Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-024-04841-8

ORIGINAL RESEARCH

Analyzing the impact of conversation structure on predicting persuasive comments online

Nicola Capuano[1] · Marco Meyer[2] · Francesco David Nota[3]

Received: 17 February 2024 / Accepted: 9 August 2024

Abstract
The topic of persuasion in online conversations has social, political and security implications; as a consequence, the problem of predicting persuasive comments in online discussions is receiving increasing attention in the literature. Following recent advancements in graph neural networks, we analyze the impact of conversation structure in predicting persuasive comments in online discussions. We evaluate the performance of artificial intelligence models receiving as input graphs constructed on the top of online conversations sourced from the "Change My View" Reddit channel. We experiment with different graph architectures and compare the performance on graph neural networks, as structure-based models, and dense neural networks as baseline models. Experiments are conducted on two tasks: (1) persuasive comment detection, aiming to predict which comments are persuasive, and (2) influence prediction, aiming to predict which users are persuasive. The experimental results show that the role of the conversation structure in predicting persuasiveness is strongly dependent on its graph representation given as input to the graph neural network. In particular, a graph structure linking only comments belonging to the same speaker in the conversation achieves the best performance in both tasks. This structure outperforms both the baseline model, which does not consider any structural information, and structures linking different speakers' comments with each other. Specifically, the F1 score of the best performing model is 0.58, which represents an improvement of 5.45% over the baseline model (F1 score of 0.55) and 7.41% over the model linking different speakers' comments (F1 score of 0.54).

Keywords Social media persuasion · Persuasive comment detection · Influence prediction · Text data

1 Introduction

The widespread adoption of social media platforms has facilitated a global forum for individuals to participate in discussions and express their viewpoints. Recently, academic interest has grown in understanding what makes individuals and their contributions persuasive in social networks (Prabhakaran and Rambow 2013; Rosenthal and Mckeown 2017; Diehl et al. 2016; Gil de Zuniga et al. 2018). A particularly intriguing aspect involves analyzing which comments lead other users to alter their views (Ta et al. 2022). In order to train artificial intelligence (AI) models and predict persuasiveness through supervised learning, datasets must provide a measure of it. However, although numerous datasets include social media conversations, they usually lack a metric for measuring persuasiveness, making research in the domain of persuasion detection challenging.

An exception is the datasets sourced from the Reddit channel /r/ChangeMyView (CMV). In this channel, a user writing a post, called "Original Poster" (OP), initiates a discussion on

Marco Meyer and Francesco David Nota have contributed equally to this work.

✉ Nicola Capuano
ncapuano@unisa.it

Marco Meyer
marco.meyer@uni-hamburg.de

Francesco David Nota
notafd.dottorando@casd.difesa.it

[1] Department of Information Engineering, Electrical Engineering, and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Italy

[2] Faculty of Humanities, University of Hamburg, Mittelweg 177, 20148 Hamburg, Germany

[3] Center for Higher Defence Studies, Defence Analysis and Research Institute, Piazza della Rovere, 83, 00165 Rome, Italy

Published online: 02 September 2024

a topic, generally by providing a personal opinion on it, and challenges other users to present arguments aimed at changing his/her view. The OPs can reward users that change their view with virtual awards called "Deltas". CMV discussions can be on any topic and are overseen by moderators to ensure the quality of discourse, fostering an environment where users openly express their views. Researchers have studied persuasiveness by developing AI models that predict the persuasiveness of comments based on CMV datasets.

Detecting persuasiveness in conversations, is of significant interest in understanding the influence of online content and could be used in several applications; for example, persuasiveness can be considered a measure of impact of disinformation (Zhang et al. 2013; Zerback et al. 2021; Hidey and McKeown 2018). Previous research has proposed AI models based on content features, to evaluate the overall influence of a speaker's comments in a conversation (Hidey and McKeown 2018), and the role of personal characteristics to predict persuasiveness (Al Khatib et al 2020). By contrast, the use of the structure of conversations as a feature of the models tested in the literature has received little attention. We close this gap by investigating the influence of the structure of conversations on persuasiveness.

We leverage a public dataset derived from Change My View (CMV) discussions to investigate the detection of persuasive comments by applying state-of-the-art graph neural networks and comparing them with baseline models and previously obtained results in the literature. Specifically, using graphs as data structure, we explore different structural representations of conversations that account for the relationships among comments to predict their persuasiveness. Our findings indicate that a self-speaker representation of the conversation structure improves models predicting persuasiveness in conversations. This graph structure links, for each comment in the conversation, the five previous comments from the same speaker to represent a speaker's contribution to the discussion.

The remainder of this paper is organized as follows: in Sect. 2 we review related research; in section 3 the dataset and the tasks involved are described; in Sect. 4 the methodology is discussed; in Sect. 5 results are compared on the tasks of persuasive comment detection and influence prediction; in Sect. 6 we discuss findings, list challenges and opportunities of our study; finally, Sect. 7 draws conclusions. The code of our experiments is available on GitHub (Fdnphd 2023).

2 Related literature

In the last decade, several studies have analyzed the dynamics of persuasion in social networks and their consequences. Diehl et al. (2016) found that news use leads to political persuasion. More interestingly, social interactive uses of social media also lead to political persuasion. Gil de Zuniga et al. (2018) found a direct relationship between political discussion disagreement and political persuasion in social media contexts and that civil reasoning also plays a moderating role in the process of political persuasion on social media. Prabhakaran and Rambow (2013) studied power relations in online written communication. Rosenthal and Mckeown (2017) created several system components that have been used to successfully detect influence in multiple online genres.

A more specific line of inquiry involves the analysis of Change My View (CMV) threads, aiming to uncover patterns associated with the successful persuasion of users. Some studies leveraged CMV datasets to identify patterns associated with a higher likelihood of persuasion, often measured by the acquisition of 'Delta' rewards. Guo et al. (2020) investigated the difference between speakers who have been awarded a Delta and those who have not. Xiao and Mensah (2022) found that the perceived persuasiveness of a comment varies systematically from the comments in the top thread level to the most nested level. Papakonstantinou and Horne (2023) found that, on average, top persuaders were more likely to provide external evidence for their claims, use morality-based reasoning, make longer comments, engage in more back-and-forth argumentation, and were less likely to use informal language in their arguments.

Similarly, Wiegmann et al. (2022) focused on analyzing the debaters' persuasion strategies, seeking to uncover the behavior, language style, and argumentative techniques that distinguish good from poor debaters in Change My View. They found that the effectiveness of persuasion improves over time for average debaters; more than two replies and comments longer than 400 characters correspond to higher chances of getting a delta. Other works studied the characteristics of persuasive comments and speakers; for example Egawa et al. (2019) analyzed persuasiveness with elementary argumentative units (EUs) in a token-level five-class scheme: testimony, fact, value, policy, and rhetorical statement. The authors proposed a Bi-LSTM-based sequence classifier for EU-labeling. They concluded that EUs indicate persuasiveness if used effectively. They found that 'fact' is the most persuasive EU.

Some research attempted to predict persuasiveness by employing AI models to identify persuasive users or strategies (Hidey and McKeown 2018; Wei et al. 2016; Tan et al. 2016; Khazaei et al. 2017; Shmueli-Scheuer et al. 2019; Jo et al. 2018). These studies also considered the relationships among comments in a conversation and the likelihood of a comment of receiving a delta. Several papers have highlighted the necessity of advanced techniques for representing conversation structures, interplay, and long-distance relationships of comments to predict persuasiveness (Guo et al. 2020; Wiegmann et al. 2022; Petruzzellis et al. 2023). In this work we face the tasks described in Tan et al. (2016) and Hidey and McKeown

(2018), as well as analysed features commonly adopted in the literature such as interplay, length of comments (Wiegmann et al. 2022), linguistic features (Khazaei et al. 2017), reputation of users (represented as the number of deltas ever received by a user Guo et al. (2020)) and structural features (Xiao and Mensah 2022). Notably, none of these studies have thoroughly explored the role of conversation structure in persuasiveness prediction.

Following advancements in conversation representations using graph neural networks (Ghosal et al. 2019), we evaluate the role of graph conversation structure applied to state-of-the-art GNN models and features. The aim is to understand if one or more graph representations of the conversation can improve the performance in predicting persuasiveness. This would demonstrate the importance of incorporating information about the conversation structure in persuasive prediction models. We perform experiments on social media conversations extrapolated from Change My View channel across two distinct tasks: persuasive comment detection and influence prediction, which we elaborate on in the following section.

3 Dataset and tasks

3.1 Dataset description

The dataset employed in this study is sourced from the Reddit "Change My View" (CMV) channel, publicly available as described by Tan et al. (2016). It is formed by a collection of conversations, each comprising the initial comment, known as the Original Poster (OP) comment, and the subsequent comments from challengers who endeavor to change the OP's viewpoint. During a conversation, the OP can assign a virtual reward called "Delta" (D) to the comment they find most persuasive (Fig. 1).

It is noteworthy that this dataset is self-labeled, with users autonomously assigning Deltas to user's comments they deem persuasive, eliminating the need for human annotators. The conversations in this dataset were accumulated over a period spanning from January 1, 2013, to September 1, 2015, resulting in a total of 3051 discussions. The resulting dataset was randomly shuffled and divided into train, validation, and test sets, accounting for 70, 20, and 10% of the total conversations.

In Table 1 we list the resulting number of comments of each set; note that some conversations have multiple comments with Deltas; furthermore, the absence of an explicitly awarded comment in a conversation does not indicate the lack of a persuasive comment in the conversation; therefore, a comment not being awarded with a Delta may be persuasive for some participants which did not award it explicitly in the conversation (Hidey and McKeown 2018).

3.2 Task in focus

In this work we conduct experiments on two tasks: (1) persuasive comment detection, aiming to predict which comments are persuasive, and (2) influence prediction, aiming to predict which users are persuasive. Influence prediction, introduced by Hidey and McKeown (2018), involves a dataset where each data point consists of the original poster comment and an attempted persuasive response, where responses consist of one or more sequential comments from the same challenger. Persuasive comment detection centers on the prediction of the most persuasive comments (those with a higher probability of receiving a Delta). Therefore, persuasive comment detection aims to identify specific comments that are more likely to persuade a user; while, Influence Prediction tries to predict persuasive users in the context of a conversation, where a persuasive user is defined as a user that must have at least one comment awarded with a Delta.

4 Methodology

We capitalize on recent advancements in representing conversations using graph neural networks (GNNs) (Ghosal et al. 2019). Our primary objective is to scrutinize whether a conversation's structural characteristics have an impact on the performance of two tasks: persuasive comment detection and influence prediction. We also endeavor to determine which conversation structure yields optimal results. To achieve these goals, we systematically construct graphs for each conversation, predict comments that have the highest likelihood of receiving a Delta (i.e. perform node classification), and compare our findings against baseline models and prior results achieved in the literature.

4.1 Conversation structure representation

Our experiments consider various graph structures to ascertain the most effective representation of conversations for influence prediction and persuasive comment detection. The graph structures we create draws inspiration from Ghosal et al. (2019), where a comment C by speaker S is linked to the previous n inter-speaker comments, i.e. comments from a different speaker, and the previous m self-speaker comments, i.e. comments authored by the same speaker. Notably, we consider "previous comments" of comment C as those occurring temporally before it, regardless of the sub-thread in which C resides within the conversation.

The graph's edge creation process involves linking each comment (node) with up to n inter-speaker comments, up to m self-speaker comments, itself (through a self-edge), and the OP comment. The link with the OP comment is based on

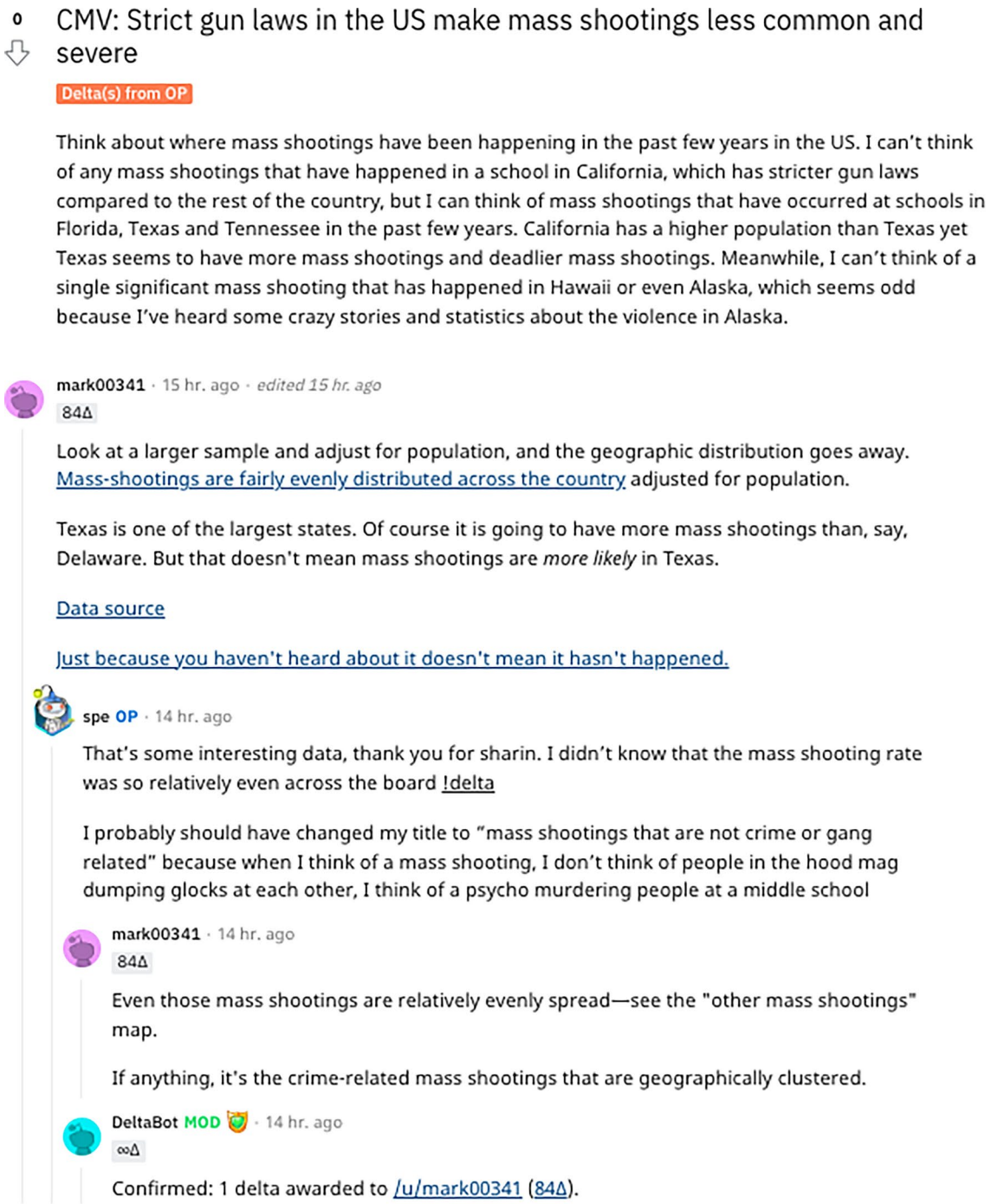

Fig. 1 Extract of conversation from CMV with the OP post and a Delta awarded comment

Table 1 Cardinality of train, validation, and test set

Identifier	Not awarded[a]	Awarded[a]
Train	174,689	3714
Validation	48,909	1160
Test	24,352	585

[a]Comments awarded with Delta are considered persuasive, while the remaining ones are not persuasive

the assumption of an influential relationship between the initial conversation comment and subsequent comments. Each comment C_i, linked to n inter-speaker and m self-speaker ones, is a representation R_j of a subpart of the conversation contextually relevant to C_i. Given all the comments of a conversation, the resulting conversation graph is the set of all their representations.

In Fig. 2, starting from a simplified example of a conversation, we show the visible conversation structure with the main thread and a sub-thread; below it, we show the link types needed to construct the graph structure which will be input of the GNNs; each node is identified with a speaker S_i a comment number c_j and a timestamp t_k; For example, $S_2c_2t_3$ is the second comment of speaker S_2 at time t_3, where: $t_m > t_n, m > n$

The bottom part of the figure shows how each node is linked with the others for inter-speaker (n) and self-speaker (m) both equal to 2.

To further clarify each link type, let us consider a conversation with Alice as the writer of the OP and two speakers, Bob and Carla. Self-links are links starting from a comment and directed to itself, such as a link starting and ending at Bob's first comment. Links with OP are links starting from a comment of a speaker and ending at the OP, for instance, a link starting

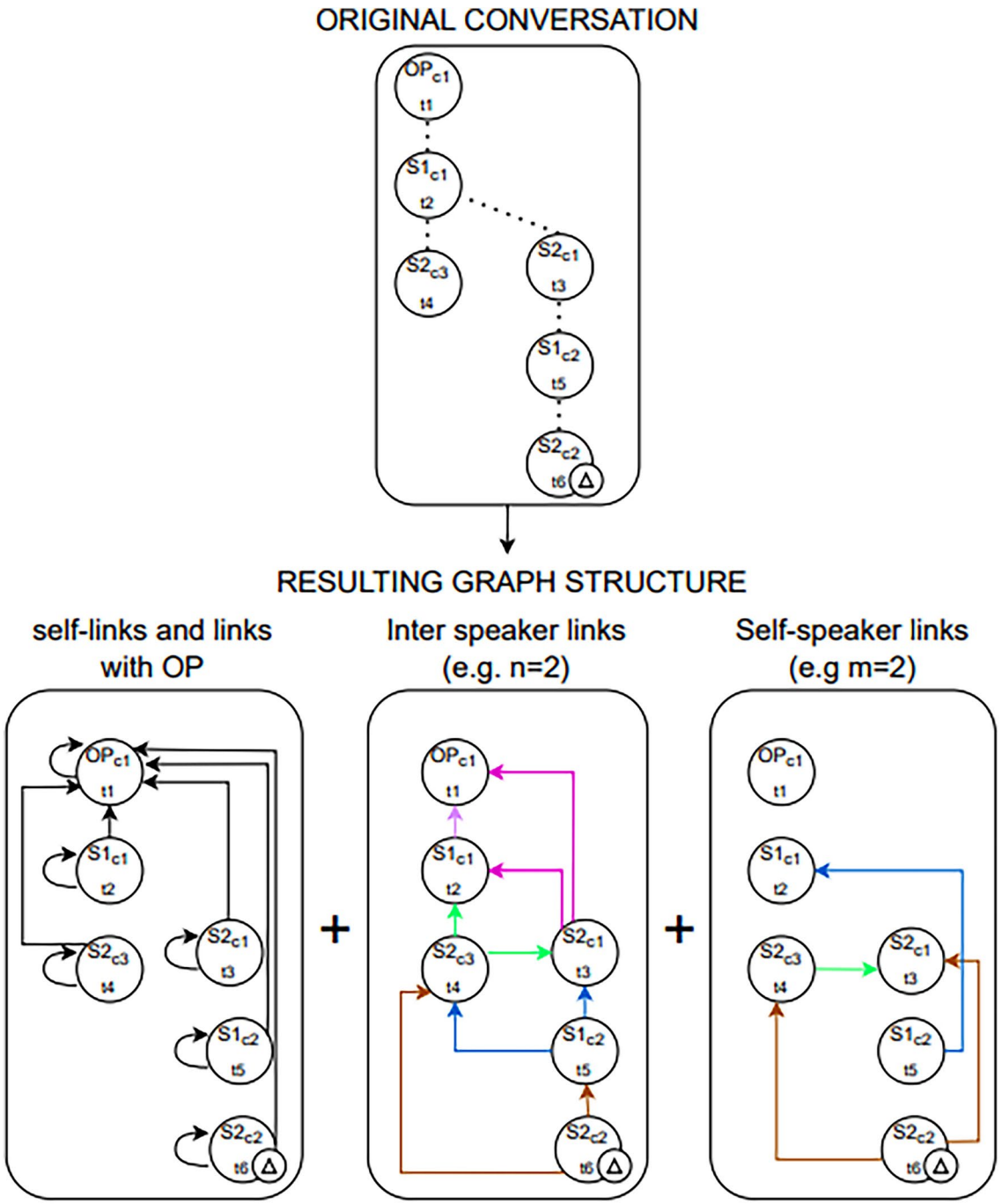

Fig. 2 Graph link types

from Carla's first comment and ending at Alice's comment representing the Original Post. Inter-speaker links start from a comment of one speaker and end at a comment of a different speaker, for example, a link starting from Carla's comment and ending at Bob's comment. Finally, self-speaker links start from a comment of a speaker and end at another comment by the same speaker, such as a link starting from Bob's second comment and ending at his first comment.

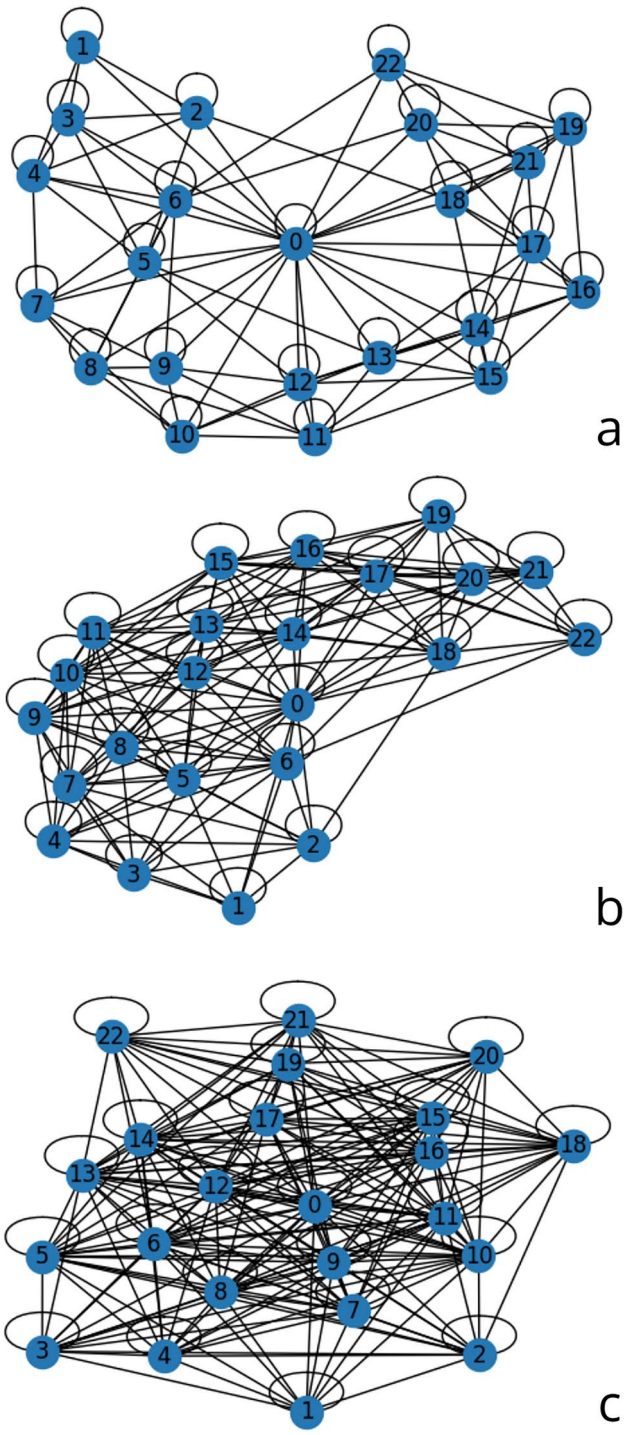

Fig. 3 Graphs representation of a real conversation **a** $\mathrm{n} = 3$, $\mathrm{m} = 2$; **b** $\mathrm{n} = 6$, $\mathrm{m} = 3$; **c** $\mathrm{n} = 10, \mathrm{m} = 5$

Inter- and self-speaker links are added to each comment in the conversation and can vary in number. For example, if there are three self-speaker links (m = 3) and Bob has four comments ordered chronologically as the 7th, 6th, 4th, and 2nd, then considering only the 7th comment, it will be connected to the 6th, 4th, and 2nd comments. Similarly, if there are two inter-speaker links (n = 2) and a conversation has nine comments, considering Carla's comment as the 7th in chronological order, the 7th comment would be connected to the 6th and 5th comments, regardless of the speaker.

In this work evaluate different graph structures created by varying n and m for inter- and self-speaker dependencies across several state-of-the-art GNN models, including graph sage (Hamilton et al. 2017), graph convolutional networks (GCN) (Kipf and Welling 2016), and Graph Attention Networks (Veličković et al. 2017). In Fig. 3, the graph structures of a real conversation composed of self-edges, inter-speaker, self-speaker, and OP links are depicted. In particular three different variants of the same conversation consisting of 22 comments are shown. The first with $\mathrm{n} = 3$ and $\mathrm{m} = 2$, the second with $\mathrm{n} = 6$ and $\mathrm{m} = 3$ and the third with $\mathrm{n} = 10$ and $\mathrm{m} = 5$ (where n and m are respectively the number of inter-speaker and self-speaker links).

4.2 Models overview

To comprehend the role of structure in predicting persuasiveness, we analyze the performance differences between state-of-the-art GNNs and baseline models. We consider two types of baseline models to evaluate the role of the conversation structure in predicting persuasiveness. The first encompasses a dense neural network which receives as input one single comment (i.e. this model ignores the structure of the conversation). The second is a GNN that receives as input a graph structure which is a simple linked list where each node is linked solely to the temporally previous one (i.e. representing only the temporal dimension of the structure of the conversation). In this section, we provide a brief overview of the models used in our experimental analyses.

GraphSAGE learns node embeddings from graph-structured data; it leverages a sampling-based approach to aggregate information from a node's local neighborhood efficiently (Hamilton et al. 2017). Graph attention network (GAT) is a GNN model that emphasizes attention mechanisms, allowing nodes to weigh the importance of their neighbors during the embedding process. This attention-based approach enables GAT to capture complex relationships and dependencies within the conversation graphs (Veličković et al. 2017).

Graph convolutional networks (GCN) is a GNN model that propagates information through graph convolutional layers. It relies on graph convolution operations to update node representations iteratively (Kipf and Welling 2016). Dense neural network (DNN) is a feedforward neural network with densely connected layers. DNNs do not consider the conversation structure and are used in this work as a baseline model. Random forest is an ensemble learning method based on decision tree classifiers. It combines the predictions from multiple decision trees to provide more robust and accurate results (Louppe 2014). We use random forest in conjunction with GraphSAGE for the task of influence prediction.

4.3 Features and text representation

Effective representation of comments features is crucial for GNN and baseline model processing. We incorporate a range of features, some common in natural language processing and others derived from prior research in persuasiveness and influence within social network conversations. A list of features and their definitions or references is presented in Table 2. For each feature, we evaluate its performance both individually and in combination with other features across different graph structures applied on GNNs and baseline models. This analysis helps elucidate the individual contributions of each feature and identifies the most effective ones.

4.4 Implementation details and evaluation metrics

In Figs. 4 and 5 we show respectively the steps needed to train and evaluate the GNNs and baseline models persuasive comment detection and influence detection. The chart depicted in Fig. 4 is divided into three parts; the first one is about choosing the features, which GNN and the number of inter-speaker and self-speaker links to be used; the second defines functions to preprocess data, construct the graph and define the models. Finally, the third part uses the defined functions to compute the graph dataset, train the models, and evaluate the performance. For what concerns influence detection, as seen in Fig. 5, the probabilities of the most performing model computed in persuasive comment detection are adopted as features together with the number of comments a speaker made in a conversation.

Note that the same user can have different labels if present in different conversations. For example, in conversation C_1 user U_1 can have a delta-awarded comment and therefore be labeled as persuasive, while the same user in conversation C_2 might not have any comment awarded with delta. Therefore, in this example, U_1 would be present in two rows in the dataset, one row representing the couple U_1-C_1 and labeled as persuasive and the other row representing the couple U_1-C_2 and labeled as not persuasive. Additional details on our implementation are related to the configuration of model and hyperparameters.

We adopted two-layered GNNs as they outperformed alternative configurations. When the dimensionality of the features exceeds 40, we set the hidden channel to a dimensionality 80; while when the input dimensionality is less than 40, the hidden channel dimensionality is set to 4. We adopt as activation function Leaky ReLU, while softmax is employed to calculate the probability for a comment of being awarded with a Delta. Adam serves as the optimizer with a learning rate of 0.01, and cross-entropy acts as the loss function. Additionally, we estimate class weights to address the dataset's significant class imbalance between Delta-awarded comments and non-awarded comments.

The evaluation metrics used for our comparative analysis are accuracy and F1 score both defined by the following formulas:

$$
\begin{aligned}
Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\
F1score &= \frac{2 \times (Precision \times Recall)}{(Precision \times Recall)} \\
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN}
\end{aligned}
$$

where TP is true positives, FP is false positives and FN is false negatives.

Table 2 Feature list

Feature name	Definition
Universal sentence encoder (USE)	A sentence embedding defined by Yang et al. Yang et al. (2020)
TF-IDF	The product of term frequency (TF) and inverse document frequency (IDF) (Sparck Jones 1972); where TF is the number of occurrences of a term in a document in the corpus and IDF is a score that measures how important a term is. Rarely occurring terms have a high IDF score
Emotions intensity (EI)	The intensity of emotions (anger, sadness, joy, and optimism) expressed in the comment, going from 0 (not intense) to 1 (strongly intense) (Barbieri et al. 2020)
Speaker's Delta (SD)	The number of Deltas a speaker has received overall during conversations in CMV. It is considered a metric of the credibility of a speaker (Wei et al. 2016)
Absolute position (AP)	The position of a comment in the conversation, e.g. 20 if the comment is the 20th temporally speaking after the OP comment
Words length (WL)	The number of words of a comment

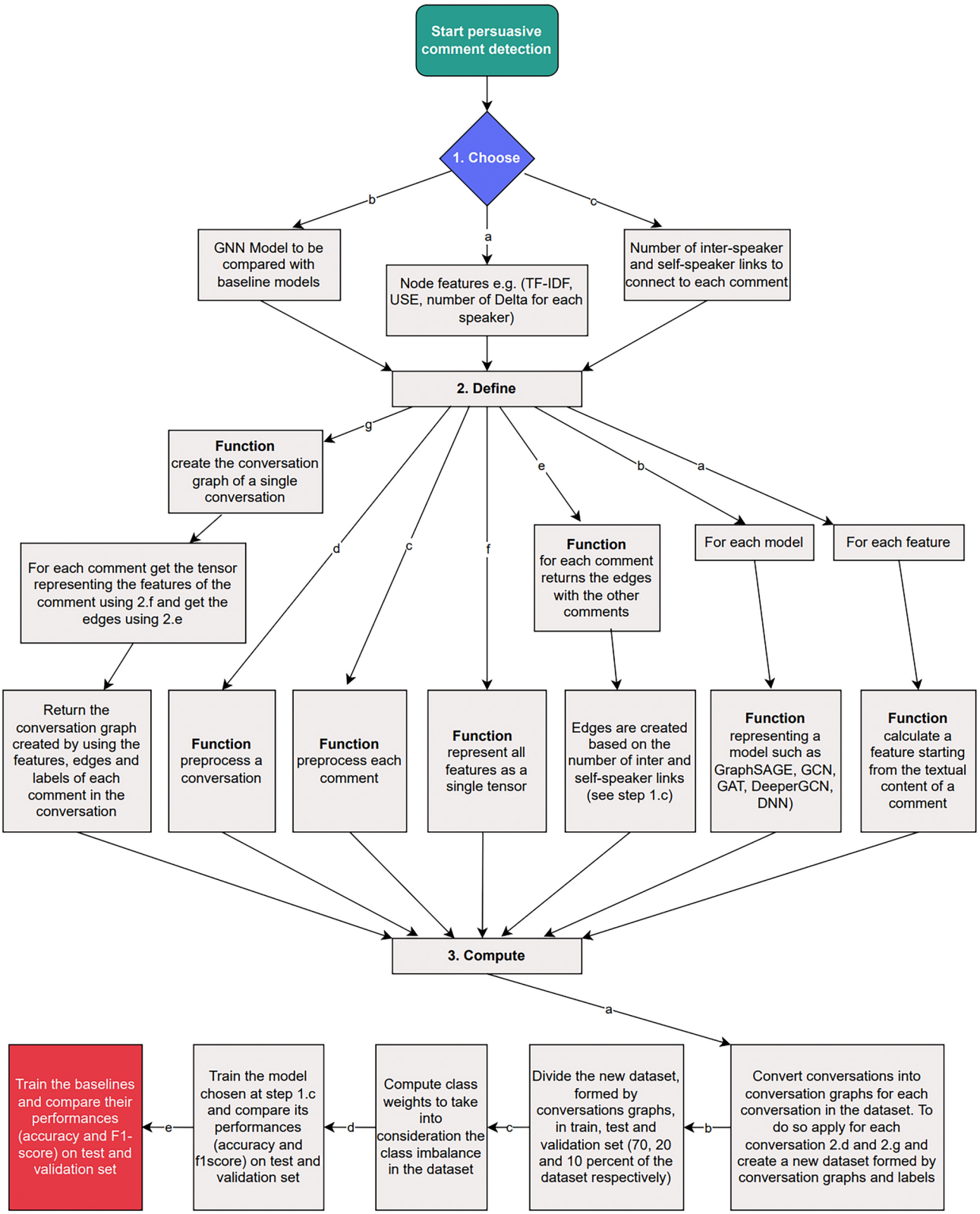

Fig. 4 Flowchart for persuasive comment detection

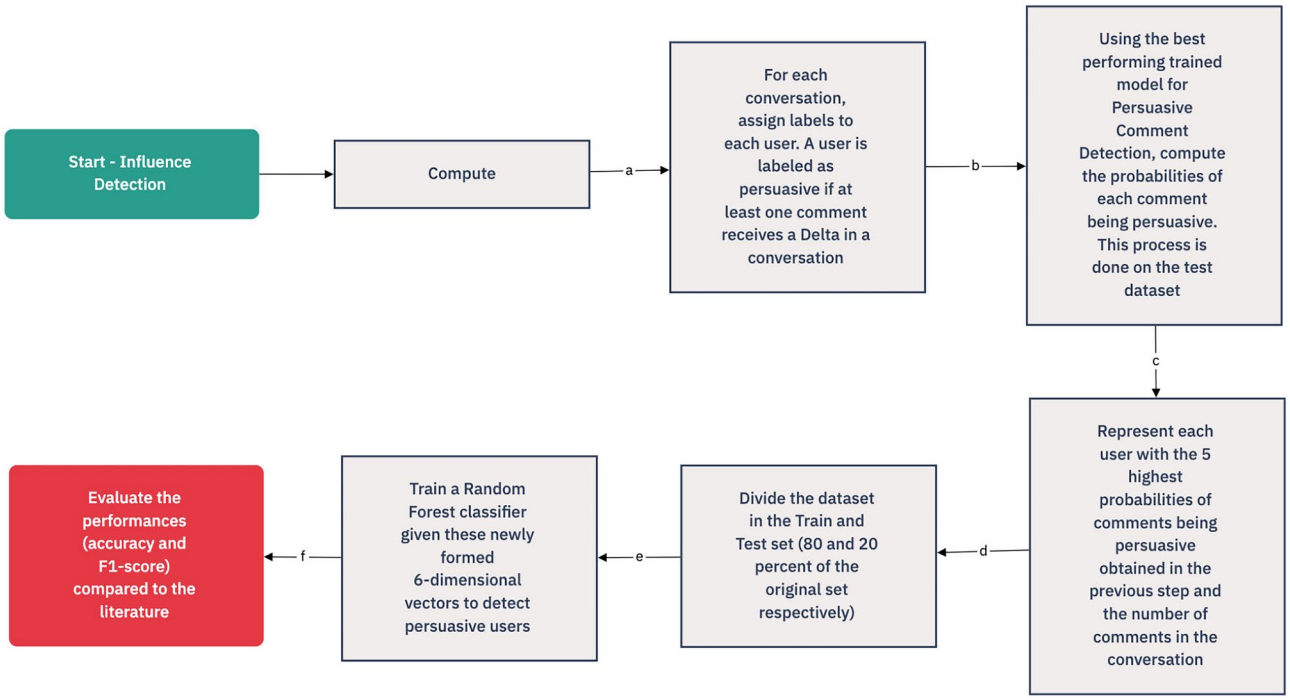

Fig. 5 Flowchart for influence detection

Table 3 Results of best-performing models

Model	n	m	Features	Accuracy	F1 score	Recall
Graph SAGE	**0**	**5**	**USE TF-IDF,**	**0.940**	**0.581**	**0.655**
Graph SAGE	0	3	USE, TF-IDF	0.972	0.558	**0.559**
Graph SAGE	8	3	USE, TF-IDF	0.946	0.540	**0.541**
Graph SAGE-baseline	1	0	USE, TF-IDF	0.976	0.521	**0.517**
Dense neural network-baseline	X	X	USE, TF-IDF	0.906	0.553	**0.585**

5 Experiments and results

In this section, we present the outcomes of our experiments on the tasks of persuasive comment detection and influence prediction. We compare GNN and baseline models on different sets of features for persuasive comment detection and benchmark our best results against prior literature for influence prediction. Furthermore, for GNNs we also compare several graph structures by varying the number of inter-speaker (n) and self-speaker (m) links for each comment.

5.1 Persuasive comment detection

We organize the results for persuasive comment detection into two tables, the overall best-performing results, and results for individual features. For all results, see appendixes A. Table 3 displays the top-performing experiments in persuasive comment detection. These models leverage the universal sentence encoder (USE) and term frequency-inverse document frequency (TF-IDF) features. Results are highlighted for the highest accuracy and macro averaged F1 score. Note the discrepancy between high accuracy and F1 score is due to the dataset imbalance formed mostly by non-persuasive comments and a minority of persuasive ones. We use n and m to indicate the number of inter-speaker and self-speaker comments linked to each node.

In Table 4 we list the performances of models trained and tested on individual features. All features other than USE were tested with a hidden channel of 4, compared to the USE where the hidden channel was 64. This is necessary to adapt the complexity of the model to the input dimensionality. For the sake of brevity, we refer G-SAGE as the graph sage model having for each node 8 previous inter-speaker links and 3 same-speaker ones; G-Line as the baseline graph sage model having each comment linked to the previous one, and, finally DNN as the dense neural network baseline.

5.2 Influence prediction

Intuitively, changing a person's view in a conversation might result from multiple interactions with one or more speakers. The idea of predicting which comment has been awarded a Delta follows the rules of the CMV channel. However, since

Table 4 Results on individual features

Feature	Model	Accuracy	F1 score	Recall
USE	G-SAGE	**0.971**	**0.531**	0.539
	G-LINE	0.978	0.517	0.516
	DNN	0.958	0.518	0.534
TF-IDF	G-SAGE	0.710	0.446	0.707
	G-LINE	0.751	0.459	0.708
	DNN	0.739	0.443	0.709
EI	G-SAGE	0.742	0.447	0.613
	G-LINE	0.776	0.458	0.595
	DNN	0.379	0.286	0.562
SD	G-SAGE	0.567	0.383	0.730
	G-LINE	0.821	0.469	0.593
	DNN	0.824	0.476	0.593
AP	G-SAGE	0.707	0.440	0.687
	G-LINE	0.503	0.352	0.690
	DNN	0.510	0.356	0.662
WL	G-SAGE	0.793	0.475	0.705
	G-LINE	0.791	0.474	0.702
	DNN	0.845	0.496	0.697

Table 5 Results on influence detection

Study	Accuracy	F1 score
Present study	0.950	0.640
Hidey and McKeown (2018)	0.810	0.607

Deltas are awarded to comments, not users, they do not identify the most persuasive user. The task of influence prediction aims at predicting the most persuasive users. To achieve coherent labeling, we labeled a user as persuasive if at least one comment of that user has been awarded a Delta; we labeled the user as not persuasive otherwise. This task has been explored in-depth by using the graph structure and model that performed best on persuasive comment detection (shown in bold in Table 3).

We predict the probability of receiving a delta for each comment in the conversation and we create a dataset where each user is represented as a vector formed by the overall count of comments in the conversation and the five probabilities of the most persuasive user comments. A Random Forest classifier, consisting of 100 estimators, is trained on this dataset, and the results, as shown in Table 5, are compared with the one achieved by Hidey and McKeown (2018).

It is important to note that, although a relationship exists between the GNN classifying individual persuasive comments and the random forest algorithm classifying persuasive users, this relationship is not always consistent. This is because random forest adopts as features the probability outputs generated by the GNN, rather than the classification of comments as persuasive or not. Consequently, a user with five comments in a conversation, none of which classified as persuasive by the GNN, may still be classified as a persuasive user by random forest. Conversely, a user classified as non-persuasive by the random forest may still have at least one comment deemed persuasive by the GNN.

5.3 Additional observations

In an effort to assist researchers in this emerging field, we summarize key findings of low-performing experiments: (a) keeping stopwords increased performance, in line with the findings of Tan et al. (2016); (b) GNNs considering edge attributes (such as SplineGNN (Fey et al. 2018) and DeeperGCN (Li et al. 2020) performed worse than the ones using only node features; on these GNNs we tested edge features that were inspired by features indicative of persuasiveness in the literature; (c) two layers of GNNs performed better than one or more than two; (d) linking future self and inter-speaker comments led to lower performances.

Note that these results are listed in detail in Appendix A. We could not compare our results with some similar works in the literature because of their dataset filtering decisions. Khazaei et al. (2017) have used a smaller and balanced dataset composed of 1720 persuasive comments and 1720 non-persuasive ones; Shmueli-Scheuer et al. (2019) did not include comments of sub-threads in the dataset independently if they received a Delta or not, and, Jo et al. (2018) have trained and tested their models only on the topics that have the highest Delta ratios. As GNNs required to include the structure of the conversation during model training, we had to experiment on the entire unbalanced dataset, including sub-threads; moreover, we decided to use all the dataset conversations independently from the topic, including low Delta ratio topics.

6 Discussion

In this section, we discuss our findings in a broader context, addressing the significance of our results, the challenges encountered in our study, and the opportunities they uncover.

6.1 Main findings

Our results underscore the significant variations in performance stemming from feature choices, model selection, and structural representations within graph neural networks (GNNs). Notably, the highest performance was attained by the graph SAGE model, leveraging a structure that links each comment of a speaker (S) with the previous five comments from the same speaker, incorporating self-edges and connecting each comment with the original post. This self-speaker structure surpasses baseline models and structures including inter-speaker links inspired by similar conversation graph representations in the literature such as DialogueGCN (Ghosal et al. 2019), emphasizing the importance of a unified representation of speaker comments in influencing a person's view.

We explored various features commonly associated with persuasiveness in the literature, yet our findings demonstrate

features traditionally used in natural language processing, such as universal sentence encoder (USE) and term frequency-inverse document frequency (TF-IDF), exhibit superior performance when used in combination, both for baseline and GNN models. Specifically, users' overall delta number, time difference between comments, length of the comment, text similarity, position of a comment in the conversation and emotion intensity (for anger, sadness, joy and optimism) did not influence positively the performance as visible in Appendix A. Thus, features representing reputation, interplay, complexity of argumentation and emotions did not improve significantly the performances of the model. Furthermore, more complex structures with a large number of links tend to perform worse; one plausible explanation is that excessive interconnectedness between nodes dilutes information as nodes mutually influence each other's content during GNN processing.

We found that linking comments distant in the conversation with each other, unless they are from the same speaker, did not increase the performance significantly, in fact in some cases it reduced it. Interestingly, as listed in Appendix A, GCN and GAT underperformed baseline models, implying that the specific implementation of the model plays a critical role in improving prediction results. The highest-performing persuasive comment detection model was subsequently employed for influence prediction, demonstrating a significant improvement compared to prior research. Certain experiments conducted in this study demonstrate a higher recall compared to what is considered the best performing model. However, in these instances, the accuracy and F1 score are significantly lower than those observed for self-speaker structures. Specifically, whenever the F1 score exceeds 0.5, the recall is below 0.55. Experiments exhibiting a recall greater than 0.7 also have an F1 score that is lower or significantly lower than 0.5, which generally indicates suboptimal performance, as an F1 score below 0.5 suggests a poor balance between precision and recall. These results suggests the importance of the self-speaker structure in detecting influential speakers within conversations.

6.2 Challenges and opportunities

The intricacies of analyzing persuasiveness in conversations give rise to several challenges while also offering opportunities for future research. In this study, we have explored various graph structures and used them in GNNs to predict persuasive comments and users. Nevertheless, the need for an automated search mechanism to identify the most effective graph structure persists. Our approach involved processing information from each conversation independently. However, considering relationships between different conversations may provide valuable insights into the dynamics of persuasiveness across diverse contexts.

It is important to acknowledge that some persuasive comments may not have been explicitly recognized with a Delta award in CMV by users. Consequently, a classifier may accurately identify a comment as persuasive, even if it was not explicitly acknowledged in the CMV context, hence the unsolved challenge of having to train models compromised by comments having characteristics of a persuasive comment, but not labeled explicitly as such. While our investigation serves as a potential indicator of persuasive behaviors in conversations, it is worth noting that user behaviors may vary when engaging in social networks other than the Reddit CMV channel. This observation highlights the need for cross-platform studies in the future. Therefore, integrating datasets that label persuasiveness in different contexts holds the promise of yielding more robust results in future research endeavors.

7 Conclusion

This work analyzes the influence of conversation structure on predicting persuasive comments in online discussions. Extensive experimentation is conducted using a publicly available dataset, complemented by features previously identified as predictive of persuasiveness in the literature, along with additional commonly used features in natural language processing. The experiments explore various ways of linking comments to represent a conversation as a graph to optimize performance of GNNs. Our results demonstrate that the best-performing features for predicting persuasiveness are TF-IDF and the universal sentence encoder embeddings and, most importantly, that only using the GraphSAGE model with a specific self-speaker graph structure consisting of 5 self-speaker links and 0 inter-speaker ones for each comment significantly outperforms baseline models for the task of persuasive comment detection and outperformed previous results in the literature for the task of influence prediction.

These results indicate that when considering a graph conversation structure the change in performances obtained by linking comments of different speakers or comments distant from each other in the conversation is negligible, while the influence of comments written by the same speaker on each other shows a significant increase in predictive power, thus highlighting the importance of including self-speaker representations of user comments to predict their persuasiveness in online conversations. To the best of our knowledge, this study represents the first attempt to examine in detail the influence of conversation structure on predicting comments' persuasiveness; the findings shed light on the importance of incorporating self-speaker structural representations for predicting persuasiveness in online conversations.

Table 6 JSON of all results with comments including stopwords

Model	Past_n	Past_m	Future_n	Future_m	Edge_Features	Node_Features	Accuracy	F1 score
GSAGE	20	10	0	0	X	embeddings, tf_idf	0.957	0.521
GSAGE	10	4	0	0	X	embeddings, tf_idf	0.974	0.527
GSAGE	1	0	0	0	X	embeddings, tf_idf	0.977	0.522
GSAGE	8	3	0	0	X	embeddings, tf_idf	0.966	0.534
GSAGE	8	3	0	0	X	embeddings, tf_idf, emotions, speaker_delta, abs_position, word_len	0.791	0.481
GSAGE	8	3	0	0	X	embeddings	0.971	0.531
GSAGE	1	0	0	0	X	embeddings	0.978	0.517
GSAGE	1	0	0	0	X	embeddings, tf_idf, emotions	0.978	0.521
GSAGE	1	0	0	0	X	tf_idf	0.746	0.457
GSAGE	8	3	0	0	X	emotions	0.742	0.448
GSAGE	1	0	0	0	X	emotions	0.777	0.458
GSAGE	8	3	0	0	X	speaker_delta	0.567	0.384
GSAGE	1	0	0	0	X	speaker_delta	0.824	0.476
GSAGE	8	3	0	0	X	abs_position	0.631	0.410
GSAGE	1	0	0	0	X	abs_position	0.672	0.426
GSAGE	1	0	0	0	X	abs_position	0.672	0.426
GSAGE	8	3	0	0	X	abs_position	0.707	0.441
GSAGE	8	3	0	0	X	word_len	0.794	0.475
GSAGE	1	0	0	0	X	word_len	0.792	0.474
GSAGE	1	0	0	0	X	abs_position, word_len	0.881	0.520
GSAGE	8	3	0	0	X	abs_position, word_len	0.858	0.508
GSAGE	8	3	0	0	X	embeddings, tf_idf, emotions	0.966	0.530
GSAGE	8	3	0	0	X	tf_idf	0.710	0.446
GSAGE	0	3	0	0	X	embeddings, tf_idf	0.973	0.559
GSAGE	0	5	0	0	X	embeddings, tf_idf	0.940	0.581
GCN	10	4	0	0	X	embeddings, tf_idf	0.757	0.443
GAT	10	4	0	0	X	embeddings, tf_idf	0.892	0.495
Baseline	X	X	X	X	X	embeddings, tf_idf, emotions	0.957	0.513
Baseline	X	X	X	X	X	embeddings, tf_idf, emotions, speaker_delta, abs_position, word_len	0.813	0.492
Baseline	X	X	X	X	X	embeddings	0.959	0.519
Baseline	X	X	X	X	X	tf_idf	0.752	0.460
Baseline	X	X	X	X	X	emotions	0.379	0.286
Baseline	X	X	X	X	X	speaker_delta	0.824	0.476
Baseline	X	X	X	X	X	abs_position	0.511	0.357
Baseline	X	X	X	X	X	abs_position	0.504	0.353
Baseline	X	X	X	X	X	abs_position	0.504	0.353
Baseline	X	X	X	X	X	word_len	0.811	0.482
Baseline	X	X	X	X	X	word_len	0.846	0.497
Baseline	X	X	X	X	X	abs_position, word_len	0.784	0.477
Baseline	X	X	X	X	X	embeddings, tf_idf	0.906	0.554
DeeperGCN	8	3	0	0	time_difference, text_similarity, delta_ratio, word_len_ratio	embeddings, tf_idf, emotions, speaker_delta, abs_position, word_len	0.780	0.480
DeeperGCN	8	3	0	0	time_difference, text_similarity, delta_ratio, word_len_ratio	embeddings, tf_idf, emotions	0.975	0.504
SplineNN	8	3	0	0	time_difference, text_similarity, delta_ratio, word_len_ratio	embeddings, tf_idf, emotions, speaker_delta, abs_position, word_len	0.824	0.493

Table 7 JSON of all results with comments not including stopwords

Model	Past_n	Past_m	Future_n	Future_m	Edge_Features	Node_Features	Accuracy	F1 score
GSAGE	1	0	0	0	X	embeddings	0.975	0.520
GSAGE	4	4	4	4	X	embeddings	0.969	0.533
GSAGE	8	4	0	0	X	embeddings	0.970	0.544
GSAGE	6	3	0	0	X	embeddings	0.971	0.526
GSAGE	8	3	0	0	X	embeddings	0.965	0.538
GSAGE	8	3	0	0	X	speaker_delta, abs_position, word_len	0.829	0.485
GSAGE	8	3	0	0	X	embeddings, speaker_delta, abs_position, word_len	0.860	0.508
GSAGE	8	3	0	0	X	embeddings, tf_idf	0.975	0.535
GSAGE	8	3	0	0	X	embeddings, tf_idf, emotions	0.970	0.563
Baseline	X	X	X	X	X	embeddings	0.958	0.519
Baseline	X	X	X	X	X	speaker_delta, abs_position, word_len	0.829	0.491
Baseline	X	X	X	X	X	embeddings, speaker_delta, abs_position, word_len	0.731	0.451
SplineNN	8	3	0	0	X	embeddings	0.973	0.517
SplineNN	8	3	0	0	time_difference	embeddings	0.972	0.517
SplineNN	8	3	0	0	time_difference, text_similarity	embeddings	0.973	0.530
SplineNN	8	3	0	0	time_difference, text_similarity	empty	0.191	0.166
SplineNN	8	3	0	0	time_difference, text_similarity, delta_ratio, word_len_ratio	embeddings	0.970	0.515
SplineNN	1	0	0	0	time_difference, text_similarity	embeddings	0.976	0.524

Appendix A

In Tables 6 and 7 the results of the experiments are listed.

Acknowledgements This study is funded by the Center for Higher Defence Studies and the Volkswagen Foundation.

Funding This study is funded by the Center for Higher Defence Studies and the Volkswagen Foundation.
Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare no conflict of interest. The funding agency had no role in the design of the study; in the collection, analyses, or interpretation of data.

References

Al Khatib K, Völske M, Syed S et al (2020) Exploiting personal characteristics of debaters for predicting persuasiveness. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 7067–7072

Barbieri F, Camacho-Collados J, Neves L et al (2020) Tweeteval: unified benchmark and comparative evaluation for tweet classification. arXiv:2010.12421

Diehl T, Weeks BE, Gil de Zúñiga H (2016) Political persuasion on social media: tracing direct and indirect effects of news use and social interaction. New Media Soc 18(9):1875–1895

Egawa R, Morio G, Fujita K (2019) Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In: Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop, pp 422–428

Fdnphd (2023) Github—fdnphd/cmv-structures-role. https://github.com/fdnphd/cmv-structures-role

Fey M, Lenssen JE, Weichert F et al (2018) Splinecnn: fast geometric deep learning with continuous b-spline kernels. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 869–877

Ghosal D, Majumder N, Poria S et al (2019) Dialoguegcn: a graph convolutional neural network for emotion recognition in conversation. arXiv:1908.11540

Gil de Zuniga H, Barnidge M, Diehl T (2018) Political persuasion on social media: a moderated moderation model of political discussion disagreement and civil reasoning. Inf Soc 34(5):302–315

Guo Z, Zhang Z, Singh M (2020) In opinion holders' shoes: modeling cumulative influence for view change in online argumentation. Proc Web Conf 2020:2388–2399

Hamilton W, Ying Z, Leskovec J (2017) Inductive representation learning on large graphs. In: Advances in neural information processing systems, vol 30

Hidey C, McKeown K (2018) Persuasive influence detection: the role of argument sequencing. In: Proceedings of the AAAI conference on artificial intelligence

Jo Y, Poddar S, Jeon B et al (2018) Attentive interaction model: modeling changes in view in argumentation. arXiv:1804.00065

Khazaei T, Xiao L, Mercer R (2017) Writing to persuade: analysis and detection of persuasive discourse. In: IConference 2017 proceedings

Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907

Li G, Xiong C, Thabet A et al (2020) Deepergcn: all you need to train deeper gcns. arXiv:2006.07739

Louppe G (2014) Understanding random forests: from theory to practice. arXiv:1407.7502

Papakonstantinou T, Horne Z (2023) Characteristics of persuasive deltaboard members on reddit'sr/changemyview

Petruzzellis F, Bonchi F, Morales GDF et al (2023) On the relation between opinion change and information consumption on reddit. In: Proceedings of the international AAAI conference on web and social media, pp 710–719

Prabhakaran V, Rambow O (2013) Written dialog and social power: manifestations of different types of power in dialog behavior. In: Proceedings of the sixth international joint conference on natural language processing, pp 216–224

Rosenthal S, Mckeown K (2017) Detecting influencers in multiple online genres. ACM Trans Internet Technol (TOIT) 17(2):1–22

Shmueli-Scheuer M, Herzig J, Konopnicki D et al (2019) Detecting persuasive arguments based on author-reader personality traits and their interaction. In: Proceedings of the 27th ACM conference on user modeling, adaptation and personalization, pp 211–215

Sparck Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc 28(1):11–21

Ta VP, Boyd RL, Seraj S et al (2022) An inclusive, real-world investigation of persuasion in language and verbal behavior. J Comput Soc Sci 5(1):883–903

Tan C, Niculae V, Danescu-Niculescu-Mizil C et al (2016) Winning arguments: interaction dynamics and persuasion strategies in good-faith online discussions. In: Proceedings of the 25th international conference on world wide web, pp 613–624

Veličković P, Cucurull G, Casanova A et al (2017) Graph attention networks. arXiv:1710.10903

Wei Z, Liu Y, Li Y (2016) Is this post persuasive? Ranking argumentative comments in online forum. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers), pp 195–200

Wiegmann M, Al Khatib K, Khanna V et al (2022) Analyzing persuasion strategies of debaters on social media. In: 29th international conference on computational linguistics, international committee on computational linguistics, pp 6897–6905

Xiao L, Mensah H (2022) How does the thread level of a comment affect its perceived persuasiveness? A reddit study. In: Science and information conference. Springer, pp 800–813

Yang Z, Yang Y, Cer D et al (2020) Universal sentence representation learning with conditional masked language model. arXiv:2012.14388

Zerback T, Töpfl F, Knöpfle M (2021) The disconcerting potential of online disinformation: persuasive effects of astroturfing comments and three strategies for inoculation against them. New Media Soc 23(5):1080–1098

Zhang J, Carpenter D, Ko M (2013) Online astroturfing: a theoretical perspective