TARGRES - Tree-based Argumetation Resolution

Julio Cezar Silva

Abstract—TARGRES, or **T**ree-based **Arg**umetation **Res**olution, is built as a bias-reducing approach for identifying the most robust answers for complex - but deterministic - debates. Specifically, it is comprised of a mathematical abstraction construct over debate resolution as a whole, as well as applying regression principles to approximate a custom argumentation robustness score. A sizeable amount of progress in precision is achieved, over the arguably new field of intelligent debate resolution.

Index Terms—Argumentation, Natural Language Processing, BERT, Udacity, Machine Learning.

1 Introduction

Debate resolution helps shape social and scientific progress since the dawn of time. In today's era this is not only factual offline - it's also one of the backbones of the internet:

- StackOverflow and over a hundred siblings
- a vast portion of Reddit
- Quora

... and the many forums in between. The whole ever-present world of Q&A was built to thrive on the resolution of various rarely one-sided debates. The goal of TARGRES is to help lead to intelligence that can receive a **debate as input**, and **output the answer** that's most robust - and likely right.

This implementation makes use of BERT's embedding features [1], `LGBMRegressor` and a custom scraped dataset of a thousand discussions from the Kialo website. Through these, this study aims to understand and rank the robustness of argumentations in a discussion, with a combination of textual context, localized social impact (feedback from ratings) and ramifications (pros & cons stemming from a given argumentation).

2 Data Acquisition

Kialo, a platform made for argument-centered open debates, has been the home of many notably divisive discussions, the likes of Should general AI have fundamental rights?, Is Donald Trump a good president? and even meta ones like Kialo won't be successful in a world where everybody is on Facebook. Modern social networks are full of divisive debates, but there's a couple of things about Kialo that set it apart from those places:

- **Discussions are organized in trees**. Any debate has its pros and cons, but any argument within debates also has its pros and cons. And that's how Kialo views its threads, both interface-wise and API-wise, as represented below:

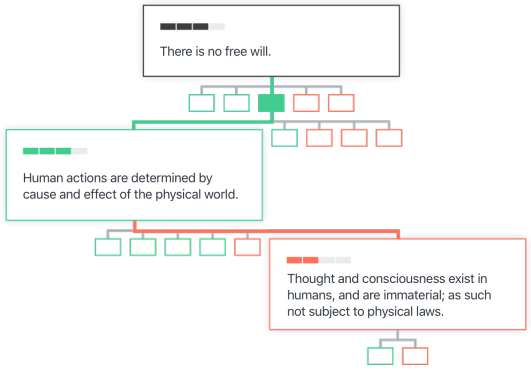

Fig. 1: Discussion tree representation from Kialo Edu - "How does Kialo work?"

- **Claims are voted upon**. Any tree node can have votes cast to it, and can be corrected or revoked by other users. Since votes won't grant a privileged position in the board[1], they're more likely to be less biased expressions of accordance by different parts of the community.
  With votes per-node weighed by the proportions of pros & cons, it's possible to create a **deterministic metric of reasoning robustness**, both globally (tree-level) and locally (node-level).

Data was collected for the top 1000 featured discussions from Kialo, and both the discussions, claims and claim/discussion ratings were aggregated in a single entity for analysis. To determine parent/child relations and reflect the discussion trees in code, claims' location metadata was backtracked.

2.1 Defining an Argumentation Tree

As seen before, Kialo already organizes its discussions as trees. For the sake of clarity, the entirety of this article is going to refer to those discussions following the nomenclature of tree as a data structure:

1. As happens in StackOverflow and Reddit, which is why TARGRES isn't primarily trained on those sources, yet means to extrapolate to them afterwards.

- **Root node**: the thesis itself, the first and central proposition within a discussion. In Kialo, it could be e.g. "A General AI should have the same rights as a human".
- **Child nodes or children**: Any claim that is a sub-claim of another.
- **Parent nodes**: Any claim that contains another as its child.
- **Pros and cons, or pro/con-nodes**: A child node that specifically supports **or** opposes its parent, whether it be another claim or the thesis itself. In its interface, Kialo highlights pro nodes with a green border, and con nodes in a red one.
- **Level or depth**: The number of vertical connections between the root and a given node. A new, unused discussion has a depth of 0, and upon having its first argumentation posted its total depth becomes 1. The first child node from that argumentation makes for a depth of 2, and so on. Here, the first level of claims in a thesis will be called **top claims** or **top arguments**.
- **Ancestors**: Ancestors are nodes that are multiple levels above in the tree in relation to the currently discussed node.
- **Subtree**: The entire group of nodes below a specific claim.

3 Preprocessing

The steps taken to prepare Kialo APIs' data into the dataset used in this study:

- Backtracked parent-child and tree depth information from metadata, replicating tree data structures in code
- Filtered useless nodes: arguments that were deleted, flagged or empty
- Parsed claim texts removing Markdown syntax, standard stop words, punctuation and casing
- Traversed the tree recursively calculating the custom metrics defined in Multi-level Robustness Metrics

4 Multi-level Robustness Metrics

In the current section, the whole structure of discussion trees will be mathematically defined, in order to construct a generically applicable concept of robustness.

4.1 Claim Rating System

By design, Kialo uses a 5 point Likert scale as the rating system for all claims therein. Users can rate a claim's impact on parent (be it a claim or the thesis), in a scale of No Impact (0), Low Impact (1), Medium Impact (2), High Impact (3), Very High Impact (4), as seen in Fig. 2.

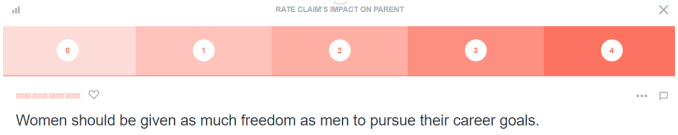

Fig. 2: Impact rating from a con-node for the Should women stay at home to raise children? discussion.

Having this type of voting layout has been long discussed as a less-biased, more in-depth alternative to dichotomous systems like [up/down]voting, common in most social networks - even though there's a possible issue with middle category consistency in Likert scales [2].

The main focus of TARGRES is mathematically understanding both the claim text as well as the ratings such claim gets, and the hidden relations in-between.

4.1.1 *Superficial Thesis Voting*

A thesis can be voted in the same way as a claim can, but in contrast to claims, a thesis is rated on its veracity, and the scale values are: False (0), Improbable (1), Plausible (2), Probable (3) and True (4)

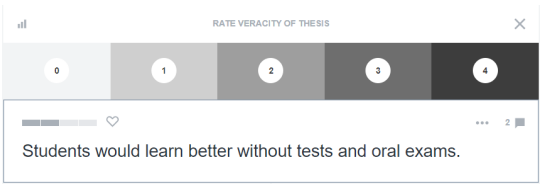

Fig. 3: Thesis rating from the Would students learn better without tests and oral exams? discussion.

This general veracity rating doesn't necessarily include the perceived veracity of all claims within a thesis, since users can vote on a thesis before ever traversing the argumentation tree. To compute a thesis' average veracity T_v:

$$T_v = \frac{\sum_{i=1}^{S} V_i}{S} \tag{1}$$

where:

V_i = number of veracity ratings for the thesis, by type ("False", "Improbable", ...)
S = number of rating types in the scale (equals 5 for Kialo's 5 point scale)
T_v = the thesis' average veracity rating

Since these votes can be easily done without any argumentation analysis - reading just the thesis is enough - a prior that is going to be assumed in this study is that thesis veracity ratings represent the **superficial, general opinion of the usebase**. This purposefully bypasses genuine, thought-through thesis ratings, in an attempt to aim for the most statistically significant scenario, extrapolated from Kialo's interface.

4.2 Calculating Thesis Robustness

The main goal of defining argumentation robustness for both sides of a discussion is to mathematically label the global consensus over a given thesis. This metric aims to represent the contextual strength of both pros and cons of all branches in a discussion tree, summing discussion topologies - as the one below - into a single number.

Fig. 4: Argument topology for the Should there be a Universal Basic Income (UBI)? discussion.
The root node is centered in gray, pro-thesis arguments in green and anti-thesis in red. Bar widths are scaled by vote counts.

As noted before, the robustness of a single argumentation node will be weighed by pro/con votes traversing the tree downwards from the root node. Within the discussion tree, the impact weight w_x for a single argumentation node x will be

$$w_x = \frac{\sum_{i=1}^{s} r_i}{s * r} + \frac{l}{l_x}, \tag{2}$$

where:

r_i = number of impact ratings for the claim, by type ("No Impact", "Low Impact", ...)
s = number of rating types in the scale (equals 5 for Kialo's 5 point scale)
r = total number of ratings in the discussion
l = total number of levels in the discussion
l_x = the level x is in; $d \geq 1$
w_x = impact of x on its parent node

This means w_x includes both the average ratings $\sum \frac{r_i}{s}$ and the level/depth l of a claim, each accordingly scaled: both r and l are normalization variables, making a thesis with more activity (and thus more votes and levels) comparable to one that's less active or just new.

By also depending on l_x instead of just r_i, w_x balances the weight of nodes that are farther down in the tree, since those nodes are less visible in the interface and thus more likely to have zero votes. If w_x was just dependent on votes, it'd result in an unrealistic weight of 0 for arguments lacking any vote. At the bottom of the tree, top arguments get broken down into increasingly specific nodes, which are essential proofs of smaller parts of the thesis itself. V_x is inversely proportional to l_x because those smaller, essential arguments prove generally less about the thesis as a whole.

The robustness score $R(x)$ is defined as

$$R(x) = w_x * \left(\frac{\sum_{i=1}^{n} R(P_{x_i})}{\sum_{j=1}^{m} R(C_{x_j})} \right), \tag{3}$$

where:

P_{x_i} = impact weights of pro children of x
C_{x_j} = impact weights of con children of x

$R(x)$ is a **recursive** measure of the whole subtree under x: proportional to the sum of $R(P_{x_i})$ and normalized by the sum of $R(C_{x_j})$, which means all pro and con children of x will also have their child nodes' robustness measured, and so forth, down to the last lone leaf (childless node) at the bottom of the tree. To finally **confirm or debunk a thesis**, both pro-thesis and con-thesis subtrees are recursively calculated and compared:

$$P = \sum_{i=1}^{n} R(P_i), \tag{4}$$

$$C = \sum_{j=1}^{m} R(C_i), \tag{5}$$

where:

P_i = the thesis' top pros (first generation of children nodes supporting the thesis)
C_i = the thesis' top cons (first generation of children nodes opposing the thesis)
P = cumulative pro-thesis robustness
C = cumulative anti-thesis robustness

All claims at all levels are accounted for - both pro and con subtrees will be traversed top-to-bottom to define P and C. The thesis robustness consensus R_T will be

$$R_T = \frac{P}{C} \tag{6}$$

R_T is a measure of global consensus regarding the macro and micro aspects of a thesis: $R_T \gg 1$ indicates the pro-thesis nodes are more robust, thus confirming the thesis; $R_T \ll 1$ indicates the con-thesis nodes as dominant, thus judging the thesis false; $R_T \approx 1$ is a middle ground between *plausible* and *probable*, indicating equally weighing points and counter-points. As a thorough metric, R_T will define the **public opinion's verdict regarding the robustness of the entire thesis**.

4.3 Comparing Robustness to Thesis Ratings

As discussed in 4.1.1, the averaged ratings on theses T_v is here considered to be a superficial representation of the public opinion, whereas the robustness R_T aims to represent a deeper metric by nature. To assess this relationship between the two, for the sake of visualization, both T_v and R_T were normalized in a $[0 - 1]$ interval, and the discussions were sorted in ascending order by the number of claims in each.

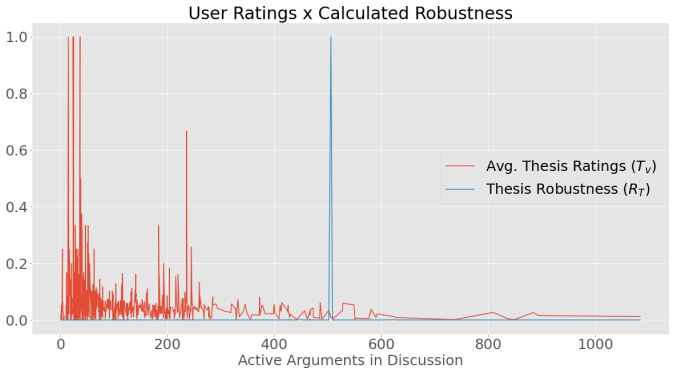

Fig. 5: Comparison of thesis ratings (orange) and this study's robustness calculations (blue).

There's a clear behavior distinction, as most of the discussions with lower claim counts (having less public interest, or being newly created) hold the highest average ratings, a tendency that settles sharply around the 300 mark.

Following the sharp decrease of T_v is the sharp increase on R_T. This numerically strengthens the point that T_v is very distinctive to an in-depth assessment of the public opinion itself - which is what R_T astrives to be. Yet, this also points to another detail about the dataset:

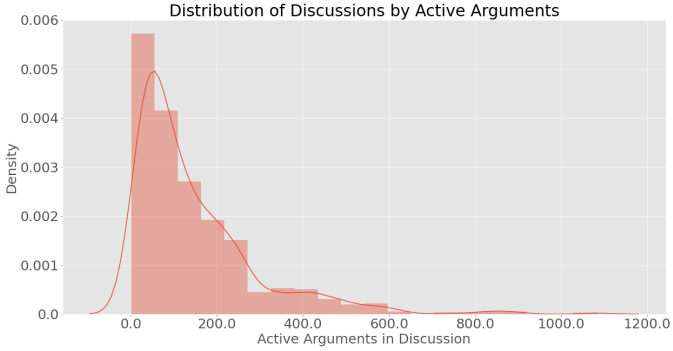

Fig. 6: Discussions' active claim count distribution.

At the time of writing, Kialo's discussion count is over 12000, so it's safe to assume the thousand discussions analyzed here are not fully representative of the platform's reality. Using 1/12th of the total database, that was retrieved in no particular order (no sorting parameter passed to Kialo's API) is arguably the biggest source of the skew seen in 6.

5 Solution Statement

This initial iteration of TARGRES is based on the following process pipeline:

- Preprocessing data and adding custom metrics for w_x and R_T calculation
- Selecting a subset of discussions (most active 1000 discussions).
- Natural language understanding of context within individual argumentations (given BERT [1] pre-trained model's embeddings as input)
- Propagating PCA-reduced embeddings of theses to their associated argumentations for broader contextual input
- Evaluating per-node predicted w_x vs actual calculated w_x to determine RMSE within an `LGBMRegressor`
- Comparing the calculated and the predicted consensus (R_T) vs superficial opinion (thesis veracity score T_v)

The RMSE evaluation metric was chosen due not only to its ease of use in benchmarking models, but mostly its assertivity on solid penalization of large errors [3]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}} \tag{7}$$

The squaring of the difference between predictions and labels helps trim the biggest predicted discrepancies on w_x, then stopping its propagation onto calculating R_T from predicted argumentation weights.

6 Discussing Model Fairness

Model fairness is an important discussion on how a given intelligence may be skewed towards a specific demographic pattern regarding the social scope of a model [4] - asking a political question only to citizens of a country where the full political spectrum isn't represented, or asking a question about the importance of feminism but only to men, are all but a few examples of this skew. Fairness is arguably even more important in a software that tries to determine the veracity of a mainly social claim. Once a TARGRES application is released, the transferable intelligence is dependent on the bases of which the models were created. So the effectiveness of TARGRES is limited by bias from the data source used for training:

- **The user ecosystem**: who are they, how many are they, from what part of what country do they come, and all demographic data. This is paramount to determining fairness.
- **The voting system**: how is it structured (up/downvotes, 5 point scales, etc), how each vote option is worded and displayed (e.g. even color distribution may shift how users vote [5]).
- **The platform structure and interface**: how theses are displayed, how the argument *tree* is displayed - i.e. how difficult and time consuming it is to properly read the entirety of arguments present in the tree, up to the last leaf.

All points above define how much bias this algorithm *will* have, since thesis robustness R_T is entirely dependent on individual argument ratings, which are done: by the **user ecosystem**; through the **voting system**; under the **platform structure**. Having access to all relevant data about these points, and applying it into normalization strategies

is adamant to maintain solid fairness, while aiming pure logical truth - supposing it exists[2].

7 Results

There was a solid progression on the evaluation metrics after the initial round of feature engineering - `creation_delta`, a time difference between the thesis' and the argument's creations, being the most prominent engineered feature, as seen in the feature importance plot from 7:

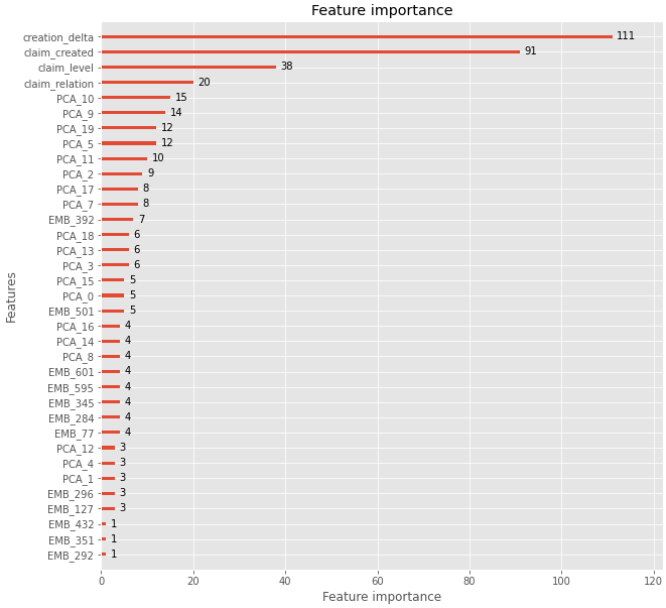

Fig. 7: Feature importances of the best regression iteration

Having the two top features be time-related reveals a lot about the cumulative, permanence-related importance of a claim: two claims that are, in theory, equally robust, will obtain considerably different absolute vote counts if one is created a long time before the other.

Next come the group of textual-context-related features. The benchmarked models all seem to have given a similar amount of importance to a select group of claim text embedding features (`EMB_#`), but, due to near-categorical capabilites, the PCA-reduced thesis embeddings were clearly dominant, effectively blocking top-10 entries before a claim embedding enters the rankings. For that reason, different amounts of components for this PCA were also benchmarked:

2. Excluding ill-advised users and practices like sockpuppetry [6], **the veracity of all votes shall be considered equal**, and that may change the attainable range of logical truths.

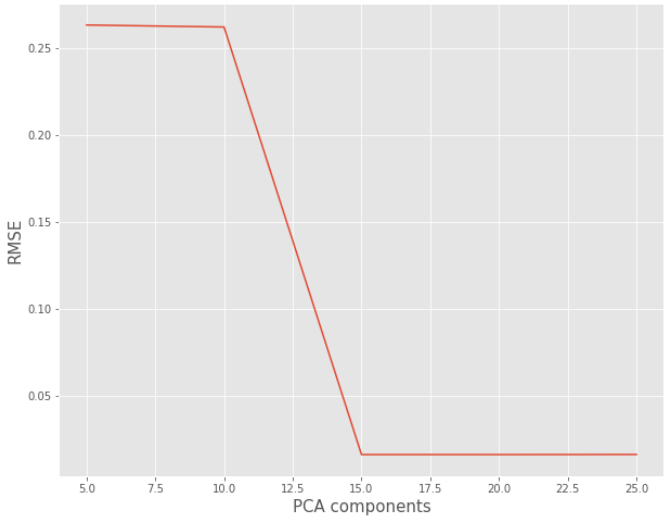

Fig. 8: RMSE for the same models under different counts of thesis PCA features.
The benchmark was done in the $[5, 25]$ interval.

There is a very clear elbow at the count of 15 PCA components, after which the RMSE starts to gradually increase once again. The context hints given by the thesis are undeniable, but given this benchmark and the already large feature space from claim text embeddings, it's clear that just joining the thesis embedding entirely would not be beneficial.

8 Conclusion / Future Work

TARGRES, currently, is more an initial exploration in the mathematical modeling of argumentation resolution, than it is a mature algorithm. Even within a small subset of a specific data source, there is much room for improvement.

The first and foremost point to improve in this algorithm, given its goal in the world of argumentation, would be model fairness. But as it is, this factor depends on Kialo being allowed and willing to share any demographic feature for research, in the first place. When adjusting data for fairness, the models' evaluation scores may suffer, but that is a very comfortable tradeoff, especially for the context of TARGRES.

From the model benchmarks, seeing how the time-related features were paramount to the best iterations of regression, it would likely be a good decision to do a proper feature engineering round, breaking down time metadata, creating new features and - if the available data allows it - **analyzing the distribution of votes and w_x weights over time**. This could even lead to a different strategy of normalization throughout the whole dataset. There were also some time-related features in Kialo's data that TARGRES did not work with, like latest dates of activity and popularity indices.

Even so, it must be observed that the exploration of tree-like representation of debates, and its resolution through mathematical abstractions and Machine Learning, is a fairly new area of research. The maturity of this study's algorithm, and of its exploratory analysis, somewhat matches the apparent maturity of the specific field itself. Novel developments, even if not pristine, are necessary for progress.

REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. Cited 2 times on pages 1 and 4.

[2] D. Andrich, "A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling thurstone and likert methodologies," *British Journal of Mathematical and Statistical Psychology*, vol. 49, pp. 347–365, Nov. 1996. Cited on page 2.

[3] W. Wang and Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," *IOP Conference Series: Materials Science and Engineering*, vol. 324, p. 012049, Mar. 2018. Cited on page 4.

[4] V. D. Warmerdam and M. Brouns, "Priors of great potential: Common sense reduced to priors," Feb. 2020. Cited on page 4.

[5] T. Whitfield and T. Whiltshire, "Color psychology: a critical review.," *Genetic, social, and general psychology monographs*, 1990. Cited on page 4.

[6] Z. Bu, Z. Xia, and J. Wang, "A sock puppet detection algorithm on virtual spaces," *Knowledge-Based Systems*, vol. 37, pp. 366–377, Jan. 2013. Cited on page 5.