Controversy Detection: a Text and Graph Neural Network Based Approach

Détection de la controverse : une approche basée sur les réseaux de neurones, appliquée aux graphes et aux textes

Benslimane Samy, Jérôme Azé, Sandra Bringay
Caroline Mollevi, Maximilien Servajean
samy.benslimane@lirmm.fr

LIRMM, University of Montpellier, Montpellier, France
AMIS, Paul Valery University, Montpellier, France,
Institut du Cancer Montpellier (ICM), Montpellier, France
IDESP, UMR Inserm - Université de Montpellier, Montpellier, France
Controversial topic: A controversial content can be defined as content which polarize attention into communities, stimulate interaction between them

- Detecting controversy: Prevent fake news / Identify hot topics / Evolution of controversy
Controversial topic: A controversial content can be defined as content which polarize attention into communities, stimulate interaction between them

- Detecting controversy: Prevent fake news / Identify hot topics / Evolution of controversy
- Social media: Public debate + different opinions

Tweet from Donald Trump about Mexico wall

With Mexico being one of the highest crime Nations in the world, we must have THE WALL. Mexico will pay for it through reimbursement/other.

8:44 AM - 27 Aug 2017
26,059 Retweets 98,660 Likes

Tweet about death penalty

@YesDeathPenalty
Because I don't believe in an eye for an eye.
6/6/16, 5:15 PM
Goal: Automatic detection of controversial topic on social media, using both structural and textual information

• Contribution
  • Graph Neural Network (GNN)-based controversy detection method
  • Experimental study: 2 different approaches, on real-world datasets
  • Incorporating textual features → To improve detection performance
Controversy detection/quantification: State-of-the-art

- **Content-based methods** → Based only on textual information & semantic
  - On Wikipedia: Use of word embeddings and sentence embeddings (word2vec, Bert)
    - + Apply models (Nearest Neighbors, LSTM, etc.)

```
“\text{I want to cancel my shoes order}”
```

[0.23, 0.56, 0.67, 0.97, 0.05, 0.98, …, 0.13]

---

Sznajder, B., Gera, A., Bilu, Y., Sheinwald, D., Rabinovich, E., Aharonov, R., Konopnicki, D., Slonim, N.: Controversy in context. CoRR (2019)

Dori-Hacohen, S., Jensen, D.D., Allan, J.: Controversy detection in wikipedia using collective classification. In: 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 797–800 (2016)

Jang, M., Allan, J.: Improving automated controversy detection on the web. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR. pp. 865–868. ACM (2016)

Jang, M., Foley, J., Dori-Hacohen, S., Allan, J.: Probabilistic approaches to controversy detection. In: 25th ACM International Conference on Information and Knowledge Management, CIKM. pp. 2069–2072 (2016)
Controversy detection/quantification: State-of-the-art

- **Structure-based methods** → Focus on user interactions

Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quantifying controversy on social media. ACM Trans. Soc. Comput. 1(1), 3:1–3:27 (2018)
Emamgholizadeh, H., Nourizade, M., Tajbakhsh, M.S., Hashminezhad, M., Esfahani, F.N.: A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). Soc. Netw. Anal. Min. 10(1), 90 (2020)
Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Reducing controversy by connecting opposing views. In: Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI. pp. 5249–5253 (2018)
Guerra, P.H.C., Jr., W.M., Cardie, C., Kleinberg, R.: A measure of polarization on social media networks based on community boundaries. In: Seventh International Conference on Weblogs and Social Media, ICWSM. The AAAI Press (2013)
Controversy detection/quantification: State-of-the-art

- **Hybrid methods** → Use both content and structural information

---

**Reddit**: sample of the comment-tree structure of a post. Both information, textual and structural, are available.

---

Zarate, J.M.O.D., Feuerstein, E.: Vocabulary-based method for quantifying controversy in social media. In: Ontologies and Concepts in Mind and Machine - 25th International Conference on Conceptual Structures, ICCS. Lecture Notes in Computer Science, vol. 12277, pp. 161–176. Springer (2020)

Hessel, J., Lee, L.: Something’s brewing! early prediction of controversy-causing posts from discussion features. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 1648–1659 (2019)

Zhong, L., Cao, J., Sheng, Q., Guo, J., Wang, Z.: Integrating semantic and structural information with graph convolutional network for controversy detection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL. pp. 515–526. Association for Computational Linguistics (2020)
Controversy detection: Pipeline

1. Graph Building

2. User feature extraction

3. Graph Embedding

4. Graph classification
Controversy detection: Pipeline

Stage 1. Create the Reddit user graph from the comment-tree structure

Stage 2. Create user node features from comments of each user

Stage 3. Represent a graph vector embedding, using 2 different approaches

Stage 4. Classify the embedding vector into controversial or not
Stage 1. Graph building

User graph of a controversial Reddit post, edges representing interactions between 2 users
Stage 2. Feature extraction

**BERT model:** transfer learning method based on transformers block

---

Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bi-directional transformers for language understanding. In: Proceedings of the NAACL-HLT Conference: Human Language Technologies, 2019, Volume 1. pp. 4171–4186. Association for Computational Linguistics (2019)
Stage 2. Feature extraction

**BERT model:** transfer learning method based on transformers block

- **PT:** last layer features of the pre-trained model (dim = 768)
- **FT_sentiment:** fine-tuned model with sentiment Reddit comment (dim = 64)
- **FT_itself:** fine-tuned with own comments, label depending on their respective post (dim = 64)

Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bi-directional transformers for language understanding. In: Proceedings of the NAACL-HLT Conference: Human Language Technologies, 2019, Volume 1. pp. 4171–4186. Association for Computational Linguistics (2019)
Stage 3. Graph embedding

APPROACH 1: Hierarchical learning representation (HRL-GCN)

- Pooling layer node representation

- Final graph representation

Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W.L., Leskovec, J.: Hierarchical graph representation learning with differentiable pooling. In: Annual Conference on Neural Information Processing Systems, NeurIPS
Stage 3. Graph embedding

**APPROACH 2: Attention-based representation (ARL-GAT)**

- At each Attention-layer $l$, for each node $i$

- At last layer $L$, learn graph embedding

$$z_G = \left\|_{l=0}^{(L)} \right( \text{READOUT} \left( \{h_i^{(l)} \mid u_i \in U \} \right) \right)$$

Zhang, S., Xie, L.: Improving attention mechanism in graph neural networks via Cardinality preservation. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 1395–1402. ijcai.org (2020)
Stage 4. Graph classification

Controversial

Non-controversial
Reddit dataset

Source: Real-world data from Reddit, collected by Hessel & Lee (2018)

- Divided into 6 datasets, corresponding to 6 subreddits (AM, AW, LT, RS, PF, FN):
  - N posts/threads by dataset
  - For each threads/post in a subreddit
    Comment-tree related to the post, w/ meta-data inside

Hessel, J., Lee, L.: Something's brewing! early prediction of controversy-causing posts from discussion features. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 1648–1659 (2019)
Reddit dataset

Source: Real-world data from Reddit, collected by Hessel & Lee (2018)

- Divided into 6 datasets, corresponding to 6 subreddits (AM, AW, LT, RS, PF, FN):
  - N posts/threads by dataset
  - For each threads/post in a subreddit
    Comment-tree related to the post, w/ meta-data inside

**Table 1.** Statistics on the 6 real-world balanced Reddit datasets.

|                       | AM   | AW   | FN   | LS   | PF   | RS   |
|-----------------------|------|------|------|------|------|------|
| Number of posts       | 3305 | 2969 | 3934 | 1573 | 1004 | 2248 |
| Average number of users by post | 72   | 67   | 76   | 79   | 47   | 48   |
| Average number of comments by post | 144  | 141  | 159  | 132  | 95   | 98   |

Hessel, J., Lee, L.: Something's brewing! early prediction of controversy-causing posts from discussion features. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 1648–1659 (2019)
Experiments set-up

• **Dataset**
  - Train/test set: 80/20%, for each of the 6 subreddit datasets

• **Baseline**
  - POST (Text+Time): only focus on the post (w/ bert) *
  - (C-{Text Rate Tree} + Post): structural + text features *
  - (DTPC-GCN): GNN model based **

• **HRL-GCN**
  - 4-layer GCN per pooling layer, with 1 and 2 pooling layers

• **ARL-GAT**
  - 2 node aggregators: Mean/sum

---

* Hessel, J., Lee, L.: Something’s brewing! early prediction of controversy-causing posts from discussion features. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, pp. 1648–1659 (2019)

** Zhong, L., Cao, J., Sheng, Q., Guo, J., Wang, Z.: Integrating semantic and structural information with graph convolutional network for controversy detection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL. pp. 515–526. Association for Computational Linguistics (2020)
Results

**Table 1.** Performance comparison of our GNN-based controversy detection with baseline. Performance is evaluated using accuracy of the validation set.

| Model | AM  | AW  | FN  | LS  | PF  | RS  |
|-------|-----|-----|-----|-----|-----|-----|
| POST (Text+Time) | 68.1 | 65.4 | 65.5 | 66.2 | 66.5 | 69.3 |
| DTPC-GCN | | | | | | 67.6 |
| POST + C-{Text_Rate_Tree} < 1 hour | 71.1 | 70 | 68.1 | 67.9 | 66.1 | 65.5 |
| POST + C-{Text_Rate_Tree} < 3 hours | **74.3** | 72.3 | 70.5 | **71.8** | **69.3** | **67.8** |
| ARL-GAT (Mean-aggr) | 65.7 | 69.2 | **72.4** | 58.4 | 53.7 | 62.9 |
| ARL-GAT (Sum-aggr) | 67.5 | 71 | 72.2 | 67 | 63.7 | 51.8 |
| HRL-GCN (pool=2) | 69 | 72.2 | 71.7 | 68.3 | 65.7 | 63.6 |
| HRL-GCN (pool=1) | 69.6 | **74.6** | 72.2 | 67.9 | 68.2 | 66.7 |

**Table 2.** Performance of our best GNN approach enriched with different user text embeddings as initial node features.

| Model | AM  | AW  | FN  | LS  | PF  | RS  |
|-------|-----|-----|-----|-----|-----|-----|
| HRL-GCN (pool=1) | 69.6 | **74.6** | **72.2** | 67.9 | 68.2 | **66.7** |
| + PT | **70.8** | 73.7 | 71 | 65.4 | **70.6** | 64.7 |
| + FT_SENTIMENT | 69.1 | 72.9 | 70.5 | **68.6** | 66.7 | 64 |
| + FT_ITSELF | 67.3 | 73.9 | 71.8 | 68.3 | **70.6** | 63.8 |
Future Work

- Work on controversy quantification, on different social media
- Node text representation improvement
- Look at quantifying controversy over time, and how to reduce controversy on topics

M. Simonovsky and N. Komodakis. Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 29–38, Honolulu, HI, July 2017. IEEE.
Future Work

- Work on controversy quantification, on different social media
- Node text representation improvement
- Look at quantifying controversy over time, and how to reduce controversy on topics

Thank you for your attention

Samy Benslimane, samy.benslimane@lirmm.fr

Controversy Detection: a Text and Graph Neural Network Based Approach

M. Simonovsky and N. Komodakis. Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 29–38, Honolulu, HI, July 2017. IEEE.
**Comment-availability across time**

![Graph showing comment availability across time]

**Fig. 4.** Impact of comments availability on controversy detection performance.
Stage 3. Graph embedding

**APPROACH 2: Attention-based representation**

At each Attention-layer \( l \), for each node \( i \)

1. Learn attention score

\[
e_{u_iu_j}^{(l)} = a \left( W^{(l)}h_{u_i}^{(l)}, W^{(l)}h_{u_j}^{(l)} \right)
\]

2. Normalize score

\[
\alpha_{u_iu_j}^{(l)} = \text{softmax}(e_{u_iu_j}^{(l)}) = \frac{\exp(e_{u_iu_j}^{(l)})}{\sum_{u_k \in \mathcal{N}(u_i)} \exp(e_{u_iu_k}^{(l)})}
\]

3. Learn new node representation

\[
h_{u_i}^{(l+1)} = \sigma \left( \sum_{u_j \in \mathcal{N}(u_i)} \alpha_{u_iu_j}^{(l)} W^{(l)}h_{u_j}^{(l)} \right)
\]

4. At last layer \( L \), learn graph embedding

\[
z_G = \left\| \bigwedge_{l=0}^{(L)} \text{READOUT} \left( \{ h_{u_i}^{(l)} \mid u_i \in U \} \right) \right\|
\]

Zhang, S., Xie, L.: Improving attention mechanism in graph neural networks via Cardinality preservation. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 1395–1402. ijcai.org (2020)