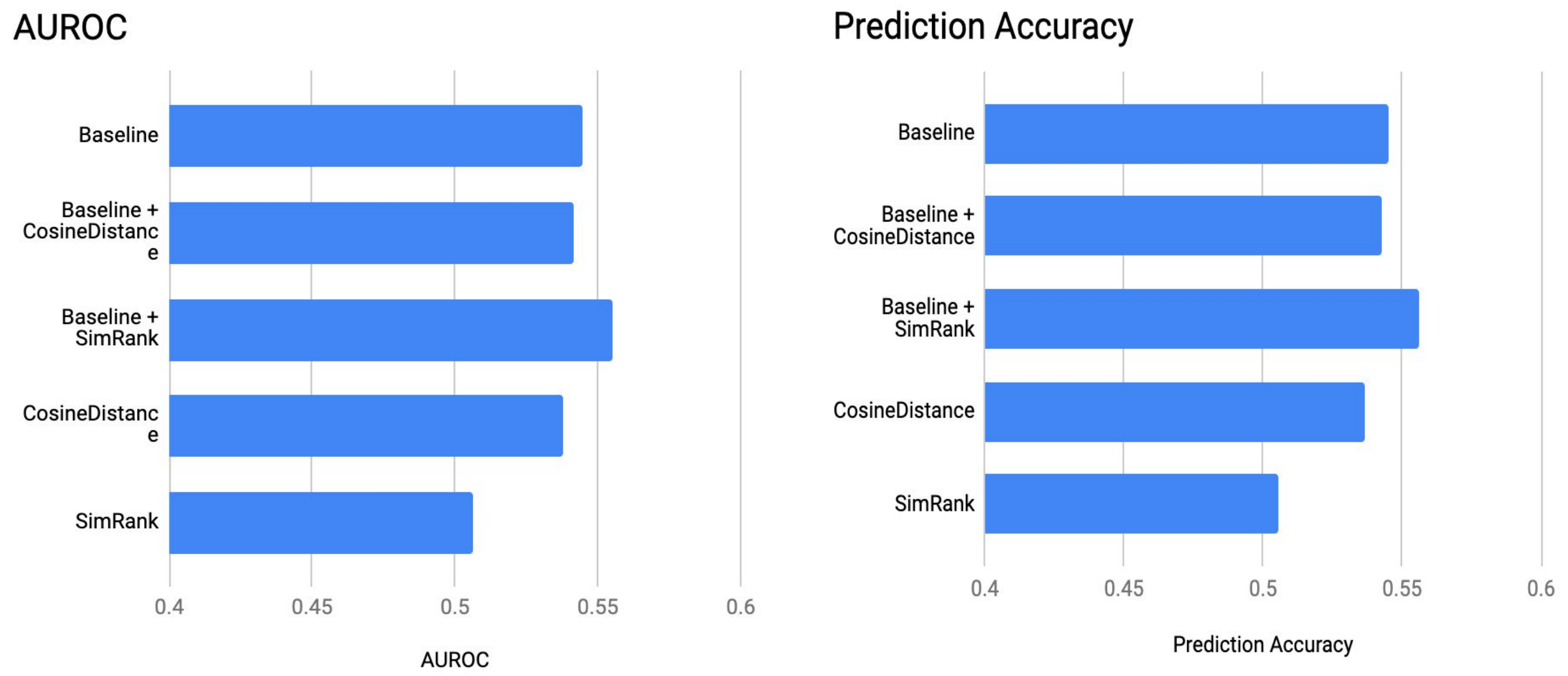


Figure 5. Evaluation of logistic regression models that use various subsets of linguistic and community similarity features.

CONCLUSION

- We used Cosine Similarity and SimRank to augment our linguistic model
- Adding Cosine Similarity does not improve the accuracy of the model
- We observed an improvement in performance after introducing user similarity based on SimRank, which measures similarity of user neighborhoods
- We see promise in the graph-based approach and wish to explore more advanced graph embeddings in the future

REFERENCES

[1] *Jeh, Glen, and Jennifer Widom. "SimRank: a measure of structural-context similarity." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.*

[2] *Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016, April). "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions." In Proceedings of the 25th international conference on world wide web.*

Reddit data source:
https://www.reddit.com/r/bigquery/comments/5z957b/more_than_3_billion_reddit_comments_loaded_on/