A Methodology for Surveying Gradients of Influence on Social Media Platforms using Multi-Media Content

Luke Gassmann[a;*], **Matthew Edwards**[a] **and Ryan McConville**[a]

[a]School of Computer Science, University of Bristol, United Kingdom.
ORCiD ID: Luke Gassmann https://orcid.org/0000-0001-9247-944X,
Matthew Edwards https://orcid.org/0000-0001-8099-0646,
Ryan McConville https://orcid.org/0000-0002-7708-3110

Abstract.

Determining the importance of influence inside a network and its impact on the behaviour of its users can provide insight into historical trends, information dispersal, and reducing the spread of misinformation. Therefore, improving research into how users perceive and interact with others is valuable. Understanding how members of a group influence each other by sharing media or holding conversations can be especially vital. These interactions (often between two group members) can lead to members adopting certain behaviours, we consider this an example of influence. In this paper, we review existing work in detecting and defining influence in social networks and we propose a methodology for three experiments using content features and transformer architecture models to measure and evaluate different types of influence at various resolutions.

1 Introduction

Online social networks provide millions of connections between people every day. In many cases, these platforms provide a place where people can voice their opinions on public matters and reach a much wider audience. These interactions take place using many different graphical, text, or platform-specific mediums, each of which provides a unique method of communicating ideas between those connected inside the network. Interactions often take place between two people, with the most common public forum interaction being a post and a reply. This type of interaction covers two types of social media users, where the influencer posts content and the potentially influenced reply. Understanding and measuring the extent of this influence, the type of content and its wider impact on the network would provide additional insight into the features of influencing content.

Many existing methods for determining influence use platform-specific features that are computationally light. Features such as 'friends', 'followers', 'hashtags' and 'likes' are architectural examples of influence indicators within groups [1], where users actively show positive engagement with an individual or group by building connections. This implies a positive correlation between the user's outgoing connections and their level of current and historic influence in the community. Although these methods are valuable, they are symptoms of past influence, local to the platform, require insight into a more extensive network's topology, and are vulnerable to phenomena like controversial influence [4] [21] [22] [13]. However, light-weight topological methods can still provide insight into individuals' roles in group influence using a ratio of proportional connections. For instance, outlier detection suggests that individuals act as the group's information source, where connections denote the ability to spread information directly and indirectly in the group (via transitive influence) [13] [5] which can often also be associated with particular account types [14]. Whilst these features are beneficial, they are limited in complexity and rely on an assumed association. More complex attention mechanism research (such as DeepInf and MRAInf [15] [19]) have provided reliable influence detection methods in networks by identifying the importance of topological connections in individuals' behaviour. However, these mechanisms require large amounts of platform-specific connection data [3] and fail to provide insight into the source of content-features that lead to a user being influenced. TAP and EIRank [20] [2] attempt to restructure topological graphs as they relate to conversational features such as topic, however, these are also limited in scope and are dependent on platform-specific content features, a trait shared by many content-based methods in addition to being limited in resolution and variations of influence [21][16].

Due to the constraints of existing influence detection research and the improvements in language feature extraction, surrounding research in influence detection could benefit from a multi-label non-topological framework. Therefore we propose a modular content-based approach, focusing on: transparency in conversational and media-related feature extraction, types of influence with a range of multi-label resolutions, and observing these principles in a larger network. Reflecting on these aims, each experiment's question can be summarised as:

- **Experiment 1:** Can resolutions of interpersonal influence be predicted using conversation content?
- **Experiment 2:** Can user engagement with an image's content be predicted based on community affiliations?
- **Experiment 3:** How do gradients of influence compare to other baselines in social network group mapping?

We define three metrics used to represent influence in our research. *Social influence* (SI): the degree of overlap in conversation partners' structural connections on a platform. *Behavioural influence* (BI): the likelihood of future signal-boosting behaviour between the two parties (e.g., re-sharing content). *Active engagement* (AE): the act of

* Corresponding Author. Email: luke.gassmann@bristol.ac.uk

engaging with a piece of content by choice.

The remainder of this paper is organised as follows: A discussion of three linked experiments relating to their wider contribution and experimentation, and a summary.

1.1 Investigating Lexical Influence

Since text is one of the most prevalent forms of communication on social networking platforms, we propose a framework that is intended to predict social and behavioural influence using dialogue between two users. This is based on the hypothesis that an exchange between two users elicits symptomatic changes in network structure (commonly misused as a source and validation of influence). Based on this issue, our first experiment proposes an interpersonal influence framework using the source of influence and to identify the importance of content-features in prediction. To maximise cross-platform compatibility, these content-features will be from a post-and-reply relationship, as we consider this to be the simplest and most common form of dialogue between two users. Our research will additionally attempt to identify interpersonal influence at three levels of resolution (simple, moderate and detailed) as we argue that existing research disregards multi-label influence. In summary, our framework's main principles are content-features, social and behavioural influence prediction, modular design, feature prediction importance, and influence resolutions.

1.1.1 Experimental Setup

As shown in Figure 1, our framework can be assessed in five stages, using a set of large social networking datasets for feature extraction and classifier training. The framework's input is restricted to post and reply text between two users. Due to the framework's modular design, we can also provide topological variables like the number of previous active encounters (exposure), however, this is not a requirement but a source for comparison.

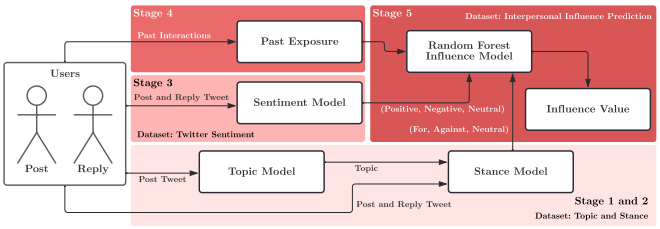

Figure 1. Diagram showing the proposed five stage social and behavioural influence prediction framework.

The first stage will use a fine-tuned BERT-base model to identify the initial topic of conversation from the post text. Reply text is not used for topic detection, as its content is contextually dependent on the initial post. The purpose of the topic model is to supply conversational context to the stance model and aid the interpretation of the user's comment in stage 2. The topic model will be trained on a set of twenty different social and political topics using topic and stance datasets [17][6][10][11][18] gathered by Li, Zhao and Carage [8]. In stage two, we will use another fine-tuned BERT-base model to identify stance features from the text. The input for this model will accept two variables, the retrieved topic extracted from the post, and the post then reply text. The stance model will be trained using the datasets from stage one. The predicted labels in stage two will indicate a for, against, or neutral stance for both the post and reply text (six features total). In stage three, we retrieve the post and reply's sentiment feature, to identify how the users are interacting. We argue that influence may be impacted by users' conversation behaviour in addition to the statements made. The sentiment model will be trained on a dataset by [9], and will indicate a positive, negative or neutral sentiment for the text and reply users (six features total). Whilst stage four's features are optional, they are intended so that we can compare prediction accuracy with and without topological factors included. Stage 4 uses the number of past encounters between the two users and is used as a feature for classification. Due to the framework's modular design, each text-specific model can have a hybrid topological and text model to compare against. This approach to combining and isolating features can be applied across all modular features. The final stage uses a random forest classifier to predict social or behavioural influence across a range of resolutions. We will provide the classifier with unique features and labels attained via the MuMiN Twitter dataset [12]. Whilst Twitter is only one social network, its dynamics represent common social behaviours and engagement patterns [7]. These thirteen features will be attained via stages 1-4, using previously unseen post-reply dialogue and used as the classifier's input. Raw influence data will be collected and clustered into groups (indicated by the resolution) as prediction labels. Raw social influence data uses the number of overlapping members in the two user's immediate neighbourhoods, whilst raw behavioural influence data uses the number of times the replying user has retweeted content by the posting user.

By applying this method we will first investigate the framework's prediction capabilities across influence types and resolutions. To compare our framework's accuracy, the difficulty of this task, and the validity of our extracted features, we will then compare our framework to a set of fine-tuned state-of-the-art large language models. Secondly, due to the framework's modularity, we will also assess feature importance via Mean Decrease Impurity, Shapley Additive Explanations, and by isolating feature sets during model training.

1.2 Investigating Visual Influence

Whilst existing text-based influence and group alignment research commonly focus on platform-specific features like hashtags and external hyperlink associations [16], image content remains an undervalued source for identifying influence as it relates to specific communities. As such, we propose an engagement prediction framework that uses popular social images (or memes) and a user's group associations to determine how a user will engage with an image. This is based on the hypothesis that a user's associated neighbourhood can act as a character profile that can determine image engagement. Our research will aim to predict three engagement characteristics to determine whether a user would *encounter* the image, *engage* with the image, and their *expressed attitude* towards the image's content. We also aim to contribute a large social media image dataset that will be used to compare state-of-the-art multi-modal large language models (MLLM).

1.2.1 Experimental Setup

Based on our designs in Figure2, our MLLM training framework can be split into three stages: data collection, training, and model comparison. Data collection will begin by determining key social media

political communities that share memes. We will attempt to train the model on a range of communities across the political spectrum. To do this, we will determine political standing by reviewing word embedding and associated news outlet links using the AllSides Media Bias Chart. From each community, we will extract the following: posted memes for model input, a list of users posting on the image forum, and the community's (direct) replies to the post for feature extraction. Once all community forum posts are collected, we retrieve each user's affiliated communities. These community affiliations will be any community that the user has publicly joined or interacted with. To prevent overfitting and reduce input size, a threshold for the most common affiliated communities in the population will be set. A user's community affiliation score will represent their activity in the community proportional to all associated community interactions. In order to prevent the model from generalising opposing viewpoints on a political topic. Our dataset labels will be generated for each image and each collected user, with the encountered label representing whether the image appears in the user's associated communities. The engaged label represents when an image appears in a user's associated communities but that they have not been active in the post. The final attitude label represents the user's sentiment retrieved from their forum reply via large language model feature extraction. The final dataset will have the following properties for training: post identifier, user identifier, image data, list of communities and the user's corresponding affiliation value, encountered label, engagement label, and the list of sentiment feature values (attitude labels).

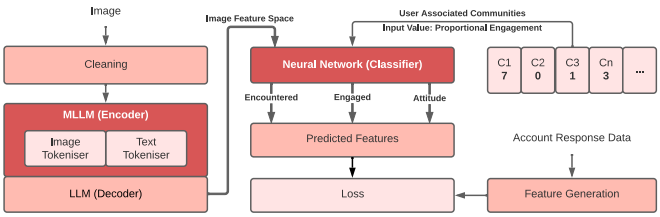

Figure 2. Diagram showing the high-level overview of the proposed image engagement framework.

As shown in Figure2, the framework's input is separated into two stages: image data and community affiliation values. The image data is first prepared and then encoded to be passed to an MLLM sequence-to-sequence encoder, this feature space is then passed to a LLM for decoding. The LLM feature space alongside the user's list of affiliated community values is then passed to a classifier to predict the encountered, engagement, and attitude labels. By applying this method we will investigate the prediction accuracy and validity of the framework across a range of multi-modal large language models, to determine whether image content and connection data are factors of engagement.

1.3 Social Network Gradients of Influence

Using the cross-platform approach from the first experiment proposal, we aim to leverage post-and-reply connections in a larger social network to review how gradients of influence relate to topological and platform behaviour. This will allow us to compare patterns of influence to state-of-the-art group detection models and topological ground truths. Additionally, we aim to contribute a large content-focused social networking dataset, the application and comparison of a content influence framework, and an assessment of how socio-political groups and account types impact gradients of influence.

1.3.1 Experimental Setup

We divide the experimental setup into three stages, data collection, applying influence, and comparison. During data collection, we will retrieve post-reply relationships across different political network communities. To achieve this, we identify community posts using news-outlet links, mentions, or associated hashtags. Labelling news outlets' political stances will be based on resources from the AllSides media bias chart. The originating post-reply relationship for each community will be an internal or external news outlet headline and a platform user's replying comment. Once the source connection is retrieved, in stage 2 we apply the Inter-Inf framework to the post-reply connection. In the event that two users have more than one connection, the influence value will be proportional to the total. We then add the replying users to an edge user list and repeat these steps using the edge user list to identify posts and corresponding replies, until a community depth has been reached. Additional data will be collected for ground truth and account type analysis. The final dataset structure will have the properties shown in Figure 3

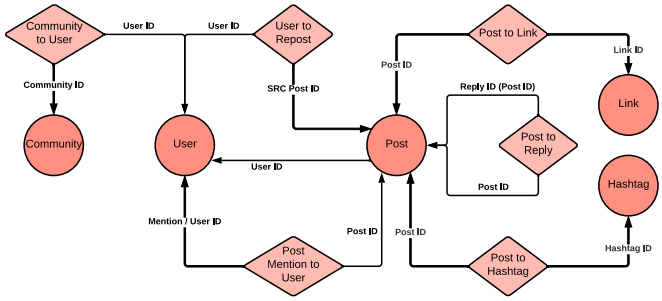

Figure 3. Diagram showing the high-level overview of the proposed image engagement framework.

The final dataset will allow us to compare influence gradient patterns across communities to platform-specific behaviours (reposts, hashtags and links), account type relationships, and against traditional topological group detection methods. Based on this method we aim to review the practical application of content-based influence and its relation to network features.

2 Summary

These proposed linked experiments aim to better understand the influence that takes place in social networks. We argue that by using shared content as the source information rather than the topological symptoms that appear later, we will better identify content and platform features that elicit influential social behaviours. The overarching proposal represents a step towards content-based influence analysis and its wider implications on online communities. We are aware that a future for network influence analysis is likely to become hybrid (to balance resource requirements). However, by demonstrating the value of content features as the originating source of influence, we aim to highlight the importance of content in future network analysis methods.

References

[1] A. Azcorra, L. F. Chiroque, R. Cuevas, A. Fernández Anta, H. Laniado, R. E. Lillo, J. Romo, and C. Sguera, 'Unsupervised Scalable Statistical Method for Identifying Influential Users in Online Social Networks', *Scientific Reports*, **8**(1), (2018).

[2] Hongbo Bo, Ryan McConville, Jun Hong, and Weiru Liu, 'Social network influence ranking via embedding network interactions for user recommendation', pp. 379–384. Association for Computing Machinery, (2020).

[3] Hongbo Bo, Ryan McConville, Jun Hong, and Weiru Liu, 'Social influence prediction with train and test time augmentation for graph neural networks', volume 2021-July, (2021).

[4] Elizabeth Dubois and Grant Blank, 'The echo chamber is overstated: the moderating effect of political interest and diverse media', *Information Communication and Society*, **21**(5), 729–745, (2018).

[5] Santo Fortunato and Darko Hric, 'Community detection in networks: A user guide', *Physics Reports*, **659**, 1–44, (2016).

[6] Lara Grimminger and Roman Klinger, 'Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection', in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 171–180, Online, (2021). Association for Computational Linguistics.

[7] Kaitlin M. Lewin, Morgan E. Ellithorpe, and Dar Meshi, 'Social comparison and problematic social media use: Relationships between five different social media platforms and three different social comparison constructs', *Personality and Individual Differences*, **199**, 111865, (2022).

[8] Yingjie Li, Chenye Zhao, and Cornelia Caragea, 'Improving stance detection with multi-dataset learning and knowledge distillation', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6332–6345, Online and Punta Cana, Dominican Republic, (2021). Association for Computational Linguistics.

[9] Wei Chen Maggie, Phil Culliton. Tweet sentiment extraction. https://kaggle.com/competitions/tweet-sentiment-extraction, 2020.

[10] Lin Miao, Mark Last, and Marina Litvak, 'Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures', in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, (2020). Association for Computational Linguistics.

[11] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko, 'Stance and sentiment in tweets', *ACM Trans. Internet Technol.*, **17**(3), (2017).

[12] Dan S. Nielsen and Ryan McConville, 'Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset', in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 3141–3153, New York, NY, USA, (2022). Association for Computing Machinery.

[13] Sancheng Peng, Yongmei Zhou, Lihong Cao, Shui Yu, Jianwei Niu, and Weijia Jia, 'Influence analysis in social networks: A survey', *Journal of Network and Computer Applications*, **106**, 17–32, (2018).

[14] Mario Alfonso Prado-Romero, Alberto Fernández Oliva, and Lucina García Hernández, 'Identifying twitter users influence and open mindedness using anomaly detection', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11047 LNCS, pp. 166–173. Springer Verlag, (2018).

[15] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang, 'Deepinf: Social influence prediction with deep learning', pp. 2110–2119. Association for Computing Machinery, (2018).

[16] Karishma Sharma, Emilio Ferrara, and Yan Liu, 'Characterizing online engagement with disinformation and conspiracies in the 2020 u.s. presidential election', *Proceedings of the International AAAI Conference on Web and Social Media*, **16**(1), 908–919, (2022).

[17] Parinaz Sobhani, *Stance Detection and Analysis in Social Media*, Ph.D. dissertation, 2017.

[18] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych, 'Cross-topic Argument Mining from Heterogeneous Sources', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 3664–3674, (2018).

[19] Zhenhua Tan, Fan Li, and Danke Wu, 'Mrainf: Multilayer relation attention based social influence prediction net with local stimulation'. Institute of Electrical and Electronics Engineers Inc., (2021).

[20] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang, 'Social influence analysis in large-scale networks', (2009).

[21] Muheng Yang, Xidao Wen, Yu Ru Lin, and Lingjia Deng, 'Quantifying Content Polarization on Twitter', in *Proceedings - 2017 IEEE 3rd International Conference on Collaboration and Internet Computing, CIC 2017*, volume 2017-January, pp. 299–308. Institute of Electrical and Electronics Engineers Inc., (2017).

[22] Sarita Yardi and Danah Boyd, 'Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter', *Bulletin of Science, Technology & Society*, **30**, 316–327, (2010).