Improving Misinformation Detection in Tweets with Abstract Argumentation

Lars Malmqvist[1], Tommy Yuan[1] and Peter Nightingale[1]

[1]Department of Computer Science, University of York, Deramore Lane, York YO10 5GH, UK

Abstract

This paper presents a new approach to improving misinformation detection in Twitter data by augmenting approaches based on massive-scale language models such as BERT or GPT-3 with a graph representation based on abstract argumentation. We outline an approach that uses stance information to construct and annotate argumentation graphs using a variety of schemes and combine this with linguistic features from language models using a graph neural network architecture for binary classification. We outline the planned experiments for this new approach, using benchmark datasets, Pheme and RumourEval, that are common in the literature. This paper presents work in progress that seeks to contribute not only by improving misinformation detection, but also by combining abstract argumentation and linguistic data in a new way.

Keywords

Abstract Argumentation, Graph Neural Networks, Misinformation Detection

1. Introduction

The problem of detecting misinformation in social network data has received increasing attention recently, not least due to the COVID-19 public health emergency [1]. It is fair to say that deep learning models based on a variety of architectures have been increasingly successful in learning to detect various types of misinformation such as rumours, fake news, and the intentional spreading of false information [7].

For Twitter data two of the most successful recent approaches have been one the one hand detection based on large-scale language models [10], using the characteristics of the language used to determine veracity, and on the other methods that use the features of the propagation graph by which a source tweet is retweeted by other actors on the social network [11], using the structure of the graph as an indicator of veracity.

Our approach seeks to combine and enrich these two approaches by combining graph structure, in the form of an abstract argumentation framework derived from stance information with linguistic node level features from a language model and argumentative features based on the acceptability status of arguments under different semantics.

This paper presents our work-in-progress towards this goal including the planned methodology and experimental setup. This research builds on earlier work presented at SAFA 2020 [13],

CMNA'21: Workshop on Computational Models of Natural Argument, September 2-3, 2021, Online

lm1775@york.ac.uk (L. Malmqvist); tommy.yuan@cs.york.ac.uk (T. Yuan); peter.nightingale@york.ac.uk (P. Nightingale)

http://www-users.cs.york.ac.uk/~tommy/ (T. Yuan); https://www-users.cs.york.ac.uk/pwn/ (P. Nightingale)

where we presented a way to approximate acceptability in abstract argumentation frameworks using graph neural networks. While the approximation task was different in that example (e.g. determining acceptability), the same GCN architecture can apply in this case.

As part of this research, we plan to make the following contributions:

- Show that adding the structure of an argumentation graph can improve detection of misinformation relative to baseline language models
- Determine the optimal way to construct abstract argumentation graphs from stance information
- Determine the best architectures for combining linguistic and argumentative features for input into Graph Convolutional Networks (GCN)

2. Background

2.1. Abstract Argumentation

Abstract Argumentation is a formalism for non-monotonic reasoning that bases its representation on the modelling of conflict [4]. It is represented in the form of a directed graph in which vertices represent arguments and edges a relation of attack.

Definition 1 (Argumentation Framework). *Formally, an argumentation framework is a tuple ($F = \langle args, atts \rangle$) in which args is a finite set of arguments and $atts \subseteq args \times args$ defines a relation of attack.*

The notion of acceptability is key to abstract argumentation.

Definition 2 (Acceptability). *An argument $A \in args$ is called acceptable with respect to an extension $ext \subseteq args$ iff for every $B \in args$ with B attacks A there is an argument $A' \in ext$ with A' attacks B.*

Extensions are subsets of argumentation frameworks that are collectively acceptable. Extensions are evaluated based on semantics that define rules for what arguments can be accepted together [14]. In this research, we will construct Argumentation Frameworks based on stance information, using a variety of schemas and determine their acceptability status using an abstract argumentation solver as an input to a Graph Convolutional Network.

2.2. Graph Convolutional Networks

Convolutional Neural Network (CNN) models have been highly successful in computer vision tasks. Graph Convolutional Networks seek to apply this kind of model to graph structured data. In the original model proposed by Kipf et al. [8], the convolutional operator is modelled by an 1-hop aggregation of neighbourhood information within the graph. While simple, this approach has proven successful in a number of context, including the approximation of acceptability for abstract argumentation. In this work, we will use an adapted version of the GCN that we have previously presented [12] for this purpose, changing the classification task but retaining much of the architecture.

2.3. Misinformation Detection on Social Networks

There is a large extant literature on misinformation detection in its various guises. The two approaches most directly relevant to our research are those that rely on large-scale language models such as GPT-3 [5] and BERT [3] to analyze tweet content [10] and those that use the patterns of tweet propagation to detect misinformation [11].

However, the most directly parallel work in the literature is found in Toni and Orascu's [2] use of argumentative features to improve a Bi-LSTM model for detection of fake reviews and to determine whether news headlines support tweets. This research showed the power of argumentative features for this problem. However, compared to our research, it relies on a more complex formalism, bipolar argumentation frameworks, and uses the argumentative features as an adjunct to a principally NLP-based model instead of having a primary focus on graph structures.

3. Method

Our approach to the problem follows a four step process:

1. Construct an abstract argumentation graph based on stance information
2. Generate linguistic features by creating embeddings for the input tweets, using a language model
3. Generate argumentative features by resolving the acceptability of the arguments in the argumentation framework
4. Train a Graph Convolutional Network using the argumentation graph and the generated features inserted at the node level

In the following section, we will explore each of these steps in more detail.

3.1. Constructing the Argumentation Framework

We will construct the argumentation framework based on existing stance information present in our source datasets. The problem of stance detection has a substantial literature of its own [9] and we do not seek to add to it with this research.

While stance categorization can vary between datasets it is generally possible to classify the relationship between a source tweet, that is the target for classification, and additional tweets in its propagation graph into a polarity of positive, negative, or neutral. We use this polarity to construct abstract argumentation frameworks containing arguments representing each tweet and attacks based on different combinations of the following schemes for mapping stance into attack relationships:

- Adding attack relationships from any node that has a negative stance towards the source node to the source node itself
- Adding attack relationships from the source node to any node that has negative stance towards it

- Adding attack relationships between nodes that have differing stance to the source node. So if tweet A is positive toward the source node and tweet B is negative, we would add either unilateral or bilateral attack relationships, depending on our chosen scheme
- Limiting these attack relationships to only those in a tweet's sub-propagation graph relative to the source node
- Adding neutral nodes to the graph both as isolated components, only linked to themselves and with attack relationships towards the source node or node with a negative stance towards the source node, thereby reclassifying neutral nodes as positive or negative for the purposes of the argumentation graph

These schemes will be tested in different experimental setups to determine which perform better in different contexts. While some of these schemes may seem prima facie strange from a common sense point-of-view, they are designed to allow different ways for neighbourhood information to aggregate in the graph convolutional network, which may help in the classification task.

3.2. Linguistic Features

We generate linguistic features by creating embeddings for all tweets in our dataset using a variety of language models. First and foremost, we use the large-scale transformer based language models such as BERT. But we will also include simpler models such as Word2Vec for comparison. We generate these using different embedding sizes for comparison. These embeddings are added as node features in the GCN model along with the argumentative features. We will, however, also train a baseline classifier using only linguistic information for the sake of comparison.

We will try different feature settings with our experimental setup to determine, which are more successful in predicting misinformative tweets.

3.3. Argumentative Features

Argumentation specific features are generated by incorporating the acceptability status both for sceptical and credulous acceptance under the Complete, Preferred, Stable, Semi-Stable, and Stage semantics. The features will be pre-calculated using an abstract argumentation solver and added as node features by encoding acceptable as 1 and unacceptable as 0.

These features will be tested in different experimental setups to determine whether the addition of acceptability information improves overall performance.

3.4. GCN Architecture

We follow the GCN Architecture outlined in Malmqvist et al. [13] and adapt it to incorporate the additional feature information from the language model and the argumentation solver.

The architecture includes the following elements:

1. Pre-computed linguistic and argumentative features along with the normalised adjacency matrix for the argumentation framework

2. An input layer receiving these inputs
3. 6 repeating blocks of a GCN layer [8] and a Dropout layer [15]
4. Residual connections feeding the original features and the normalised adjacency matrix as additional input at each block
5. An layer that aggregates the embeddings generated by the GCN layers for graph level classification. We will experiment with different aggregation functions during the final phase of the research
6. A sigmoid layer that represents the probability that the source tweet is true

We treat the problem of veracity as a binary classification problem at the level of the graph. That means we aggregate information from all nodes and the embeddings generated by the GCN in order to judge whether the source tweet is true or not.

4. Planned Experiments

We will test the model on two datasets that contain both veracity and stance information and are commonly used in the research literature: Pheme [16] and RumourEval [6]. Both datasets contain tweets relating to controversial current events.

We will construct a baseline model using only linguistic features and one using only the argumentation graph and argumentative features. Then we will generate different model variants using various combinations of language models and argumentation graph construction schemes some including and some excluding argumentative features.

We will compare the results of these models based on their performance on the benchmark datasets to establish what combinations yield the best results for detecting misinformation. Finally, we will present our results in the context of other work using the same datasets.

5. Discussion

This paper presents the work-in-progress for our experiments in applying abstract argumentation in conjunction with linguistic features for detecting misinformation. We hope that this will extend the applicability of argumentation based methods for misinformation detection as well as give greater understanding of how argumentation can be used to enrich and improve deep learning architectures. This will also help decrease the gap between the formal world of abstract argumentation and natural language expressions of argument by incorporating both types of information in the same deep learning model. So far early work has demonstrated the technical feasibility of this approach and we aim to be able to present concrete results in September 2021.

References

[1] Jennifer L Bonnet and Senta Sellers. The COVID-19 Misinformation Challenge: An Asynchronous Approach to Information Literacy. *Internet Reference Services Quarterly*, 24(1-2):1–8, 2019.

[2] Oana Cocarascu and Francesca Toni. Combining Deep Learning and Argumentative Reasoning for the Analysis of Social Media Textual Content Using Small Data Sets. *Comput. Linguist.*, 44(4):833–858, dec 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-Person Games. *Artificial Intelligence*, pages 321–357, 1995.

[5] Luciano Floridi and Massimo Chiriatti. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[6] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, 2019.

[7] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1):82, 2020.

[8] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, sep 2019.

[9] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020.

[10] Sebastian Kula, Michał Choraś, and Rafał Kozik. Application of the BERT-Based Architecture in Fake News Detection BT. In Álvaro Herrero, Carlos Cambra, Daniel Urda, Javier Sedano, Héctor Quintián, and Emilio Corchado, editors, *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, pages 239–249, Cham, 2021. Springer International Publishing.

[11] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor Detection on {T}witter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia, jul 2018. Association for Computational Linguistics.

[12] Lars Malmqvist. Approximate Solutions to Argumentation Frameworks with Graph Neural Networks. *Online Handbook of Argumentation for AI*, page 32, 2021.

[13] Lars Malmqvist, Tommy Yuan, Peter Nightingale, and Suresh Manandhar. Determining the acceptability of abstract arguments with graph convolutional networks. *CEUR Workshop Proceedings*, 2672:47–56, 2020.

[14] Baroni Pietro, Martin Caminada, and Giacomin Massimilliano. An Introduction to Argumentation Semantics. *The Knowledge Engineering Review*, 00(January):1–24, 2004.

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[16] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLOS ONE*, 11(3):1–29, 2016.