# Predicting Interpersonal Influence from Conversational Features

1st Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

***Abstract*—Interpersonal influence has a radical impact on the dissemination of information in online social media. Methods for measuring this influence between online conversation partners are often over-reliant on platform-level features, rendering them inoperable in other settings. We propose a novel and portable solution using Transformers to derive features of conversations that indicate influence. In an evaluation across a diverse discussion dataset, we show that our framework competes with existing state-of-the-art large language models, being able to predict both social and behavioural measures of influence accurately, and at different levels of resolution, with a Macro-F1 above 0.91 in all cases of social influence.**

***Index Terms*—Artificial neural networks, Social network services, Natural language processing, Text analysis**

## I. Introduction

Modern online social media enables billions of conversations between people. These conversations take place in a variety of platform architectures, via a number of modalities, and their content can cover the entire range of human interests. A common factor in many human conversations is that one or both of the parties uses the conversation to exert influence over the other party's point of view or future behaviour. Predicting the magnitude of this influence between two conversation partners is an important problem in a range of application areas, including enabling effective information dissemination [1], understanding historical social trends [2], halting the spread of misinformation [3], and commercial advertising.

Many existing platform architectures include structural indicators that can be predictive of influence. Depending on the platform, a 'friend' or 'following' relationship can indicate a greater trust or dependence between parties, which can translate to a greater effect on the viewpoint or behaviour of a connection. Some users can also be identified as more generally influential within a network because of the number of other users with whom they have such structural connections. These indicators are powerful, but limited, neglecting influential online interactions that are not (or not yet) recognised with a structural indicator, and they can be highly platform-dependent [4]. Whilst the growth of social media means that increasing amounts of data is available on online interactions [5], [6], methodologies using these structural indicators for modelling influence can struggle when they need to be applied in other platforms, such as in traditional web forums, or interactions in large chatrooms [6]. Structural indicators also neglect 'controversial influence' [7]–[10], in which social media users interact and make connections to but are not influenced by people whom they actively disagree

In this paper, we propose a novel framework that uses features of the conversation itself to predict the influence two parties may have on each other. We chose this approach because we believe it to be the most portable and generally applicable, with the conversation features we use being extractable from any online interaction, regardless of the platform.

As this form of conversation-based interpersonal influence prediction (unlike predicting general degree of influence in a network) is a novel task, there are no standard metrics or existing frameworks with which to evaluate performance. Therefore, we define influence metrics using *social* and *behavioural* markers of influence. We use matching followers as a measurement of 'social influence' due to the impact social processes have on social network topology. We argue that the act of two people gravitating to similar neighbourhoods indicates an influenced user desiring more exposure to another user or their network. We believe a future retweet is an intuitive example of an influential source impacting another user's behaviour. This behaviour is an active choice by someone to project a message or show support, both of which indicate a motivation to associate themselves with the user or their message. A system that is able to predict both social and behavioural influence outcomes from a conversation, at a range of resolutions, is, we argue, modelling interpersonal influence in that conversation.

Due to the novelty of this task, there is no existing approach that tackles this conception of influence, with the most similar literature tackling the identification of influential nodes within social networks (which does not address whether a node exerted influence in a particular conversation) or predicting certain behaviours such as connection-forming at a macro-level. To provide context for our framework designed as a solution to this task, we assess the task's difficulty by comparing our framework's performance against that of several large language classifier models (BERT, RoBERTa, DistilBERT, GPT2) and baseline classifiers using our own feature set (most frequent class, stratified random, and uniform random).

We train and evaluate our model using *social influence* (SI): the degree of overlap in conversation partners' structural

connections on a platform, and *behavioural influence* (BI): the likelihood of future signal-boosting behaviour between the two parties (e.g., re-sharing content). Alongside this, we assess the level of contribution to BI and SI for each conversation factor under different circumstances, in an evaluation carried out on a large social-networking dataset. In accordance with our aim for a method that is not constrained to a particular domain, our evaluation includes discussions across twenty different topics, as well as different levels of resolution in quantifying influence. In short, the research questions we address are:

- **RQ1:** To what extent can interpersonal influence be predicted by an online conversation's content?
- **RQ2:** Is there a significant change in model accuracy when predicting different levels of social and behavioural influence?
- **RQ3:** How do conversation features contribute to predicting interpersonal influence?
- **RQ4:** How well can conversation features predict influence in isolation?

The remainder of this paper is organised as follows. In Section II we survey existing methods for predicting wider-group influence in social networks. Section III describes our datasets and influence-modelling methodology. Section IV presents the results of our experiments as applied within a large Twitter corpus. Finally, we conclude with some key observations.

## II. Related Work

Topology-based detection posits that the types of relationships between nodes that can lead to behavioural or social changes can be detected by looking at a network's structural features. For instance, outlier detection identifies influential nodes in a network by the ratio of ingoing and outgoing node connections across the network [11]. In a social network, these influential nodes act like a heat source, where actions have a transitive cascading effect, diffusing into the community [10]. The features used to determine nodes' in/out influence ratios are often platform dependent features, such as the number of followers, the number of posts or the number of retweets [12]. Gaussian outlier detection can demonstrate this principle by detecting anomalous and influential nodes with more internal community edges than the network average, and shows that news outlets and politicians hold the highest number of connections, indicating greater influence [1].

Whilst there are examples of effective topological methods for monitoring and predicting interpersonal influence of an individual based on a wider group, these methods are often binary in resolution and depend heavily on structural connections, neglecting much of the content of interactions (which can lead to controversial influence false-positives [7] [8] [9] [10]). Qiu et al. developed the DeepInf framework [13] predicts an influential chain-reaction with a latent space that reflects network structure features and a binary status (indicating an action) for neighbouring nodes. A Graph Attention Transformer (GAT) layer in the DeepInf framework provides the attention coefficients to measure the contributed importance between graph vertices. As topological features are dependent on structure, smaller graphs with fewer edges provide less data for models to learn graph features. As a solution, models such as AugInf use graph augmentation during training and testing to create additional edges to a network's sub-groups via a Variational Graph Auto Encoder [14]. These edges are based on their predicted likelihood and provide more latent space detail for smaller networks. Similar to DeepInf, the desire to increase node structural details within latent feature spaces is also researched in the MRAInf framework [15]. The MRAInf framework was designed in response to the 1-Weisfeiler-Lehman restriction by introducing a local stimulation mechanism containing multiple 1-Dimensional convolution blocks to provide discriminate reinforcement between feature maps.

Content-based influence detection traditionally uses a sample of media content to predict a single resolution of social or behavioural influence [8]. In principle using content provides a tailored set of assets for each node that are less dependent on structural information and platform architecture (which can contain edges between nodes for reasons other than shared viewpoints [4]). Several methods can identify linguistic characteristics of a discussion. Most commonly, low resource-intense methods are used to extract discussion characteristics using keyword extraction to determine generalised endorsements (i.e. hashtags or links to political websites) [3]. This method whilst requiring less resources, provides limited resolution, has dependencies on platform features (hashtag implementation), and requires a wider network to draw conclusions on behaviour. However, whilst more resource-intensive, NLP Transformers can be trained to provide semantic and local context to words within a sentence using word-embeddings [16]. After fine-tuning, these models can classify discussion characteristics such as sentiment and stance [17]. Although resource intensive, we argue that the use of word embedding (over traditional low resource methods) allow us to map words to lower dimensional space, capturing unique semantic and structural features which provide insight into otherwise unknown influential principles, at a range of resolutions, and without dependencies on platform specific content or the wider network [18].

Whilst network structures and conversational features provide insight into the influence of a node within a network, research in Topic Affinity Propagation (TAP) has provided a combined approach including real world examples of topic separation and influential clustering [19]. Whilst network structures can indicate the amount of influence a node has, real world social networks are noisy, with users having differing levels of influence on different topics. TAP uses affinity propagation on a topical factor graph model, where it receives the structure and topic features for a sub-network of nodes. The model is trained via distributed learning for topic sensitive clustering. In applications beyond topic sensitivity, the EIRank model embeds sub-networks representing different types of interactions (retweets, replies, mentions, favourites, following) such that interaction types have differing levels of contribution

for showing signs of influence [20]. The features related to the interaction between nodes in an influence cascade [21] expands this definition further, and uses platform related interactions as features to indicate influence: initiation, contribution, sharing. The types of behavioural influence being used across the cascading levels are visualised through a Sankey diagram, where users are divided into levels based on their distance from the central node. The network's structure and features allow us to identify cultural and behavioral differences in the methods people use to influence other's politics on social media [22], [23].

## III. Methodology

In this study we base our framework first and foremost around content-based features, to uniquely provide a non-platform-dependent influence prediction framework. A conversation's attributes can be defined as $a = \{stance, sentiment, exposure\}$. That is: for each speaker, the stance they take on the topic of conversation, the tone of their contribution to the conversation, and their prior exposure to this conversation partner. For instance, consider the graph dataset $G = (U, C)$, where $U$ is the set of users (nodes) with an individual user being represented as $u_i \in U$, and $C$ is a set of conversations (directed edges) between users, with an individual conversation being represented as $c_i \in C$. In conversation $c_1$, if user $u_1$ responds to a statement by $u_2$, it is possible for $c_1$ to contain features that indicate whether $u_1$ is being influenced. We define influence in two proxy categories: social influence $s$ as the number of matching followers between $u_1$ and $u_2$, and behavioural influence $b$ as the number of retweets by $u_1$ from $u_2$. This relationship can be summarised as: $a_1 \frown a_2 \Rightarrow s_1, b_1$ or $c_1 \Rightarrow s_1, b_1$.

As shown in Figure 1, our machine learning framework has five stages and is trained and validated on a conglomerate of large social-networking datasets. The first stage uses the text of a post to predict the topic of conversation in each interaction, with an interaction defined as a post and a reply to that post. Once the conversation topic is predicted, in the second stage we predict each conversation partner's stance on the topic. The third stage then examines both post and reply to understand the users' conduct as expressed in the sentiment of their text. In the fourth stage, we count the number of past encounters to understand the amount of previous exposure these two users have to one another. The final stage uses stance, sentiment, and exposure as conversational features in a Random Forest classifier to predict the degree of social or behavioural influence. Our motivation in using these modular classifiers across the framework's 5 stages [Figure 1] provides novel and measurable insight into each factor's contribution to SI and BI. Thus we evaluate each model's performance and compare to state-of-the-art LLMs to provide confidence in our modular features and the framework's ability to answer our RQs. All models were trained until experiencing a plateau on the validation macro F1 score.

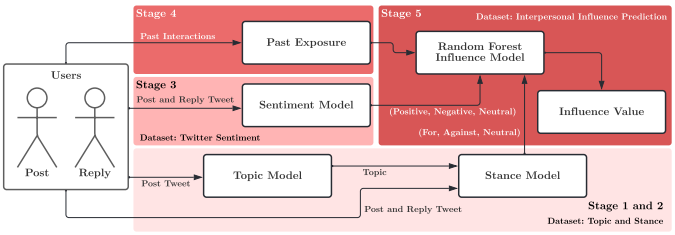


Fig. 1: Diagram showing the proposed social and behavioural influence prediction framework in five stages. Stages 1-4 retrieve conversation features from the text and Stage 5 derives an influence score from these features.

### A. BERT Classifier

BERT models provide the basis for our word embeddings and vectorisation in Stages 1-3 [Figure 1]. Each dataset is divided into training (70%), validation (20%), and testing (10%) [17], [24]. We adopt a pre-trained *bert-base-uncased* model [25] that has been trained in a self-supervised manner using the English Wikipedia site and Book Corpus (11038 unpublished books) [25]. We use the model's input structure and syntax *[CLS] Sentence A [SEP] Sentence B [SEP]* to distinguish between sentences [17], [26] and train the BERT multi-class-classifier at a leaning rate of 3e-5 (concluded from Bayesian Optimisation tests). The external linear layer uses an AdamW Optimizer algorithm and Cross-Entropy Loss function [26], [27].

### B. Stage 1 and 2 - Topic and Stance Model

*1) Dataset* - Topic and Stance*:* Motivated to have a large corpus of political topics and nuance behaviours, for Stages 1 and 2, we use five datasets gathered by Li, Zhao and Caragea for stance and topic detection [27]:

- The MT-Stance dataset [28] contains multi-labelled Twitter posts collected during the United States 2016 presidential election. The influence of election cycles provide topical relevance in detecting western political figures.
- The Trump-Biden dataset [29] covers the six weeks before the United States 2020 presidential election. This data expands upon MT-Stance, providing more political candidates as well as topical events.
- The COVID dataset [30] contains labelled Tweets from the United States during the COVID-19 global pandemic with stances and sentiment on COVID-19 restrictions.
- The Controversial-Six dataset [31] contains Tweets from the United States between 2016-2017 regarding stances on six controversial political topics.
- The Controversial-Eight dataset [32] provides seven additional controversial topics, increasing the relevance of the topic detection model.

The content of the five datasets were combined, translated, and condensed to Tweet, topic, and stance. We also augmented the dataset with a random sample of Tweets to reflect non-political discussions. We merge column labels from different

datasets that have identical topics (e.g. politician names), and remove any unnatural message structures like Tweet @ symbols, hyperlinks, and unidentifiable characters. We keep hashtag text but the octothorpe symbol is removed and underscore characters are replaced with spaces. The final dataset consists of 44500 Tweets covering 21 topics[1].

*2) Dataset* - ADDITIONAL TOPIC AND STANCE DATASET (ATS)*:* To increase our confidence about the generalisability of our model trained on the stance and topic datasets provided by Li, Zhao and Caragea, we created an additional topic and stance (ATS) dataset, which consists of 1000 Tweets posted internationally between 2021 and 2022, for each of the 21 topics from the *Topic and Stance* dataset. Tweets were retrieved through searches using associated positive and negative keywords.

*3) Model Training:* The topic detection model ($f_a$) for Stage 1 was trained using the *Topic and Stance* dataset to identify political topics in a sentence. Conversation sentences were tokenised ($\boldsymbol{G} \in \mathbb{N}^{m \times n}$) and applied to the topic model ($\boldsymbol{X} = f_a(\boldsymbol{G}) | \boldsymbol{X} \in \sigma(\mathbb{R}^{m \times 21})$) providing a sigmoid topic probability vector for each conversation. Our motivation for topic retrieval was to increase the conversation's context. Thus, the most likely topic was applied with the original sentence to the stance classifier model ($f_b$) to predict the labels: *against, for, neutral* ($\boldsymbol{Z}_1 = f_b(\boldsymbol{G}, \boldsymbol{X}) | \boldsymbol{Z}_1 \in \sigma(\mathbb{R}^{m \times 3})$). After fine-tuning, our topic model has a Macro-F1 test score of 0.671. The stance detection model for Stage 2 was fine-tuned on the same dataset and achieved a Macro-F1 test score of 0.671. We then test both models using the ATS dataset. The topic detection model achieves a Macro-F1 score of 0.756. Each of the stance topics' were independently tested and averaged at 0.837 with a standard deviation of 0.037.

### C. Stage 3 - Sentiment Model

For Stage 3's sentiment model training, we use the *Twitter Sentiment* dataset [33] containing 27481 Tweets labeled for sentiment classification. The sentiment model ($f_c$) is also trained using tokenised word-vectors ($\boldsymbol{Z}_2 = f_c(\boldsymbol{G}) | \boldsymbol{Z}_2 \in \sigma(\mathbb{R}^{m \times 3})$) and evaluates social conduct in conversation based on three sentiment labels (positive, negative, and neutral), and was split into training (70%), validation (20%), and testing (10%) datasets. After classifier training, the model has a Macro-F1 test score of 0.738.

### D. Stage 4 and 5 - Interpersonal Influence Model

To train our interpersonal influence model, we extract content features (stance ($\boldsymbol{Z}_1$), sentiment ($\boldsymbol{Z}_2$), exposure ($\boldsymbol{e} \in \mathbb{N}^m$)) from the MuMiN dataset [34], which provides training data ($\boldsymbol{F} = \boldsymbol{Z}_1 \frown \boldsymbol{Z}_2 \frown \boldsymbol{e}^T | \boldsymbol{F} \in \mathbb{R}^{m \times 13}$)) for the social and behavioural influence classifiers. This dataset consists of 20000 users and 71000 conversations and required updated follower details.

[1]Abortion, Atheism, Bernie Sanders, Climate Change is a Real Concern, Cloning, the Death Penalty, Donald Trump, Face Masks, Anthony Fauci, the Feminist Movement, Gun Control, Hillary Clinton, Joe Biden, Marijuana Legalization, the Minimum Wage, Nuclear Energy, School Uniforms, School Closures, Stay at Home Orders, Ted Cruz, and Random Tweets belonging to none of the previous categories.

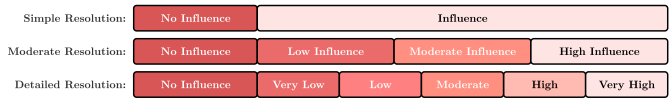


Fig. 2: Resolutions of interpersonal influence. We derive these thresholds using K-Means clustering on the influence proxy variables.

Exposure is assigned to each conversation from the number of post-reply relationships in the dataset between the two interacting users. Motivated to measure categorising capabilities, our Random Forest classifier[2] assesses interpersonal influence at three different resolutions. We use K-means clustering to categorise resolution training labels [Figure 2] which are assigned by: $\sum_j^n [\boldsymbol{g}_j < h]$, where $\boldsymbol{g} \in \mathbb{R}^n$ holds the lower boundaries for each resolution and the scalar $h$ represents the two users' matching follower or retweet count. Training results will be discussed in Section IV.

## IV. RESULTS

In this section we present answers to our main research questions, including whether the features of a brief conversation can predict measures of social and behavioural influence (RQ1) across ranges of detail (RQ2). We also examine the relative importance and isolated predictive power of subsets of conversation features (RQ3, RQ4). We compare our framework, the complexity of the task, and the validity of our extracted features against state-of-the-art LLMs and baseline classifier comparisons (most frequent class, uniformly random, stratified random). Compared to LLMs and baseline classifiers, our framework shows good and more consistent prediction rates across label resolutions. Further investigation indicates that the exposure feature has the most significant impact on predicting social and behavioural influence, followed by post features. However, it should be noted that exposure struggles to predict influence independently and that other conversation features like sentiment perform better [Table II].

### A. Predictions Across Label Resolutions (RQ1, RQ2)

The social influence tests [See Table I] show good performance across label resolutions, with the detailed social influence model achieving a Macro-F1 score of 0.9334. The most accurate social influence model was moderate, followed by detailed and simple. Both precision and recall scores across all social influence model categories remain above 0.91, with precision averaging 0.9402, whilst recall averaged 0.9211. The poorest recall, precision, and Macro-F1 trend across detailed and moderate models were the lowest social influence labels (low and very low). In comparison, the most accurate behavioural influence test was the simple model achieving a Macro-F1 score of 0.891. Both simple and moderate behavioural models had larger Macro-F1 label inconsistency, with the Moderate model ranging from 0.7058 (High) to 0.9986 (None).

[2]We evaluated other classifiers, with the Random Forest performing best.

| | Social | | | Behavioural | |
|---|---|---|---|---|---|
| | S | M | D | S | M |
| InterInf Framework Labels (F1 Score) | | | | | |
| None | 0.919 | 0.922 | 0.918 | 0.998 | 0.998 |
| Very Low | 0.914 | 0.893 | 0.875 | 0.783 | 0.741 |
| Low | | | 0.986 | | |
| Medium | | 0.977 | 0.909 | | 0.903 |
| High | | 0.967 | 0.944 | | 0.705 |
| Very High | | | 0.966 | | |
| Model Comparison (Macro F1 Score) | | | | | |
| **InterInf-Framework** | **0.916** | **0.94** | **0.933** | 0.891 | **0.837** |
| BERT-Base | 0.91 | 0.91 | 0.226 | **0.996** | 0.249 |
| DistilBERT | 0.898 | 0.848 | 0.823 | 0.988 | 0.339 |
| GPT2 | 0.904 | 0.84 | 0.834 | 0.983 | 0.261 |
| RoBERTa | 0.0 | 0.168 | 0.112 | 0.996 | 0.249 |
| Simple Baseline Classifier (Mean Average) | | | | | |
| Stratified | 0.492 | 0.247 | 0.167 | 0.503 | 0.253 |
| Uniform | 0.507 | 0.249 | 0.17 | 0.501 | 0.25 |
| Most Frequent | 0.49 | 0.08 | 0.005 | 0.006 | 0.0 |

TABLE I: Table showing interpersonal influence tests across resolutions: Simple (S), Moderate (M), Detailed (D). Including label and model comparisons, with row-spans representing the resolution's label categories [See Figure 2]. Our framework's best Macro-F1 label and resolution is the social influence moderate model.

The prediction results were aligned with our expectations for RQ1 and RQ2, as the framework outperforms simple baseline classifiers, and is comparable to state-of-the-art large language models (outperforming a standard BERT model in most categories). Considering this, we are confident that the features we extract (stance, sentiment, and exposure) are justified to be assessed as indicators of influence and can provide more transparent assessment than a black-box feature space. Furthermore, considering the framework's combined modular size, it is far more consistent in prediction accuracy across categories compared to single large language models. We also note that we suspect the RoBERTa model would perform significantly better in social categories if given more suitable hardware during training.

### *B. Prediction Feature Importance (RQ3, RQ4)*

In this section we review RQ3 and RQ4, by identifying the contribution of conversation feature sets when determining the social and behavioural models' prediction results. We explore the extent of this contribution under two circumstances: the feature's *collective contribution* alongside others; and the feature's *isolated contribution*, where we measure an independent feature or feature set's prediction accuracy[3]. To retrieve a feature or feature set's isolated accuracy, new interpersonal influence models were trained.

In answering RQ3, Figures 3 and 4 test results provide us with details on collective feature importance. Consistently across models, exposure was the dominant feature in predictions, with its contribution ranging from 23% to 28%.

[3]We use the following features and feature sets: post stance, post sentiment, all post, reply stance, reply sentiment, all reply, and exposure.

| | Social | | | Behavioural | |
|---|---|---|---|---|---|
| | S | M | D | S | M |
| all features | 0.916 | 0.94 | 0.933 | 0.891 | 0.837 |
| all sentiment | **0.9014** | **0.9117** | **0.8952** | **0.8609** | **0.8288** |
| all stance | 0.8891 | 0.89 | 0.8264 | 0.7125 | 0.7247 |
| all reply | 0.8469 | 0.8785 | 0.873 | 0.8112 | 0.749 |
| reply sentiment | 0.8452 | 0.8695 | 0.8573 | 0.8076 | 0.7045 |
| all post | 0.8637 | 0.8372 | 0.7382 | 0.6113 | 0.4895 |
| post sentiment | 0.8626 | 0.8270 | 0.7362 | 0.6114 | 0.4896 |
| post stance | 0.8497 | 0.8105 | 0.6586 | 0.6086 | 0.4887 |
| reply stance | 0.8152 | 0.7606 | 0.7008 | 0.5157 | 0.5352 |
| exposure | 0.7589 | 0.6661 | 0.54 | 0.4068 | 0.2624 |

TABLE II: Table showing Macro-F1 tests of isolated conversation features and feature sets across resolutions: Simple (S), Moderate (M), Detailed (D). This test demonstrates that the *all sentiment* feature set has the highest prediction accuracy and exposure has the poorest. Isolated conversation feature sets can be compared to feature models demonstrated in Table I.

We found that whilst there was a positive association with exposure in social influence models, the reverse was true for behavioural influence. However, both show a negative association with post positive sentiment scores. Post features are also consistently considered more important across models over reply features. We also find no consistent trend in stance and sentiment dominance toward final predictions. In answering RQ4, we find that exposure and independent stances are poorer at predicting social influence when isolated, with an average Macro-F1 score below 0.78, in comparison to reply features (such as reply sentiment) which reach an average Macro-F1 score of 0.8573 [Table II]. We notice an additional divide when comparing all post (0.813) and all reply (0.8661) average Macro-F1 values. Furthermore, when comparing stance (all stance) and sentiment (all sentiment), sentiment has the best Macro-F1 score for social influence averaging at 0.9027. Behavioural influence reflects these results but consistently has a lower Macro-F1 score in each category, as well as the all reply feature set having a higher Macro-F1 score compared to all stance by up to 10%. The impact of behavioural influence features has also changed [Figure 4] for positive (post neutral stance and sentiment) and negative associations (post stance against).

The results of the collective feature importance tests [Table 3] were aligned with our expectation that exposure has the most significant impact in predicting both social and behavioural influence. We hypothesise that the reason posts have higher feature importance in comparison to replies, is that a conversation relies on the composure of the initial statement and that the reply in most circumstances follows a pattern based on this. When isolating these features [Table II], we see a switch in importance, with reply sentiment data having better Macro-F1 scores in comparison to post and exposure features. This is consistent with the discussed theory and would suggest that isolated reply features become more important for prediction.

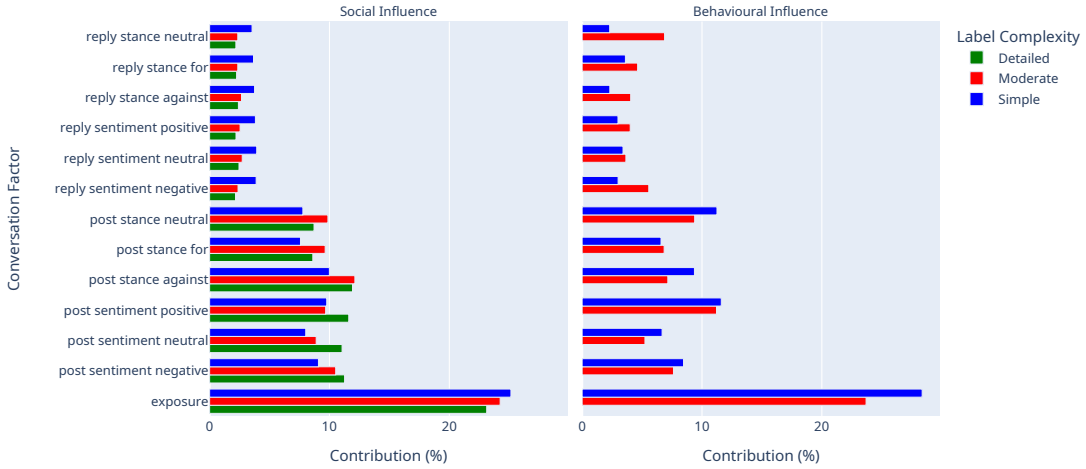


Fig. 3: Bar Charts showing collective feature importance tests for predicting social and behavioural influence. In both examples, the exposure value has the most significant impact on interpersonal influence, followed by post features.

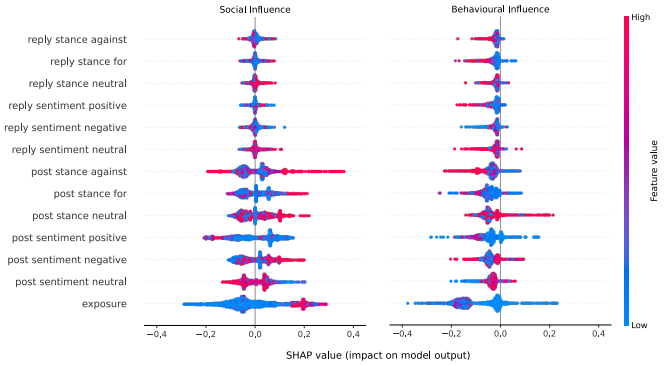


Fig. 4: SHAP diagram of the simple resolution for the social and behavioural influence model showing how each features' value impacts influence label prediction.

## V. Conclusion

Determining the extent of online influence between two people is a challenging issue and analysis can often be limited to analysing platform dependent structures and features that correlate to changes in behaviour. Our research addresses this issue by providing a novel way to measure a user's political and social-economic influence using universal factors that are non-platform-specific. Examining conversations between pairs of users, we contribute a novel content-based approach, that avoids dependence on topological features and was trained and tested on large social-networking datasets to extract key features of the conversation: its topic, sentiment, the prior exposure of the two speakers, and their stance on the topic. From these features, we demonstrate that we can predict measures of social and behavioural influence with a Macro-F1 score of 0.94 and 0.89, with our model showing more consistent performance than state-of-the-art large language models. Our research also provides novel insight into each factor's interpersonal influence contribution, finding that prior exposure between people has the highest impact on social and behavioural influence predictions within the model, but this feature is a poor predictor by itself. We also discover that that a post's conversation features have more impact on the final prediction than those of the reply. In summary, our research is intended as an important stepping stone in determining online social influence based on universal content seen both online and offline. Our research shows strong predictive performance for social and behavioural influence and gives insight into feature importance across twenty political topics.

## References


[1] M. A. Prado-Romero, A. F. Oliva, and L. G. 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)*, vol. 11047 LNCS. Springer Verlag, 2018, pp. 166–173.

[2] D. S. Liu and G. Y. Si, "A social influence model based on interpersonal relationship," in *Proceedings - 2011 International Conference on Instrumentation, Measurement, Computer, Communication and Control, IMCCC 2011*, 2011, pp. 581–584.

[3] K. Sharma, E. Ferrara, and Y. 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*, vol. 16, no. 1, pp. 908–919, 2022.

[4] G. Villa, G. Pasi, and M. Viviani, "Echo chamber detection and analysis," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 78, 2021.

[5] M. S. Rahman, S. Halder, M. A. Uddin, and U. K. Acharjee, "An efficient hybrid system for anomaly detection in social networks," *Cybersecurity*, vol. 4, no. 1, p. 10, 2021.

[6] Eeti, A. Singh, and H. Cherifi, "Centrality-Based Opinion Modeling on Temporal Networks," *IEEE Access*, vol. 8, 2020.

[7] E. Dubois and G. Blank, "The echo chamber is overstated: the moderating effect of political interest and diverse media," *Information Communication and Society*, vol. 21, no. 5, pp. 729–745, 2018.

[8] M. Yang, X. Wen, Y. R. Lin, and L. Deng, "Quantifying Content Polarization on Twitter," in *Proceedings - 2017 IEEE 3rd International Conference on Collaboration and Internet Computing, CIC 2017*, vol. 2017-January. Institute of Electrical and Electronics Engineers Inc., 2017, pp. 299–308.

[9] S. Yardi and D. Boyd, "Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter," *Bulletin of Science, Technology & Society*, vol. 30, pp. 316–327, 2010.

[10] S. Peng, Y. Zhou, L. Cao, S. Yu, J. Niu, and W. Jia, "Influence analysis in social networks: A survey," *Journal of Network and Computer Applications*, vol. 106, pp. 17–32, 2018.

[11] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.

[12] 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*, vol. 8, no. 1, 2018.

[13] J. Qiu, J. Tang, H. Ma, Y. Dong, K. Wang, and J. Tang, "Deepinf: Social influence prediction with deep learning." Association for Computing Machinery, 2018, pp. 2110–2119.

[14] H. Bo, R. McConville, J. Hong, and W. Liu, "Social influence prediction with train and test time augmentation for graph neural networks," vol. 2021-July, 2021.

[15] Z. Tan, F. Li, and D. Wu, "Mrainf: Multilayer relation attention based social influence prediction net with local stimulation." Institute of Electrical and Electronics Engineers Inc., 2021.

[16] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware BERT for language understanding," in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020.

[17] J. W. Sirrianni, X. Liu, and D. Adams, "Predicting Stance Polarity and Intensity in Cyber Argumentation with Deep Bidirectional Transformers," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 3, pp. 655–667, 2021.

[18] V. Morini, L. Pollacci, and G. Rossetti, "Toward a standard approach for echo chamber detection: Reddit case study," *Applied Sciences (Switzerland)*, vol. 11, no. 12, 2021.

[19] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," 2009.

[20] H. Bo, R. McConville, J. Hong, and W. Liu, "Social network influence ranking via embedding network interactions for user recommendation." Association for Computing Machinery, 2020, pp. 379–384.

[21] C. Senevirathna, C. Gunaratne, W. Rand, C. Jayalath, and I. Garibay, "Influence cascades: Entropy-based characterization of behavioral influence patterns in social media," *Entropy*, vol. 23, pp. 1–26, 2021.

[22] G. V. Lukyanova and D. S. Martyanov, "Influence of Communication Strategies on the Structure of Political Discussions," in *Proceedings of the 2021 Communication Strategies in Digital Society Seminar, ComSDS 2021*. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 126–129.

[23] S. Bell and M. Kornbluh, "Networking in the digital age: Identifying factors that influence adolescents' online communication and relationship building," *Applied Developmental Science*, vol. 26, no. 1, pp. 109–126, 2022.

[24] G. K. W. Huang and J. C. Lee, "Hyperpartisan News and Articles Detection Using BERT and ELMo," in *2019 International Conference on Computer and Drone Applications (IConDA)*, 2019, pp. 29–32.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186.

[26] S. X. Lin, B. Y. Wu, T. H. Chou, Y. J. Lin, and H. Y. Kao, "Bidirectional Perspective with Topic Information for Stance Detection," in *Proceedings - 2020 International Conference on Pervasive Artificial Intelligence, ICPAI 2020*. Institute of Electrical and Electronics Engineers Inc., 2020, pp. 1–8.

[27] Y. Li, C. Zhao, and C. Caragea, "Improving stance detection with multi-dataset learning and knowledge distillation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 6332–6345. [Online]. Available: https://aclanthology.org/2021.emnlp-main.511

[28] P. Sobhani, "Stance Detection and Analysis in Social Media," Ph.D. dissertation, 2017.

[29] L. Grimminger and R. 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*. Online: Association for Computational Linguistics, 2021, pp. 171–180.

[30] L. Miao, M. Last, and M. 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: Association for Computational Linguistics, 2020.

[31] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Trans. Internet Technol.*, vol. 17, no. 3, 2017.

[32] C. Stab, T. Miller, B. Schiller, P. Rai, and I. Gurevych, "Cross-topic Argument Mining from Heterogeneous Sources," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 3664–3674, 2018.

[33] W. C. Maggie, Phil Culliton, "Tweet sentiment extraction," https://kaggle.com/competitions/tweet-sentiment-extraction, 2020.

[34] D. S. Nielsen and R. 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*, ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3141–3153.