Perceived and Intended Sarcasm Detection with Graph Attention Networks

Joan Plepi and **Lucie Flek**
Conversational AI and Social Analytics (CAISA) Lab
Department of Mathematics and Computer Science, University of Marburg
http://caisa-lab.github.io

Abstract

Existing sarcasm detection systems focus on exploiting linguistic markers, context, or user-level priors. However, social studies suggest that the relationship between the author and the audience can be equally relevant for the sarcasm usage and interpretation. In this work, we propose a framework jointly leveraging (1) a user context from their historical tweets together with (2) the social information from a user's conversational neighborhood in an interaction graph, to contextualize the interpretation of the post. We use graph attention networks (GAT) over users and tweets in a conversation thread, combined with dense user history representations. Apart from achieving state-of-the-art results on the recently published dataset of 19k Twitter users with 30K labeled tweets, adding 10M unlabeled tweets as context, our results indicate that the model contributes to interpreting the sarcastic intentions of an author more than to predicting the sarcasm perception by others.

1 Introduction

Sarcasm is a form of non-literal language, in which the intended meaning of the utterance differs from the literal meaning, fulfilling a social function in a discourse (Dews et al., 1995; Riloff et al., 2013). Sarcasm detection poses a challenge for numerous NLP tasks, such as sentiment or stance prediction (Maynard and Greenwood, 2014).

Early sarcasm detection systems are based on lexical and syntactic cues (Carvalho et al., 2009; Davidov et al., 2010; Tsur et al., 2010; González-Ibáñez et al., 2011; Reyes et al., 2013; Ghosh et al., 2015). However, sarcasm interpretation requires context, even for humans (Wallace et al., 2014). More recent works hence incorporate discourse information such as contrast (Riloff et al., 2013; Khattri et al., 2015; Joshi et al., 2015; Rajadesingan et al., 2015; Tay et al., 2018), and contextualize the post by using features from user history (Bamman and Smith, 2015; Amir et al., 2016; Oprea and Magdy, 2019; Hazarika et al., 2018). The relationship between an author and the audience has been given comparably less attention, despite its relevance for the sarcasm interpretation (Rockwell and Theriot, 2001; Gibbs, 2000; Dress et al., 2008; Marwick and Boyd, 2011; Bamman and Smith, 2015). In this work, we propose a graph neural network framework jointly leveraging a user context from their historical tweets together with the social information from a user's neighborhood modeled by heterogeneous graph structures.

The key contributions of this paper are:

(1) We present the first graph attention-based model to identify sarcasm on social media by explicitly modeling users' social and historical context jointly, capturing complex relations between a sarcastic tweet and its conversational context.

(2) We demonstrate that exploiting these relationships increases performance in the sarcasm detection task, reaching state-of-the-art results on the recent SPIRS dataset (Shmueli et al., 2020), which we expand with user history. We examine the impact of different parts of the context, captured by attention weights, in modeling sarcastic utterances.

(3) We find that even with user-based models, detecting sarcastic intentions of the author is easier than identifying the sarcasm perception by others.

2 Related Work

Leveraging user history Several previous works contextualize a sarcastic post by using features from user history - employing past tweets to identify a user's behavioral traits (Rajadesingan et al., 2015), encoding user sentiment priors over different entities (Khattri et al., 2015), or manually crafting user interaction features (Bamman and Smith, 2015). Amir et al. (2016) introduce the user2vec model, applying paragraph2vec (Le and Mikolov, 2014) over user history. Hazarika et al. (2018) propose an alternative user embedding approach, encoding style and personality features.

Leveraging user network An emerging line of research makes use of social interactions to encode information about the user induced by neural architecture (Grover and Leskovec, 2016; Qiu et al., 2018). Network information improves performance on detecting cyberbullying (Mathew et al., 2019), abusive language use (Qian et al., 2018), suicide ideation (Mishra et al., 2019) or fake news (Chandra et al., 2020). To the best of our knowledge, graph network based approaches have not been used in the sarcasm detection task so far.

Perceived and intended sarcasm Perceiving sarcasm in text is not trivial even for humans, not only due to the lack of acoustic markers (Bänziger and Scherer, 2005; Woodland and Voyer, 2011) but also due to the sociocultural diversity (Rockwell and Theriot, 2001; Dress et al., 2008) where in many cases the audience may misinterpret a sarcastic statement as sincere. This has been only recently reflected in sarcasm detection models (Hazarika et al., 2018; Shmueli et al., 2020).

3 Proposed Approach

3.1 Tweet Embeddings

We denote the current tweet to be assessed $t_i \in T = \{t_1, t_2, \dots, t_N\}$, where N is the total number of tweets. We utilize SentenceBERT embeddings (Reimers and Gurevych, 2019) to encode the tweets. Formally, $\mathbf{t_i}' = SentenceBERT(t_i)$ where $\mathbf{t_i}' \in \mathbb{R}^{768}$, and SentenceBERT computes the mean of all tokens' representation. We forward this representation into a linear layer to transform in dimension d, $\widetilde{\mathbf{t_i}} \in \mathbb{R}^d$.

3.2 User Embeddings (Historical Context)

Let $u^{t_i} \in U = \{u^{t_1}, u^{t_2}, \dots, u^{t_M}\}$ be the author of tweet t_i, from now on we keep only the index i for brevity. Each user u^i is associated with a set of historical tweets $\mathcal{H}^i = \{(H^i_1, \tau^i_1), \dots, (H^i_m, \tau^i_m)\}$, where H^i_j is a historic tweet posted at a time τ^i_j by the user u^i. We adopt user2vec (Amir et al., 2016) to compute the initial user representation $\widetilde{\mathbf{u}}_i \in \mathbb{R}^d$ of user u^i based on their corresponding historical tweets $\mathcal{H}^i$, optimizing the conditional probability of texts given the author.

3.3 Social Graph (Network Context)

Apart from the importance of surrounding context to understand sarcasm (Wallace et al., 2014), certain understanding is needed between the audience and the author (Gibbs, 2000; Dress et al., 2008). Our goal is to model relations between users and their past tweets, interactions between users, and relations between tweets in one conversation. We model these relationships as a graph $\mathcal{G} = (V, E)$, where $V = \{U \cup T\}$ contains two types of nodes - Users and Tweets (Figure 1). We use three edge types $E = \{\mathrm{e}^U \cup \mathrm{e}^T \cup \mathrm{e}^C\}$, where e^U represents the social interaction between users. This involves quotes, mentions, or replies in the user history. e^T denotes the edges between tweets that are involved in one discussion thread, with all tweets connected with each other, and e^C is the relation between a tweet and its author.

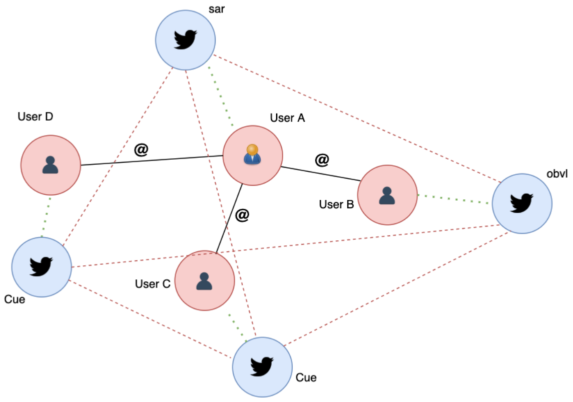

Figure 1: An example of a heterogeneous user and tweet social graph extracted from one conversation.

Representation Learning: We use Graph Attention Networks (GATs, (Veličković et al., 2018)) to exploit the neighborhood of each node to compute the final representations.[1] GAT uses a self-attention mechanism (Bahdanau et al., 2015; Vaswani et al., 2017) to assign an importance score to the connections that contribute more to the detection of sarcastic or non-sarcastic tweets. We initialize the user and the tweet nodes of the GAT with their corresponding embeddings $\widetilde{\mathbf{u}}_i$ and $\widetilde{\mathbf{t}}_i$. The initial node representation of each node $v \in V$ is linearly transformed by a weight matrix $\mathbf{W} \in \mathbb{R}^{d' \times d}$ into a vector $\mathbf{h}_v \in \mathbb{R}^{d'}$. Following, the attention weights e_{vn} of each node v are computed as:

$$e_{vn} = att(\mathbf{h}_v \| \mathbf{h}_n) \tag{1}$$

where $n \in \mathcal{N}(v)$ is a node in the neighborhood of v and att is the attention mechanism function

[1] We ran early experiments with Graph Convolutional Networks as well, obtaining inferior and less interpretable results.

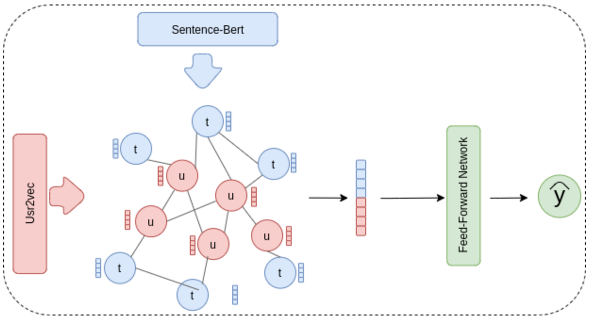

Figure 2: The social graph is initialized with user and tweet embeddings (user2vec and sentence-BERT), and tuned by GAT to take into account relationships between them. The output representations are then fed into the classification layer.

which is a single-layer feedforward neural network, parameterized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2d'}$ with a LeakyReLU nonlinearity.

The final node representation $\mathbf{h}'_v \in \mathbb{R}^{K \cdot d'}$ is computed as:

$$\mathbf{h}'_v = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{n\in\mathcal{N}(v)}\alpha^k_{vn}\mathbf{W}^k\mathbf{h}_n\right) \tag{2}$$

where K is the number of attention heads, σ is the ReLU nonlinear function, $\mathbf{W}^k \in \mathbb{R}^{d' \times d'}$ a weight matrix and $\alpha^k_{vn} = softmax(e^k_{vn})$ the normalized attention weights from the k-th attention mechanism att^k.

3.4 Classification model

The user and tweet representations learned by GAT layer are concatenated and forwarded through a two-layer feed-forward network parameterized by weight matrices $\mathbf{W}^c_1 \in \mathbb{R}^{d_1 \times 2d'}$ and $\mathbf{W}^c_2 \in \mathbb{R}^{o \times d_1}$, where d_1 is the dimension of projected embeddings, and o is equal to the number of classes. The final prediction of the model is given by:

$$\hat{y} = softmax\left(\mathbf{W}^c_2\left(\sigma\left(\mathbf{W}^c_1[h_t||h_u]\right)\right)\right) \tag{3}$$

4 Experimental Setup

4.1 Dataset

For our experiments, we use a recently published SPIRS sarcasm dataset (Shmueli et al., 2020). It utilizes *cue tweets*, conversation replies which point out the sarcastic nature of a previous post. In addition, the dataset also provides *oblivious tweets*, questioning the sarcastic nature of a given example, and *elicit tweets*, being the original start of the conversation. Non-sarcastic posts were collected randomly in equal numbers. The labeled dataset contains in total 15,000 sarcastic tweets (10,000 self-reported and 5000 perceived cues), 15,000 non-sarcastic, 10,000 oblivious and 9156 elicit tweets.

User context We extend SPIRS with over 10 million past tweets of the authors in the dataset in order to compute the user embeddings.

Social network Our graph consists of the three types of connections described in Sec.3.3.To avoid the bias coming from cue tweets, we exclude these from our graph. Our final social network consists of 108K nodes with 0.00002 density and 32% homophily, defined as the percentage of connections between authors of tweets with the same label.

4.2 Comparison Baselines

The baselines introduced by (Shmueli et al., 2020) are a Convolutional Neural Network, a Bidirectional LSTM, and a fine-tuned pre-trained BERT model. We compare our model with BERT, which performs the best of these. We add two baselines which incorporate user information. First, we extend BERT by simply concatenating the tweet embeddings with their respective user2vec author representation ('BERT + user2vec'). As a second baseline ('BERT + user-only GAT'), we build a social graph with only user nodes and their interactions (quotes, mentions, or replies) e^U as edges, and apply the GAT initialized with user2vec embeddings. The implementation of the models and the results are made publicly available, to facilitate reproducibility and reuse[2].

5 Results and Analysis

Our proposed GAT base model significantly outperforms all the baselines (Table 1) despite having fewer trainable parameters (500K) than the BERT model (110M). First, by simply concatenating the user2vec embeddings to BERT, we obtain 3.4% f1 score improvement on the BERT model, indicating the importance of user context in sarcasm detection. Moreover, we introduce the GAT module in the model. We first experiment with only tweet to tweet connections in the graph based on the conversations on Twitter and trained on top of

[2] https://github.com/caisa-lab/sarcasm_detection

Sarcasm Detection			
Model	**P**	**R**	**F1**
BERT	70.1%	69.7%	69.9%
BERT + user2vec	73.6%	73.2%	73.4%
BERT + tweet-tweet GAT	70.4%	69.9%	70.1%
BERT + user-only GAT	74.2%	78.1%	76.1%
User+tweet GAT (no cues)	**84.7%**	**83.7%**	**84.2%**
User+tweet GAT, no elicit	83.2%	80.8%	82.0%
User+tweet GAT, no oblivious	82.4%	80.4%	81.4%
User+tweet GAT + cue tweets	94.7%	94.3%	94.5%

Table 1: Mean overall precision (P), recall (R), and F1 score (F1) of each model over 10 runs with varying seeds, detecting sarcasm on the SPIRS dataset.

the fine-tuned BERT. In this case, the GAT layer only bring 0.2% improvement due to the sparse and disconnected nature of the constructed graph. In addition, we replace user2vec with GAT embeddings tuned on user-only social graph, and we achieve 6.1% improvement on BERT and 3% over 'BERT + user2vec', presumably thanks to exploiting the homophily relations between users. Finally, applying GAT on the full heterogeneous user and tweet graph (as per Figure 2) provides a large performance boost thanks to incorporating the conversational thread context between tweets.

User representation We compare the initial user embeddings initialized by user2vec with the final representations computed from the GAT. The representations are projected in 2-dimensional space using T-SNE (Van der Maaten and Hinton, 2008). In Figure 3 and 4 we visualize the initial representations with user2vec and computed representation by GAT layer respectively. While in user2vec representations sarcastic users cannot be distinguished from non-sarcastic ones, in the GAT representations we can observe communities of users sharing the same sarcastic tendency.

Conversation context For comparison, we construct three more social graphs where: 1) We remove the elicit tweets which triggered the sarcastic comment (GAT - elicit tweets), 2) We remove the oblivious tweets which interpreted the comment as serious (GAT - oblivious tweets), 3) We add the original cue tweets, revealing that the post was sarcastic (GAT + cue tweets). As expected, adding the cue tweets in the social graph leads to an almost perfect F1 score of 94.5%. Removing oblivious and elicit tweets causes just a small performance drop (2-3%). In the way the SPIRS dataset is annotated, an oblivious tweet typically triggers a cue tweet ("*c mon, dude, it was just sarcasm*"). We hypothesize that even with the cue tweets removed, the model is able to learn the predictive relation between oblivious and sarcastic tweets. This is in line with the original paper (i.e. without user context), where a 3.4% drop in prediction accuracy was observed, when the oblivious tweets were removed.

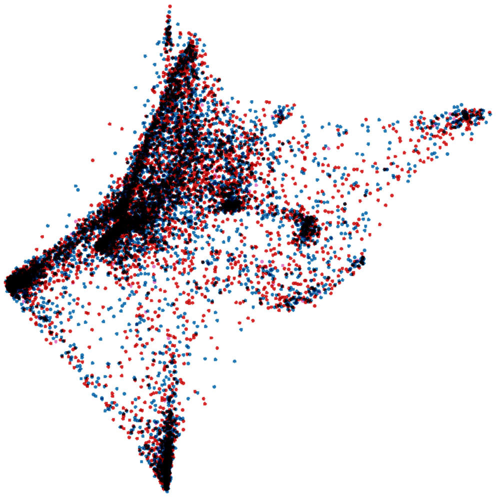

Figure 3: Initial representations of users (user2vec) projected in 2D space with T-SNE. Red color denotes sarcastic users, blue non-sarcastic.

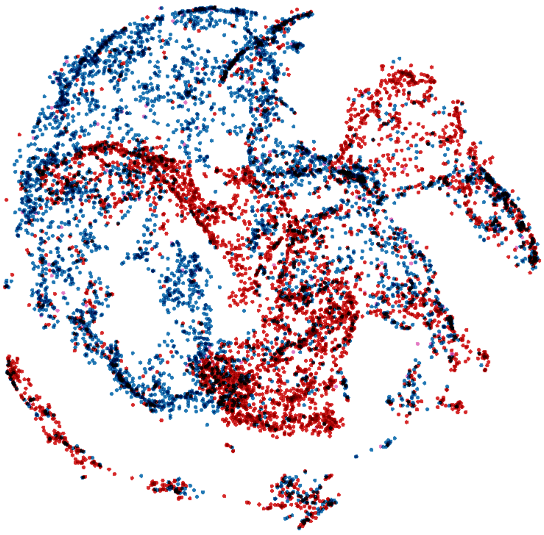

Figure 4: Learned representations by our social network module (GAT) projected in 2D space with T-SNE. Red color denotes sarcastic users, blue non-sarcastic.

Attention weights The attention mechanism of GAT is able to assign varied weights to different

Sarcasm Perception			
Model	**P**	**R**	**F1**
BERT	73.2%	68.0%	69.0%
User+tweet GAT (no cues)	75.0%	67.7%	71.2%

Table 2: Mean overall precision (P), recall (R), and F1 score (F1) over 10 runs classifying self-reported (intended) and perceived sarcasm on the SPIRS dataset.

nodes in the neighborhood, dynamically encoding of the user by their homophily relations, which boosts the effect of authors in tweet representations (Flek, 2020). We confirm this by examining users with a larger number of tweets in the dataset. When users tend to be sarcastic in most of the posts, the attention weight of their non-sarcastic tweets is smaller. In these cases, the attention weights give more importance to the surrounding user context over the conversation thread. Overall, the largest source of information for the model are the user nodes and the tweet that is being classified. We note the normalized attention weights are smaller for the oblivious and elicit tweet edges, and higher for the edges that connect tweets with their respective author. In other words, the conversational context only plays a decisive role in case of insufficient or inconsistent user-level priors.

Sarcasm Perception Cue tweets can be either authored by the same user as the sarcastic post (*intended sarcasm*) or a different one (*perceived sarcasm*).We observe that in the sarcasm detection task, the error rate on perceived sarcasm is 20% while in the self-reported sarcasm it is only around 15%. We therefore test our model on distinguishing between perceived and self-reported sarcasm. Our GAT model brings an improvement of 2.2% over the BERT baseline, with the perceived sarcasm being harder to detect (F1 56%) than the self-reported one (F1 84%). These results are aligned with the conclusions from (Oprea and Magdy, 2019). In most cases, the perceived sarcasm is misclassified as self-reported, which is present more often (70%) in the data. Perceived sarcasm is dependent on the readers rather than the author of the tweet, therefore we hypothesize that modeling the authors' context is less useful. It could be of benefit to model more robust recipient user profiles as well, to better predict how each individual will react.

Limitations Modeling the social networks with GAT is affected by several factors. First, the low graph density, as the original dataset wasn't collected by following relationships between users, hence many users across different conversation threads are not related to each other. Second, the homophily degree is only 32%, users with sarcastic tendency have few connections among them.

6 Conclusions

In this work, we explore social networks of user interactions, and contextual information to interpret sarcastic intentions in social media. We propose a graph attention-based model, which combines contextual information of users, linguistic features, and social networks. The heterogeneous social network modeling dynamically exploits relationships between users and tweets in a conversation and significantly improves the state-of-the-art results.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) as a part of the Junior AI Scientists program under the reference 01-S20060. We would like to thank Flora Sakketou and all the anonymous reviewers for their valuable input.

Ethical Considerations

The ability to automatically approximate personal characteristics of online users in order to improve natural language classification algorithms requires us to consider a range of ethical concerns, including: (1) privacy and user consent, (2) representativeness of the data for generalization, and (3) user vulnerability to a potential model or data misuse or misinterpretation.

Use of any user data for personalization shall be transparent, and limited to the given purpose, no individual posts shall be republished (Hewson and Buchanan, 2013). Researchers are advised to take account of users' expectations (Williams et al., 2017; Shilton and Sayles, 2016; Townsend and Wallace, 2016) when collecting public data such as Twitter. In this case, when we expand the original dataset with more extensive user history, we utilize publicly available Twitter data in a purely observational (Norval and Henderson, 2017), and non-intrusive manner. All user data is kept separately on protected servers, linked to the raw text and network data only through anonymous IDs.

Shah et al. (2020) identify four different sources of bias in NLP models: selection bias, label bias, model overamplification, and semantic bias. While we can't exclude any of those, the selection bias should be kept in mind in particular, when reusing the presented model, as it is unclear to which extent the augmented SPIRS dataset with user history represents a sample of the overall population on Twitter. The user selection was based solely on the available sarcasm annotations, and doesn't include any sociodemographic information.

In addition, any user-augmented classification efforts risk invoking stereotyping and essentialism, as the algorithm may lean towards label people rather than posts (e.g. "this is a sarcastic person"). Such stereotypes can cause harm even if they are accurate on average differences (Rudman and Glick, 2008). These can be emphasized by the semblance of objectivity created by the use of a computer algorithm (Koolen and van Cranenburgh, 2017). It is important to be mindful of these effects when interpreting the model results in an own end-application context.

References

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177, Berlin, Germany. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9.

Tanja Bänziger and Klaus R. Scherer. 2005. The role of intonation in emotional expressions. *Speech Communication*, 46(3):252–267. Quantitative Prosody Modelling for Natural Speech Description and Generation.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy";-). In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, TSA '09, page 53–56, New York, NY, USA. Association for Computing Machinery.

Shantanu Chandra, Pushkar Mishra, Helen Yannakoudakis, Madhav Nimishakavi, Marzieh Saeidi, and Ekaterina Shutova. 2020. Graph-based modeling of online communities for fake news detection. *arXiv preprint arXiv:2008.06274*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Shelly Dews, Joan Kaplan, and Ellen Winner. 1995. Why not say it directly? the social functions of irony. *Discourse Processes*, 19(3):347–367.

Megan L. Dress, Roger J. Kreuz, Kristen E. Link, and Gina M. Caucci. 2008. Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27(1):71–85.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Lisbon, Portugal. Association for Computational Linguistics.

Raymond W. Gibbs. 2000. Irony in talk among friends. *Metaphor and Symbol*, 15(1-2):5–27.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for 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*, KDD '16, page 855–864, New York, NY, USA. Association for Computing Machinery.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Claire Hewson and Tom Buchanan. 2013. Ethics guidelines for internet-mediated research. The British Psychological Society.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. 2015. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 25–30, Lisboa, Portugal. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam, a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, volume 1412.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Alice E. Marwick and Danah Boyd. 2011. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 173–182, New York, NY, USA. Association for Computing Machinery.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4238–4243, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 147–156, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Norval and Tristan Henderson. 2017. Contextual consent: Ethical mining of social media for health research. *CoRR*, abs/1701.07765.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, New Orleans, Louisiana. Association for Computational Linguistics.

Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 459–467, New York, NY, USA. ACM.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 97–106, New York, NY, USA. Association for Computing Machinery.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Patricia Rockwell and Evelyn M. Theriot. 2001. Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, 18(1):44–52.

Laurie A Rudman and Peter Glick. 2008. The social psychology of gender: How power and intimacy shape gender relations.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Katie Shilton and Sheridan Sayles. 2016. " we aren't all going to be on the same page about ethics": Ethical practices and challenges in research on digital and social media. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1909–1918. IEEE.

Boaz Shmueli, Lun-Wei Ku, and Soumya Ray. 2020. Reactive Supervision: A New Method for Collecting Sarcasm Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2553–2559, Online. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics. *University of Aberdeen*, 1:16.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm — a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):162–169.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. Association for Computational Linguistics.

Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

Jennifer Woodland and Daniel Voyer. 2011. Context and intonation in the perception of sarcasm. *Metaphor and Symbol*, 26(3):227–239.

Appendix

A Configurations

We perform a stratified 90/10 train-test split. We sample 10% of the training data for validation. All splits have the same class distribution and different sets of tweet authors. We use 3 GAT layers, with number of heads $K = 4$. The initial dimension is $d = 400$ and the final output dimension $d' = 100$. To train our model we set learning rate to $1e - 4$, and dropout 0.4 (Srivastava et al., 2014), and use the Adam optimization algorithm (Kingma and Ba, 2015) for 500 training epochs with early stopping. For the GAT layers, we compute the mean of the outputs from each attention head instead of concatenation. All experiments are run in Nvidia A100 40 GB GPUs.

B User Context

To incorporate user context, we first extract all user IDs for all the tweets in the dataset. In the dataset, due to different tweet types with different users, we get in total 57K users. We fetch the tweet post timeline for each user, and we end up with a total of 104M tweets, in average 1800 posts per user. For user2vec training, we take into account only the users with a minimum of 50 posts in their timeline, and we limit the total number of posts to 1000. After filtering, the amount of tweets in the context is 10M. Every tweet is pre-processed by removing all links, user mentions are replaced with "@user", emojis and hashtags are cleared. We train user2vec (Amir et al., 2016) for 12 epochs, with learning rate 1e-4. For those users which are filtered, or we cannot extract history, we initialize them as the mean representation of his user neighbors in the social network. We used the history tweets only for creating the user-to-user edges, and those are not present in the constructed graph, but are already encoded in the initial user representation. We experimented with various history length settings, and found almost no difference in the performance between using interactions throughout all history and interactions during the last year. Hence, we omitted older interactions to ease the computations.