STEntConv: Predicting Disagreement with Stance Detection and a Signed Graph Convolutional Network

Isabelle Lorge [1], **Li Zhang** [1,2] , **Xiaowen Dong**[1], **Janet B. Pierrehumbert**[1]
[1] Department of Engineering, University of Oxford
[2] Institute of Finance and Technology, University College London
isabelle.lorge@psych.ox.ac.uk, ucels07@ucl.ac.uk,
xiaowen.dong@eng.ox.ac.uk, janet.pierrehumbert@oerc.ox.ac.uk

Abstract

The rise of social media platforms has led to an increase in polarised online discussions, especially on political and socio-cultural topics such as elections and climate change. We propose a simple and novel unsupervised method to predict whether the authors of two posts agree or disagree, leveraging user stances about named entities obtained from their posts. We present STEntConv, a model which builds a graph of users and named entities weighted by stance and trains a Signed Graph Convolutional Network (SGCN) to detect disagreement between comment and reply posts. We run experiments and ablation studies and show that including this information improves disagreement detection performance on a dataset of Reddit posts for a range of controversial subreddit topics, without the need for platform-specific features or user history.

Keywords: GCN, social media, stance

1. Introduction

Social media now form an integral part of many people's lives. While these tools have allowed users unprecedented access to shared content, ideas and views across the world, they have also permitted the fast rise and spread of harmful forms of communication, such as fake news, abuse and communities acting as radicalising echo chambers at unseen scales (Terren and Borge, 2021). It is then of high interest to investigate the polarisation of opinions as a reflection of ever-shifting political and socio-cultural dynamics which have direct impact on society. For example, detecting disagreement between users can help assess the controversiality of a topic, give insight into user opinions which would not be obtainable from their post in isolation or provide a way to estimate numbers for sides of a debate.

Online communities constitute an ideal terrain for this investigation, as they are likely to foster various tensions and debates and allow researchers to examine them in real time or longitudinally (Alkhalifa and Zubiaga, 2022). Previous work on detecting disagreement has focused on supplementing textual information with user network information, either gathered through platform-specific features such as Twitter's following system, retweets and hashtags (which cannot be generalised across platforms) (e.g., Darwish et al., 2020 or through user-user interaction history (which is not necessarily available) (e.g., Luo et al., 2023). Instead, to the best of our knowledge we are the first to represent users through a user-entity signed graph weighted by stance. In addition to being generalisable to any platform and not requiring user interaction history, our method has the potential to provide more explainable representations for users by explicitly tracing disagreements back to entities they feel positively or negatively about. Furthermore, the graph can easily be adapted to various controversial topics by selecting entities relevant to that topic, is able to accommodate different amounts of information per user, and a user-entity signed network constitutes a natural and explicit representation of polarising allegiances (cf. figure1). Finally, we derive stance towards entities in an unsupervised manner which means there is no need to obtain costly manual labels.

We make the choice to use BERT both for unsupervised stance detection and for the textual part of the model itself. This choice is partially to be able to directly assess the additional contribution of our signed network to the former best model from Pougué-Biyong et al. (2021) for this dataset, but also because we build it with the view of a relatively lightweight model with potential for real time applications. In addition, large language models have been shown to underperform smaller state-of-the-art fine-tuned models on specific tasks, for example in the biomedical domain (Ateia and Kruschwitz, 2023). However, we do provide results for the performance of an open-source large language model (Falcon, Almazrouei et al. 2023) on the task as a comparison.

Our main contributions are as follows[1]:

1) We offer a simple, unsupervised method to

Work completed while Li Zhang was a postdoctoral assistant at the University of Oxford.

[1] We make all our code and data available at `https://github.com/isabellelorge/contradiction`

LREC-COLING 2024, pages 15273–15284
20-25 May, 2024.

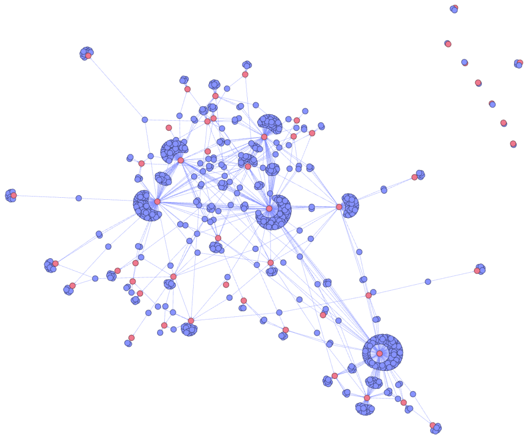

Figure 1: User-entity graph visualised with Gephi (Bastian et al., 2009) (positive edges). We apply a force atlas layout. Pink nodes are entities, blue nodes are users which we can see clustered around the target entities they expressed a positive stance for.

extract user stances towards entities by leveraging sentence-BERT.

2) We build a model using a weighted Signed Graph Convolutional Network on a user-entity graph with BERT embeddings to detect disagreement, improving on previous state-of-the-art results on a dataset of Reddit posts.

3) We present various model ablation studies and demonstrate the robustness of the proposed framework.

We start by outlining current research regarding stance and signed graphs which is relevant to our task. We then move on to describing the dataset used and graph extracted from our data, the architecture and various parameters of our model. We finally present experimental results and discuss their implications.

2. Background

2.1. Stance

The word stance refers to the intellectual or emotional attitude or position of an author towards a specific concept or entity, such as atheism or the legalisation of abortion (Mohammad et al., 2016). This is different from the concept of sentiment as it is usually defined in sentiment analysis, where it refers to the overall emotion expressed by a piece of text. A given text can then have one sentiment value but express multiple positive and negative stances, the target of which is not necessarily explicitly mentioned in the text.

Some concepts lend themselves more easily to the elicitation of stance. For example, consider the following quote from the Wikipedia section on Donald Trump:

> **Donald John Trump** (born **June 14, 1946**) is an **American** politician, media personality, and businessman who served as the **45th** president of **the United States** from **2017 to 2021**. **Trump**'s political positions have been described as populist, protectionist, isolationist, and nationalist. He won the **2016** presidential election as the **Republican** nominee against **Democratic** nominee **Hillary Clinton** despite losing the popular vote.

It would be strange to ask whether the author is in favour or against *born, media, businessman, who, is, from, described, presidential* or *popular*. On the other hand, the words in bold type [2] (with the exception of numerals, i.e. dates and ordinals) seem like more appropriate targets for holding an opinion. These words are generally referred to as *Named Entities* or NEs, a term first coined for the Sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1996). The category aims to encompass expressions which are *rigid designators*, as defined by Kripke (1980), i.e., which designate the same object in all possible worlds in which that object exists and never designate anything else. In other words, these expressions refer to specific instances in the world, including but not limited to proper names. Common NE categories are: organisations, people, locations (including states), events, products and quantities (including dates, times, percent, money, quantity, ordinals and cardinals).

With the exception of the last category, most of these constitute valid targets for extracting author stance because, unlike other terms (e.g., verbs, adverbs, prepositions, some nouns and adjectives) they can be involved in debates and elicit diverging intellectual or emotional viewpoints. Communities tend to be created on the basis of shared traits which are either given (e.g., race, gender, nationality) or acquired (preferences and opinions). Thus, for many contentious issues, agreement and disagreement between individuals are likely to crystallise around attitudes towards a few key entities which define membership identity in a form of 'neo-tribalism' (Maffesoli, 1995). Stance can be modelled at the post or author level. Here we choose to leverage all posts from a known author to determine their stance toward a specific entity.

[2] Bold typed words are all word spans identified by Spacy (Honnibal and Montani, 2017) as named entities.

2.2. Signed Graphs

Graphs, defined as combinations of nodes and edges, are useful abstractions for a variety of structures and phenomena. They can take several forms: directed (e.g., Twitter following) vs. undirected (e.g., Facebook friends); signed (e.g., likes and dislikes) vs. unsigned (e.g., retweets), homogenous vs. bipartite (with nodes of different types where there is no between-type edges, e.g., employees and companies they worked for). In the current paper, given we model user-entity stances, the graph constructed is a signed bipartite graph. While it is technically directed (the stance is from user towards entity), there are no edges in the opposite direction (i.e., from entity to user), thus we treat the graph as undirected for simplicity.

Various methods have been developed for node representation in graphs. When no node features are available, methods relying on connectivity and random walks such as DeepWalk (Perozzi et al., 2014) and node2vec (Grover and Leskovec, 2016) can produce low-dimensional node embeddings using a similar algorithm to skip-gram in Word2vec, i.e., by predicting a node given previously encountered nodes. Graph Neural Networks (GNNs), on the other hand, can leverage both connectivity and node features from a local neighbourhood to produce node representations. Among these, Graph Convolutional Networks (GCNs) were first introduced by Kipf and Welling (2016) and constitute a popular option akin to a generalisation of Convolutional Neural Networks (CNNs), by performing a first-order approximation of a spectral filter on a neighbourhood.

GCNs were originally designed to handle unsigned graphs. However, in the case of stance, as well as in many other applications related to social media, we encounter networks which are signed, i.e., which involve positive and negative edges. Processing these types of graphs and producing meaningful node representations is not straightforward, as there is an intrinsic qualitative difference between the two types of edges which cannot be optimally resolved by e.g., treating them alike, ignoring negative edges, or ignoring edges that cancel each other. One solution is to keep positive and negative representations separate from the graph neural network separate and simply concatenate them. Another way suggested by Derr et al. (2018) relies on assumptions from *balance theory* (Heider, 1946; Cartwright and Harary, 1956), which comes from social psychology and formalises intuitions such as 'an enemy of an enemy is a friend'. Thus, for each layer l, the aggregation function would gather on the positive side not only friendly nodes, but friends of friends and enemies of enemies, and similarly on the negative side get information from enemies but also friends of enemies and enemies of friends. The positive and negative convolutions are then concatenated together to produce the final node representations as in the simpler model. In our experiments we test both the simple signed model and the model with additional aggregations based on balance theory.

3. Dataset

We use the *DEBAGREEMENT* dataset for our experiments (Pougué-Biyong et al., 2021), a dataset of 42894 Reddit comment-reply pairs from 5 different subreddits (*r/Brexit, r/climate, r/BLM, r/Republican* and *r/democrats*) with each pair given one of three labels: *agree/neutral/disagree* (see dataset statistics in tables 1 and 2). The pairs of posts were labelled by crowdsourcers who received intensive training on the issues discussed in the various subreddits. The disagreement prediction task consists in predicting which of the three labels describes the relation between the comment and reply posts.

This is a very difficult task for a number of reasons. First, assessing disagreement around issues such as those discussed in the selected subreddits requires expert knowledge (hence the need for specific training of crowdsourcers). Second, there is a high level of subjectivity involved, which is evidenced by the 'clean' version of the dataset still containing over 60% labels where only 2 out of 3 crowdsourcers agreed. Because of the latter, we choose after examining the data to work with the portion of the dataset that was given the same label by all three crowdsourcers (16723 pairs of posts). Finally, it is worth noting that most previous works focusing on disagreement use Twitter data exclusively (e.g., Darwish et al., 2020; Trabelsi and Zaiane, 2018; Zhou and Elejalde, 2023). Many other platforms like Reddit lack network features such as common hashtags, user following and retweets which are highly useful for detecting endorsement between users. It is then much harder to create user representations indicative of polarisation. This emphasises the crucial need to find alternative features, such as user-entity stances, which can generalise across platforms. While the task has been tackled using user network features (Luo et al., 2023), no previous works have attempted to improve performance without leveraging user interaction features which may not always be available.

4. Framework

4.1. User-Entity Graph Construction

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be a signed undirected bipartite graph where $\mathcal{U} \in \mathcal{N}$ is the set of user nodes, $\mathcal{A} \in$

	r/Brexit	*r/climate*	*r/BlackLivesMatter*	*r/Republican*	*r/democrats*
start date	Jun 2016	Jan 2015	Jan 2020	Jan 2020	Jan 2020
agree	0.29	0.32	0.45	0.34	0.42
neutral	0.29	0.28	0.22	0.25	0.22
disagree	0.42	0.40	0.33	0.41	0.36

Table 1: DEBAGREEMENT statistics per subreddit and period

	comment-reply count	*avg length (comment)*	*avg length (reply)*
r/Brexit	15745	45	40
r/climate	5773	43	41
r/BlackLivesMatter	1929	41	39
r/Republican	9823	38	35
r/Democrats	9624	38	37

Table 2: DEBAGREEMENT post counts and word lengths

$\mathcal{N}$ is the set of entity nodes and $\mathcal{E}$ the set of edges between users and entities, with $\mathcal{E}+$ the set of positive edges and $\mathcal{E}-$ the set of negative edges. Since this is a bipartite graph, there are no edges between users or between entities, and the set of positive and negative edges are defined to be mutually exclusive (ie., there is at most one edge, either positive or negative, between a user and an entity) (see figure 2).

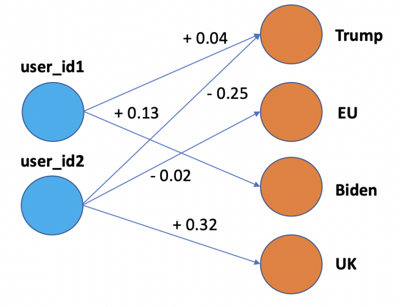

Figure 2: Example user-entity graph. The network is signed, with each edge representing user stance towards an entity.

We build the graph in the following way. First, we extract named entities for each comment and reply post using Spacy (Honnibal and Montani, 2017), discarding entities which pertain to the categories 'CARDINAL', 'DATE', 'ORDINAL', 'WORK_OF_ART', 'PERCENT', 'QUANTITY' and 'MONEY'. Since we do not have ground truth for the stance of each author for each extracted entity, we devise an unsupervised method to obtain a proxy of it by leveraging Sentence-BERT (Reimers and Gurevych, 2019). For each entity, we create 'pro' and 'con' sentences using the templates *I am for X* and *I am against X*. We then compute the cosine similarity between each SBERT-embedded post sentence and each SBERT-embedded template sentence and subtract the 'con' cosine similarity from the 'pro' cosine similarity [3]. Finally, we take the mean of all cosine differences for an entity across all user posts[4]. The advantage of this method is that it is almost entirely unsupervised and does not require prior domain knowledge or manual selection of relevant topics or entities (only the subreddit titles, see below), these naturally arise among the most frequent entities extracted from the corpus. Therefore, we define our stance measure as:

$$stance_{u,e} = \sum_{i \in P}^{N} \frac{\sum_{s \in S}^{M} \frac{cos_s^+ - cos_s^-}{|S|}}{|P|} \qquad (1)$$

where $stance_{u,e}$ is the stance of user u towards entity e, $i \in P$ is the ith post contributed by user u, $s \in S$ is a sentence in the post, cos^+ is the cosine similarity of the post sentence with the 'pro' embedded template sentence and cos^- is the cosine similarity of the post sentence with the 'con' embedded template sentence. We notice a negative bias in our extracted cosine similarities whereby the mean μ of stance values lies around -0.02 and accordingly split edges into positive edges ($stance_{u,e} >= \mu$) and negative edges ($stance_{u,e} < \mu$) after verifying that the stances follow a normal distribution and the median is close to the mean. The statistics of the resulting graph can be seen in table 4.

When examining the extracted entities, it appears that most entities which occur only a few times in the corpus will be irrelevant to our task. We thus apply a combination of two filters: we

[3] While this method has to our knowledge not been previously used, we manually examine the results for 100 sentences by calculating the stance for each named entity extracted using the method described and rating it as correct or incorrect and find satisfactory performance (0.68 accuracy).

[4] In addition, we also mean center all edge weight values by subtracting the mean.

american, antifa, aoc, asian, backstop, bernie, biden, black, blm, brexit, brexiteers, brown, christian, cnn, communist, con, confederate, conservative, corbyn, cuomo, dem, democrat, democratic, dems, dnc, fascist, fbi, floyd, george, gop, greta, holocaust, jew, kkk, leave, leftist, liberal, libertarian, maga, marxist, mcconnell, moderate, moron, msm, muslim, nazi, party, patriot, pete, poc, progressive, propaganda, qanon, racist, referendum, remainers, republican, riot, romney, sander, senate, statue, tory, trump, tucker, warren, white

Table 3: Extracted target entities

	$\|\mathcal{U}\|$	$\|\mathcal{A}\|$	$\|\mathcal{E}+\|$	$\|\mathcal{E}-\|$	$\|\mathcal{D}\|$	$\|\mathcal{D}(\mathcal{U})\|$	$\|\mathcal{D}(\mathcal{A})\|$	$\|\mathcal{CN}(\mathcal{U})\|$	$\|\mathcal{CN}(\mathcal{A})\|$
train	7107	67	3997	4615	0.001	1.83	194	0.32	5.67
test	1513	67	863	866	0.002	1.48	37	0.20	0.60

Table 4: User-entity graph statistics for full training and test datasets. $|\mathcal{U}|$: number of users; $|\mathcal{A}|$: number of entities; $|\mathcal{E}+|$: number of positive edges; $|\mathcal{E}-|$: number of negative edges ; $|\mathcal{D}|$: graph density; $|\mathcal{D}(\mathcal{U})|$: average degree (users); $|\mathcal{D}(\mathcal{A})|$: average degree(entities); $|\mathcal{CN}(\mathcal{U})|$: average common neighbors (users); $|\mathcal{CN}(\mathcal{A})|$: average common neighbors (entities)

keep entities which are amongst the 5000 most frequent entities and whose embeddings have a cosine similarity above 0.5 to at least one subreddit title embedding (*Brexit, climate, BLM, Republican* and *democrats*) (the embeddings used are the initial features for the GCN which are Word2vec embeddings trained on our dataset, see section Training). We obtain both values by conducting a sensitivity correlation analysis on the training set in the following way: we model a negative and positive entity vector for each author respectively as the sum of the negative stance and positive stance entities, concatenate them and measure the Kendall τ rank correlation between the cosine similarity of the author vectors for a given comments pair and the label (0,1 or 2 as disagreement, neutral and agreement). There is a clear peak in correlation with entities which have over 0.5 cosine similarity with at least one subreddit title and are within the 5000 most frequent entities, thus we select these as our threshold values. We also filter out multiword entities which often show redundancy and misextractions. The final set of 67 target entities can be seen in Table 3, a heatmap of cosine similarities to each subreddit can be found in Appendix A and a visualisation of the user-entity graph can be seen in figure 1.

To get a fair assessment of our model and be able to directly compare it with the performance of the GCN model alone, we subset the training dataset to comment-reply pairs which mention at least one of our target entities (for other comment-reply pairs the GCN would not have any features). The final dataset is made of 1770 comment-reply pairs. While this constitutes only 10% of the original full agreement dataset, we notice that disagreements which are most closely related to the subreddit's controversial topic will often contain the target entities [5] . We also believe that given the difficulty of the task (especially on Reddit data where many network features available on Twitter cannot be used), an improvement on a subset of disagreement types is worthwhile and holds promise for applicability. We do provide results for the subset of the dataset where only either comment or reply mentions one of our target entities, which constitutes about 40% of the full agreement dataset or 6174 comment-reply pairs, however we cannot run a comparison with the GCN model alone in this case.

4.2. STEntConv

We adopt the Signed Graph Convolutional Network proposed in Derr et al. (2018) and modify it to integrate edge weights for our stance values so that the positive and negative convolutions are as follows:

$$\begin{aligned}\mathbf{h}_i^{B(l)} = \sigma(\mathbf{W}^{B(l)}[\sum_{j\in\mathcal{N}_i^+}\frac{\mathbf{h}_j^{B(l-1)}}{|\mathcal{N}_i^+|}\mathbf{w}_j,\\ \sum_{k\in\mathcal{N}_i^-}\frac{\mathbf{h}_k^{U(l-1)}}{|\mathcal{N}_i^-|}\mathbf{w}_k, \mathbf{h}_i^{B(l-1)}]),\end{aligned} \tag{2}$$

[5] While our assumption that the model leverages information from entities not present in the text suggests we should be able to use posts which do not mention target entities, we find empirically that this is not the case. However, the better performance over the BERT baseline suggests the model does make use of information not present in the comment-reply pair. We hypothesise that this is because of a correlation between the presence of target entities in specific pairs of posts and the amount of additional information in the entity graph (ie., posts which do not contain target entities tend to come from authors for whom there is little/no entity information).

$$\mathbf{h}_i^{U(l)} = \sigma(\mathbf{W}^{U(l)}[\sum_{j \in \mathcal{N}_i^+} \frac{\mathbf{h}_j^{U(l-1)}}{|\mathcal{N}_i^+|}\mathbf{w}_j, \sum_{k \in \mathcal{N}_i^-} \frac{\mathbf{h}_k^{B(l-1)}}{|\mathcal{N}_i^-|}\mathbf{w}_k, \mathbf{h}_i^{U(l-1)}]), \tag{3}$$

where $\mathbf{h}_i^{B(l)}$ is the weighted aggregation of positive edges for layer l, $\sum_{j \in \mathcal{N}_i^+} \frac{\mathbf{h}_j^{B(l-1)}}{|\mathcal{N}_i^+|}\mathbf{w}_j$ is the weighted sum of 'friends of friends', $\sum_{k \in \mathcal{N}_i^-} \frac{\mathbf{h}_k^{U(l-1)}}{|\mathcal{N}_i^-|}\mathbf{w}_k$ is the weighted sum of 'enemies of enemies' and $\mathbf{h}_i^{B(l-1)}$ the previous layer's positive edges aggregation. Similarly, $\mathbf{h}_i^{U(l)}$ is the aggregation of negative edges for layer l, $\sum_{j \in \mathcal{N}_i^+} \frac{\mathbf{h}_j^{U(l-1)}}{|\mathcal{N}_i^+|}\mathbf{w}_j$ is the weighted sum of 'enemies of friends', $\sum_{k \in \mathcal{N}_i^-} \frac{\mathbf{h}_k^{B(l-1)}}{|\mathcal{N}_i^-|}\mathbf{w}_k$ is the weighted sum of 'friends of enemies' and $\mathbf{h}_i^{U(l-1)}$ the previous layer's negative edges aggregation. We run experiments both with the additional aggregations from *balance theory* and without (i.e., only aggregating direct friends for positive edges and direct enemies for negative edges), in which case the respective weighted aggregations are simply:

$$\mathbf{h}_i^{B(l)} = \sigma(\mathbf{W}^{B(l)}[\sum_{j \in \mathcal{N}_i^+} \frac{\mathbf{h}_j^{(l-1)}}{|\mathcal{N}_i^+|}\mathbf{w}_j, \mathbf{h}_i^{(l-1)}]). \tag{4}$$

$$\mathbf{h}_i^{U(l)} = \sigma(\mathbf{W}^{U(l)}[\sum_{j \in \mathcal{N}_i^+} \frac{\mathbf{h}_j^{(l-1)}}{|\mathcal{N}_i^-|}\mathbf{w}_j, \mathbf{h}_i^{(l-1)}]), \tag{5}$$

This is also the definition of the aggregations for the first layer l = 1. We build the weighted version of the algorithm by adapting the unweighted Derr et al. (2018) implementation from PyTorch geometric (PyG) (Fey and Lenssen, 2019). The rationale for integrating edge weights to the convolutional layer is that, given our unsupervised method for calculating stance, a high absolute value is more reliable and thus considered more informative. Positive/negative edge and node features should then be weighted accordingly when performing message passing (e.g., a small edge weight is more likely to denote a stance close to neutral). The output of the GCN is concatenated to the output of a BERT layer for comment and reply posts and fed to a one-layer feed-forward network. The final model architecture can be seen in figure 3.

5. Baselines

5.0.1. BERT

As a baseline, we fine-tune a BERT (base, uncased) layer as was originally used by Pougué-Biyong et al. (2021), i.e., we ablate the graph convolutional layer from our model and feed this to the linear layer for classification.

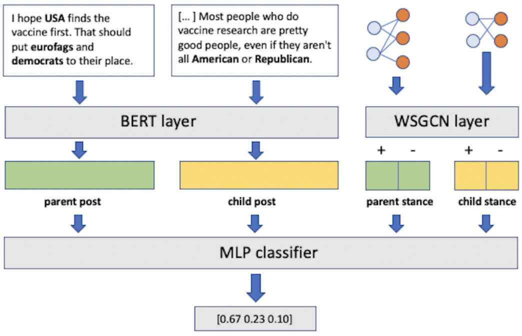
I hope **USA** finds the vaccine first. That should put **eurofags** and **democrats** to their place.
[...] Most people who do vaccine research are pretty good people, even if they aren't all **American** or **Republican**.
BERT layer
WSGCN layer
+
-
+
-
parent post
child post
parent stance
child stance
MLP classifier
[0.67 0.23 0.10]

Figure 3: Model architecture.

5.0.2. GCN only

In addition, we conduct the opposite ablation, i.e., we use only the GCN to assess how the model performs relying only on the positive and negative edges of the stance graph without any access to the text of the posts.

5.0.3. StanceRel

Luo et al. (2023) improve on previous results on the DEBAGREEMENT dataset by building a graph autoencoder and training it on a signed undirected user-user interaction graph which creates user representations based on their previous interactions (agreement or disagreement, i.e., positive or negative) and using the decoded user features along with textual features to detect disagreement. While a direct comparison with our approach may not be fully informative (as the two models use different features which may be available in different situations), we train their model on our subset of data and provide results on our test set to give a picture of the differential value of various user features.

5.0.4. FALCON

We also test an instruct-trained version of Falcon (Almazrouei et al., 2023), specifically *vilsonrodrigues/falcon-7b-instruct-sharded* implemented through the transformers library (Wolf et al., 2019). Prompt and hyperparameters used can be found in Appendix B.

6. Training

We train 100-dimensional word2vec[6] embeddings on the full dataset and use the resulting vectors

[6] We also experiment with GloVe embeddings but the performance is worse.

as initial features for our entities. User features are initialised as 100-dimensional vectors of zeros. We obtain contextual text embeddings for comment and reply posts through the *transformers* library (Wolf et al., 2019) implementation of a BERT (base, uncased) layer whose output we mean pool, excluding special tokens, and concatenate these together with the output of our weighted Signed Graph Convolutional Network layer before feeding this to a one-layer linear classifier. Since the classes are not entirely balanced, we compute class weights and weigh class loss accordingly during training. We use cross entropy as loss function, a batch size of 16, a hidden size of 300 for the first convolutional layer, a learning rate of 3e-5 and the Adam optimiser with weight decay 1e-5. We split the data into 0.80 train, 0.10 dev and 0.10 test and train for 6 epochs (models with BERT layers) and 11 epochs (GCN only). We experiment with number of convolutional layers (one versus two), type of aggregation (balance theory or only direct friends and enemies), edge weights (binary versus weighted) and sentences used to calculate stance (full post versus only sentences containing target entity). We train the models with three different random seeds and average the results.

7. Results

Results can be found in Table 5. The best performing model uses one convolution layer, only direct friends and enemies, weighted edges, and the full text of the post for stance extraction. As can be seen, on the *(c&r)* subset the addition of the user stance graph information helps improve model performance by 7 points on average compared to the BERT baseline and 6 points over the StanceRel model which previously obtained the best results on this task. While the improvement in performance is weaker for the version of STEntConv trained on the *(c|r)* dataset (since this dataset includes authors for whom the model has no relevant stance information), the model still achieves a 3 point increase over the BERT baseline.

In the non-multiple aggregation model, the boost from the stance graph information is lowest for *r/Republican*. This is consistent with additional analyses we conduct on the test dataset showing that the *r/Republican* subreddit has the highest ratio of target entities present in posts versus all entity information available for authors, meaning that there is little extra information from the GCN to be used and most relevant information is already available to the BERT model. Thus, it is likely that the better performance of our model is due to STEnTConv being able to leverage stance information about entities not present in the comment-reply pair being classified.

Adding the second aggregation from (Derr et al., 2018) performs better on the *r/Brexit* and *r/Republicans* subreddits. This would tend to indicate that the additional aggregations for these subreddits remain relevant to the task whereas they introduce noise for *r/climate*. This is supported by *r/climate* subreddit having the lowest cosine similarity on average to the target entities. Falcon performs particularly poorly, in addition to requiring over an hour for inference on the test set (vs. a few seconds for BERT-based models and GCN). StanceRel, while above the BERT baseline, underperforms our model on the test set. Given that the model leverages user history which represents a strong additional signal, we expected it to perform better. The lower performance may be due to sparseness in the user interaction graph or to the model requiring a larger dataset to learn features from interaction history. In addition, the graph StanceRel uses is weighted by frequency which may be prone to biases (e.g., user A from the Green Party agreed with user B from the Green Party many times, but they disagreed on an issue involving a specific entity, e.g. French nuclear company AREVA). Finally, they leverage the assumptions from balance theory which we show may not be useful for this particular task. Unsurprisingly, the neutral class is hardest to classify, with an f1 of 0.4 versus 0.8 for the agree/disagree classes. Examining the data indicates one reason for this might be the heterogenous nature of the class, with some neutral pairs of posts discussing unrelated topics and others agreeing and disagreeing in equal amounts. The confusion matrix between classes can be seen in figure 4.

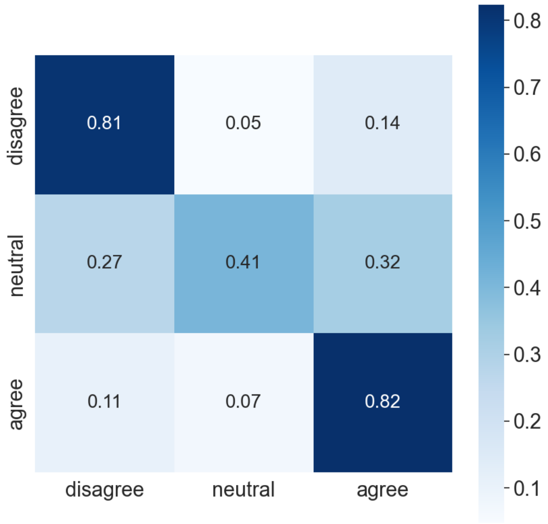

Figure 4: Confusion matrix (STEntConv)

	r/Brexit	*r/climate*	*r/Republican*	*r/democrat*	*r/BLM**	**all (sd)**
(c|r) ***BERT***	.75	**.79**	.73	.69	.72	.72 (0.03)
(c|r) ***STEntConv***	**.78**	.78	**.76**	**.71**	**.75**	**.75** (0.02)
(c&r) ***FALCON***	.40	.25	.45	.38	1.0	.42 (0.28)
(c&r) ***BERT***	.58	.54	.69	.63	.67	.64 (0.06)
(c&r) ***StanceRel***	.67	.30	.67	.60	1.0	.65 (0.22)
(c&r) ***STEntConv*** *(GCN)*	.36	.44	.44	.37	.67	.43 (0.11)
(c&r) ***STEntConv*** *(m.agg)*	**.70**	.41	**.73**	.69	1.0	.70 (0.18)
(c&r) ***STEntConv***	.62	**.64**	.70	**.74**	**1.0**	**.71** (0.14)

Table 5: Macro averaged F1 for each model and subreddit. **STEntConv** = our model enhanced with entity stances; **BERT**= BERT model (base, uncased); **StanceRel** = relation graph model from Luo et al. (2023) **FALCON**: Falcon model (instruct trained, 7B); *GCN* = STEntConv without BERT component; *m.agg*: multiple aggregations, i.e. using the 'friend of friend' additional aggregation from Derr et al. (2018). (c&r) = dataset with target entity in comment and reply; (c|r) = dataset with target entity in comment or reply. Best in bold.*The (c&r) test set only contained one comment-reply pair from the *r/BLM* subreddit.

8. Related Work

There have been a number of studies aimed at modelling stance and disagreement. Regarding stance, a supervised task was introduced at SemEval 2016 along with a small dataset of 4870 tweets annotated for stance against five topics (*Atheism, Climate Change is a Real Concern, Feminist Movement, Hillary Clinton* and *Legalization of Abortion*) (Mohammad et al., 2016). Winning teams used CNNs and ensemble models. However, the limited data and specificity of both style and topics make it difficult to extend a model trained on this data to detecting stance in other domains. Regarding unsupervised methods for both stance and disagreement, most methods have focused on modelling users through dimensionality reduction and clustering of content (Darwish et al., 2020) or through platform-specific (Trabelsi and Zaiane, 2018; Zhou and Elejalde, 2023) or non platform-specific (Luo et al., 2023) network features from user-user interactions. To the best of our knowledge, no previous method used a user-entity graph to model user representations.

Importantly, as stated in the introduction, most previous works use Twitter data which contains platform-specific network features. Thus, our work aligns with other efforts to build alternative network features in polarised communities when user endorsement features (e.g., retweets and follows) are not available, as is the case for Reddit data. For example, Hofmann et al. (2022) build a graph of edges between social entities (concepts and subreddits) to identify the level of polarisation for a given concept. In addition, even previous works which do not use platform-specific data still rely on features which may not always be available, such as user interaction history (Luo et al., 2023) and restrict ability to use graph features to known users. By contrast, our method can leverage entity stances at inference time to model novel users which have similar allegiances to users seen in training. To our knowledge, the current paper is the first to use representations from a signed graph between users and entities for the purpose of predicting disagreement.

9. Conclusion

We presented a simple, unsupervised and domain-agnostic pipeline for creating graph user features to improve disagreement detection between comment-reply pairs of social media posts. We ran several experiments against baselines and performed ablations to examine the contribution of model components and parameters. Our model uses GCN convolutions over a signed bipartite graph of automatically extracted user-entity stances. STEntConv can be leveraged to create comprehensive user representations which take into account stance towards various target entities in order to better predict whether two users are likely to agree or disagree on a range of controversial topics, regardless of availablity of platform-specific network features or user interaction history. As a next step, this method could easily be extended to target entities beyond named entities to include common nouns which are particularly relevant to the controversial topic, especially in cases where the topic is less likely to involve named entities.

10. Limitations

Scale. We acknowledge that the improvement we demonstrate over the baseline applies to only a subset of the original dataset. However, given the difficulty of the task and the lack of additional network features for Reddit data we believe this improvement is still worthwhile. Furthermore, our method could potentially be extended to include stance towards relevant common nouns in addition to named entities.

Domain. While the dataset we used covers a range of controversial topics from socio-cultural to political issues, it was extracted from 5 specific subreddits and thus it is uncertain to what extent our results would apply to other topics of disagreement and whether the model could be generalised to other domains. However, we note that our pipeline for extracting relevant entities is entirely domain-agnostic and thus we believe it could be applied successfully to any forum debating a controversial topic.

Acknowledgements

The authors would like to acknowledge support from the EPSRC (EP/T023333/1). X.D. acknowledges support from the Oxford-Man Institute of Quantitative Finance. This study was supported by EPSRC Grant EP/W037211/1.

11. References

Rabab Alkhalifa and Arkaitz Zubiaga. 2022. Capturing stance dynamics in social media: open challenges and research directions. *International Journal of Digital Humanities*, 3(1-3):115–135.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Samy Ateia and Udo Kruschwitz. 2023. Is ChatGPT a biomedical expert?–exploring the zero-shot performance of current GPT models in biomedical tasks. *arXiv preprint arXiv:2306.16108*.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, volume 3, pages 361–362.

Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of Heider's theory. *Psychological review*, 63(5):277.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

Tyler Derr, Yao Ma, and Jiliang Tang. 2018. Signed graph convolutional network. *CoRR*, abs/1808.06354.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Fritz Heider. 1946. Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112.

Valentin Hofmann, Xiaowen Dong, Janet Pierrehumbert, and Hinrich Schuetze. 2022. Modeling ideological salience and framing in polarized online groups with graph neural networks and structured sparsity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 536–550, Seattle, United States. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. SpaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Saul A. Kripke. 1980. *Naming and Necessity: Lectures Given to the Princeton University Philosophy Colloquium*. Cambridge, MA: Harvard University Press.

Yun Luo, Zihan Liu, Stan Z Li, and Yue Zhang. 2023. Improving (dis) agreement detection with inductive social relation information from comment-reply interactions. *arXiv preprint arXiv:2302.03950*.

Michel Maffesoli. 1995. The time of the tribes: The decline of individualism in mass society. *The Time of the Tribes*, pages 1–192.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. DEBAGREEMENT: A comment-reply dataset for (dis) agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ludovic Terren and Rosa Borge. 2021. Echo chambers on social media: A systematic review of the literature.

Amine Trabelsi and Osmar Zaiane. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zhiwei Zhou and Erick Elejalde. 2023. Stance inference in Twitter through graph convolutional collaborative filtering networks with minimal supervision. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1030–1038.

A. Subreddit-entity heatmap

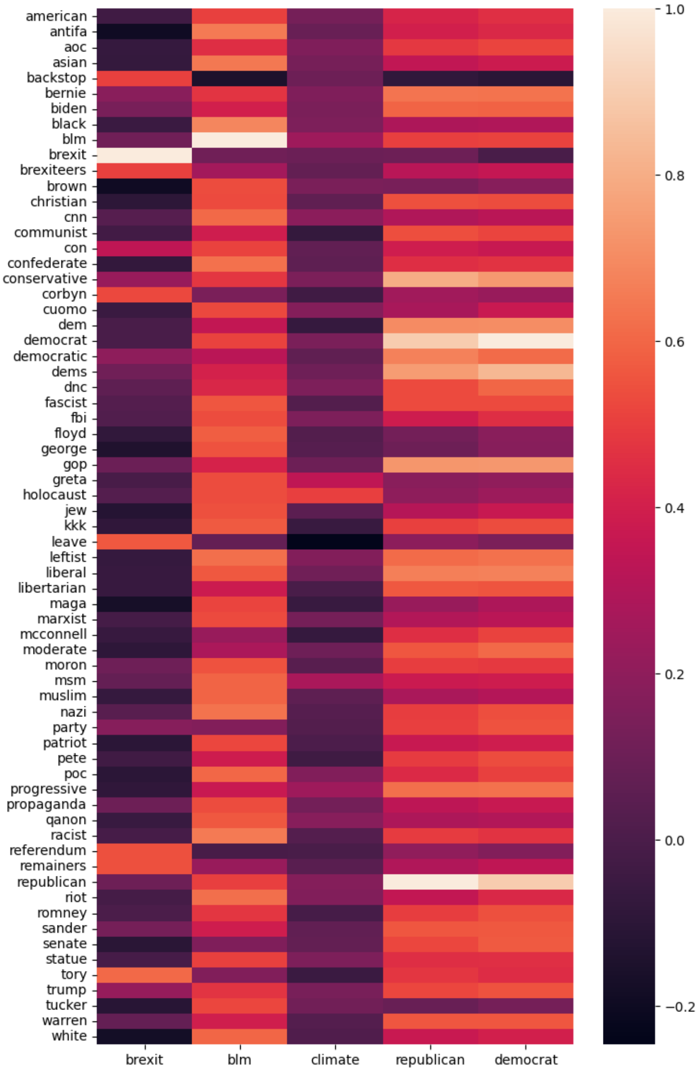

Figure 5: heatmap of cosine similarities between each subreddit name and target named entities.

B. Falcon prompt and hyperparameters

```
quantization_config = BitsAndBytesConfig(load_in_4bit=True,
    bnb_4bit_compute_dtype=torch.float16,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_use_double_quant=True)


model_id = "vilsonrodrigues/falcon-7b-instruct-sharded"


model_4bit = AutoModelForCausalLM.from_pretrained(model_id,
        device_map="auto",
        quantization_config=quantization_config,
        trust_remote_code=True)
```

```
tokenizer=AutoTokenizer.from_pretrained(model_id)


pipeline=transformers.pipeline("text-generation",
          model=model_4bit,
          tokenizer=tokenizer,
          use_cache=True,
         device_map="auto",
          max_length=38 + len(comment)+ len(reply),
         do_sample=False,
          top_k=10,
          num_return_sequences=1,
         eos_token_id=tokenizer.eos_token_id,
          pad_token_id=tokenizer.eos_token_id)


  prompt = f‘‘‘
    Here are a social media COMMENT and a REPLY.

    Say whether the reply is agreeing, disagreeing or neutral towards the comment:

    COMMENT: {comment}

    REPLY: {reply}

    The reply is
    ’’’
```