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

Isabelle Lorge, Li Zhang, Xiaowen Dong and Janet B. Pierrehumbert

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)

May 2024

# Context: social media and polarisation

- Social media now form **integral part of many people's lives**
- Alongside many positives, there are **negatives**, e.g. increased **polarisation of communities** and echo chambers online (Terren and Borge, 2021)
- **Detecting disagreement** between users can help assess the **controversiality** of a topic, give insights into **user opinions** not obtained from isolated post or provide a way to **estimate numbers** for sides of a debate

# Limitations of previous works

- Supplement textual information with **user network information** 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).

# Our method

- **Generalisable to any platform**
- **Does not require user interaction history**
- **Potential for explainability**
- **Can easily be adapted to various controversial topics**
- **Features are gathered in unsupervised way without requiring any labels**

# User-entity graph visualisation

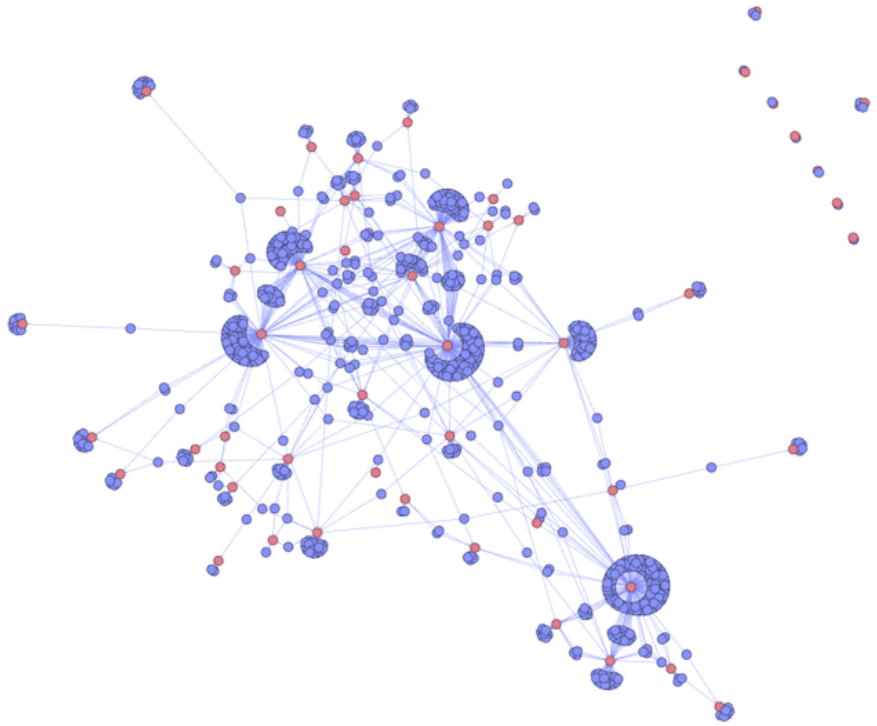

Figure 1: User-entity graph visualised with Gephi (Bastian et al., 2009) (positive edges). We apply a force atlas layout. Pink nodes are entities, blue nodes are users which we can see clustered around the target entities they expressed a positive stance for.

# Contributions

1. We offer a **simple, unsupervised method** to 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
4. We make **all our code and data available** at `https://github.com/isabellelorge/contradiction`

# Stance and entities

***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.*

- Stance != Sentiment
- Some concepts lend themselves better to eliciting stance
- NEs can elicit diverging intellectual or emotional viewpoints.
- Disagreement likely to crystallise around attitudes towards a few key entities

# Signed Graphs

- Graphs can be **directed vs. undirected; signed vs. unsigned, homogenous vs. bipartite**
- Our graph: **undirected, signed and bipartite**
- We use a **signed GCN** by Derr et al. (2018) based on *balance theory* (ie., friend of friend = friend; enemy of enemy = friend, etc.)

# Dataset

| | *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

**DEBAGREEMENT dataset (Pougué-Biyong et al., 2021): 42894 Reddit comment-reply pairs from 5 different subreddits with labels agree/disagree/neutral**

# User-Entity Graph Construction

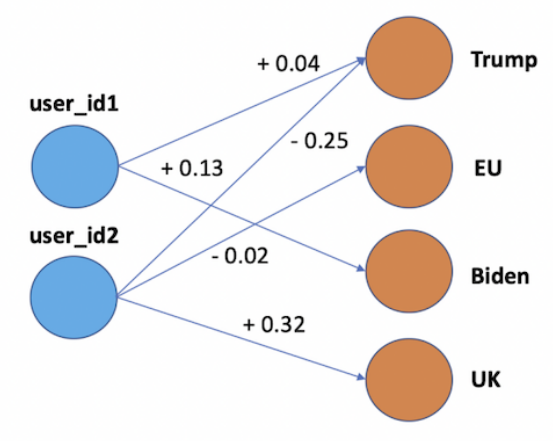


Figure 2: Example user-entity graph. The network is signed, with each edge representing user stance towards an entity.

$\mathcal{G} = (\mathcal{N}, \mathcal{E})$ signed undirected bipartite graph where $\mathcal{U} \in \mathcal{N}$ = set of user nodes, $\mathcal{A} \in \mathcal{N}$ = set of entity nodes and $\mathcal{E}$ = set of edges between users and entities, with $\mathcal{E}+$ the set of positive edges and $\mathcal{E}-$ the set of negative edges.

Bipartite graph = **no edges between users or between entities**, and set of positive and negative edges defined to be **mutually exclusive** (ie., at most one edge, either positive or negative, between a user and an entity)

# Graph dataset

| |
|---|
| *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

| | $\lvert\mathcal{U}\rvert$ | $\lvert\mathcal{A}\rvert$ | $\lvert\mathcal{E}+\rvert$ | $\lvert\mathcal{E}-\rvert$ | $\lvert\mathcal{D}\rvert$ | $\lvert\mathcal{D}(\mathcal{U})\rvert$ | $\lvert\mathcal{D}(\mathcal{A})\rvert$ | $\lvert\mathcal{CN}(\mathcal{U})\rvert$ | $\lvert\mathcal{CN}(\mathcal{A})\rvert$ |
|---|---|---|---|---|---|---|---|---|---|
| *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)

# STEntConv

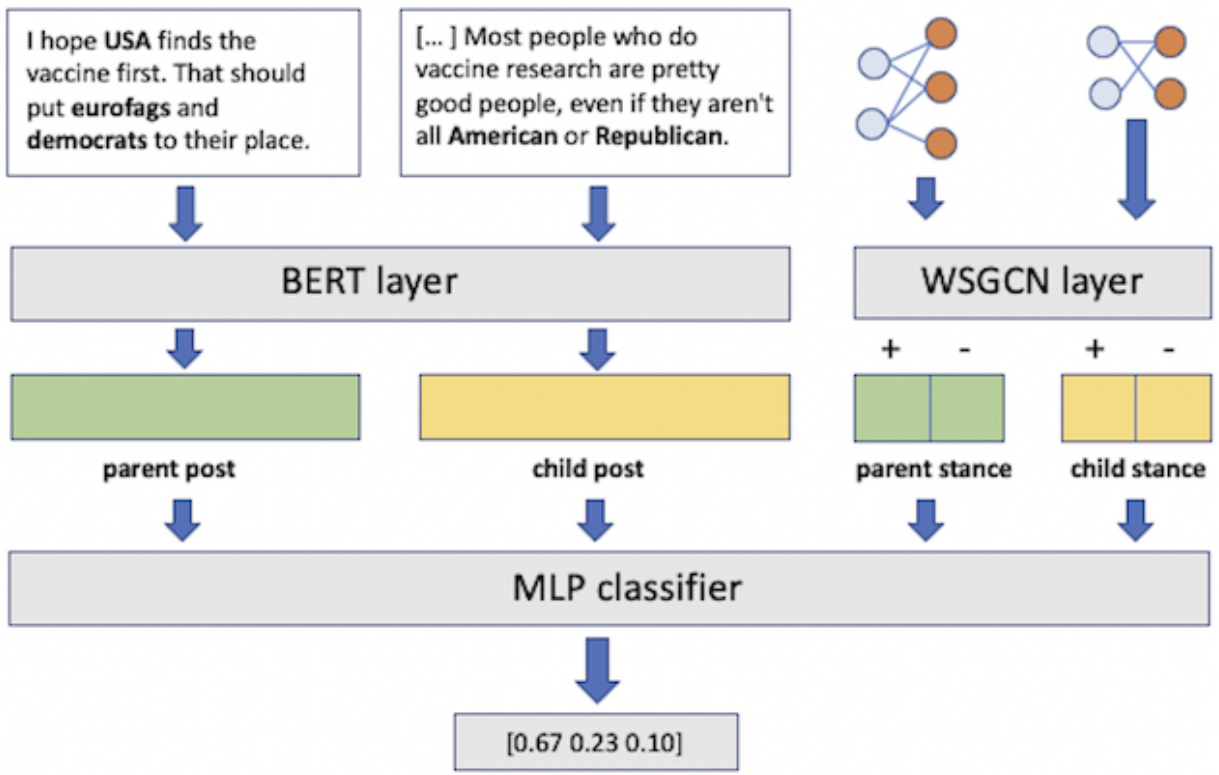


Figure 3: Model architecture.

# Baselines

- BERT
- Falcon
- StanceRel
- GCN only

# Training

- 100-dimensional word2vec embeddings trained on full dataset as initial features for our entities
- User features initialised as 100-dimensional vectors of zeros
- Cross entropy as loss function, batch size of 16, hidden size of 300 for first convolutional layer, learning rate of 3e-5, Adam optimiser with weight decay 1e-5.
- Data split into 0.80 train, 0.10 dev and 0.10 test
- Train for 6 epochs (models with BERT layers) and 11 epochs (GCN only)

# Results

| | *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.

# Confusion matrix

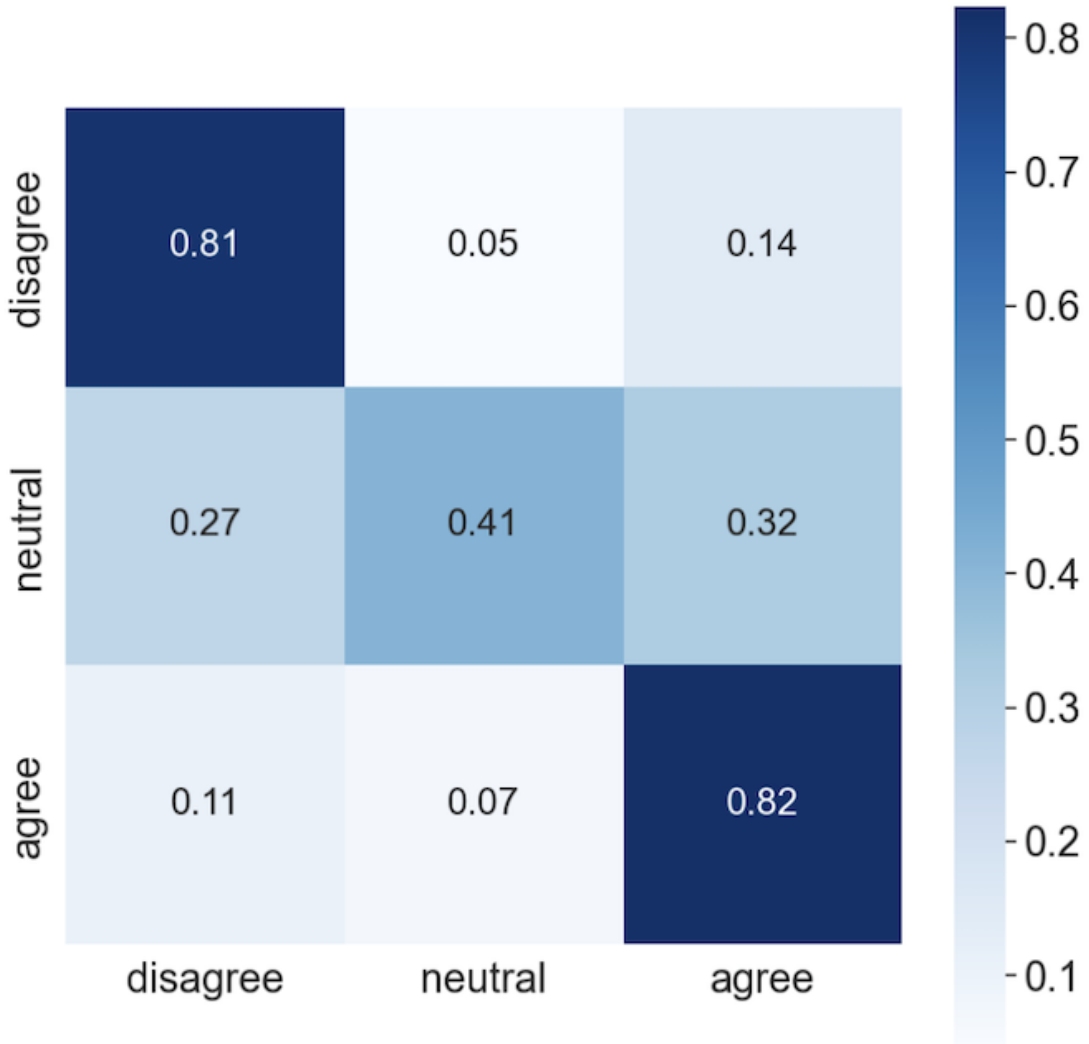


Figure 4: Confusion matrix (STEntConv)

# Conclusion

- **STEntConv outperforms all baselines**
- Possible to **add features which are not platform-specific** and **not relying on previous user-user history** to improve disagreement detection
- STENtConv also **outperforms model relying on user-user interaction history**
- **Multiple aggregations** not helpful here
- **Poor performance** of non fine-tuned LLM (Falcon)
- Potential to **combine features when available**, as most increase in performance only for posts mentioning target entities