Discourse, Context & Media 67 (2025) 100929

Contents lists available at ScienceDirect

Discourse, Context & Media

journal homepage: www.elsevier.com/locate/dcm

Developing a discourse space for analysing online discourse

Kateryna Krykoniuk *, Cleo Hopkin-King, Seán G. Roberts *

School of English, Communication and Philosophy, Cardiff University, John Percival Building, Colum Drive, Cardiff CF10 3EU, United Kingdom

ARTICLE INFO

Keywords:
Discourse space
Thematic strategies
Topic derailment
Online discourse
YouTube

ABSTRACT

Understanding the dynamics of online discourse is crucial for dealing with disinformation, radicalisation and hate speech. However, there are few formal models of how commenters orient their messages to each other to create online discourse. We introduce the concept of a 'discourse space'—a novel conceptual framework that serves as an abstract meta-representation of discourse. It provides an opportunity to quantify discourse and explore its dynamics by leveraging a range of possible discourse strategies, spanning four key aspects: cohesion, attitude, logic quality and coherence. With this view, discourse strategies emerge as generalised techniques for linguistically shaping thoughts based on the social context. To construct an empirical space from real data, 1,684 message pairs from 50 YouTube video comment sections were tagged for 25 discourse strategies. Using an advanced dimension-reduction method (t-distributed stochastic neighbour embedding, t-SNE), we demonstrate that a systematic discourse space can be constructed from the data. Specifically, the relations between individual social media messages can be positioned within the discourse space and that messages which attempt to derail the discourse occupy a specific part of this space. Furthermore, there are distinct patterns of discourse derailment within this discourse space that an automatic system could detect.

1. Introduction

A significant challenge in discourse analysis is tracking topic development throughout a discourse. How does a polite conversation about the weather transition to a heated argument about politics? Especially in online social media settings, understanding the dynamics of discontinuous discourse is critical for dealing with issues such as disinformation (e.g. Sabbah, 2024; Igwebuike & Chimuanya, 2021), radicalisation (Wignell et al., 2021; Williams & Tzani, 2024) and hate speech (Döring and Mohseni, 2020; Govers et al., 2023; Sagredos & Nikolova, 2022). However, there are two major barriers to progress: a lack of formal models of how online discourse dynamics work and a lack of suitable gold-standard dataset of comments annotated at the discourse level on which to test these models.

In addition, many discourse features are currently identified primarily at a lower linguistic level, with a predominant focus on specific lexical items, their frequency and their distribution within a text. Thus, a persistent challenge in discourse analysis is the absence of robust methodologies capable of generating a comprehensive, meta-discourse representation of a text or register. In response to this gap, the primary objective of this study is to develop a method for mapping discourse at a higher level—one that enables a holistic interpretation of the phenomenon under investigation and elucidates its connections with other discourse features. This method is then applied to evaluate its potential in detecting discourse derailment—instances of discontinuous discourse—which may function as a strategic technique commonly used in disinformation efforts. Additionally, we examine how the affordances and constraints of YouTube's commenting system shapes emergent discourse and highlight the platform's distinctive communicative features, which offers insights into the nature of mediated discourse and the ways in which technological mediation influences language use.

In this paper, we make progress on overcoming these barriers. Firstly, we suggest that discourse relations between online social media comments can be represented in a 'discourse space' (more formally, a multi-dimensional manifold): discursive transitions between comments can be mapped as points in a conceptual space, where some parts of the space represent more disruptive or extreme changes in topic. In section 2, we explain how previous theories feed into this novel conceptualisation of discourse.

Secondly, to test this conceptual model, we create a new dataset of social media comments that is tagged by human annotators for elements of discourse dynamics. Since this is the first dataset of its kind, the process includes the creation of a data tagging scheme by combining different measures of discourse elements from multiple

* Corresponding authors.
E-mail addresses: krykonyukk@cardiff.ac.uk (K. Krykoniuk), robertss55@cardiff.ac.uk (S.G. Roberts).

https://doi.org/10.1016/j.dcm.2025.100929
Received 4 December 2024; Received in revised form 25 July 2025; Accepted 27 July 2025
Available online 30 July 2025

theories. These measures include a novel measure of the extent that a comment derails the discourse from the previous topic. In section 3, we collect and tag data from YouTube comments according to this scheme. Section 4 and 5 analyse the results to map out comment relations in a discourse space. We show that this space exhibits meaningful patterns and useful in identifying discourse derailment. We hope that the progress on theoretical and practical levels will help create a roadmap for developing an automatic system for the detection of discourse derailment.

2. Background

This section describes the theoretical foundation for the data tagging scheme and for the framework. Subsection 2.1 engages with a model of topic development, and subsection 2.2 introduces discourse aspects that can viewed as dimensions along which discourse evolves. Subsection 2.3 then describes the discourse space as an abstract representation of discourse in a multi-dimensional space.

2.1. Topic development

The study of discourse topic development began in the 1970 s and 1980 s (e.g., Brown & Yule, 1983). Numerous studies within the three distinct theoretical traditions explore various facets of discourse topic development (Watson Todd, 2016: 70): conversation analysis (e.g., Sacks et al., 1974), rhetorical structure theory (e.g., Potter, 2008) and centering theory (e.g., Miltsakaki & Kukich, 2004). In addition, Herring's (1999) seminal work on interactional coherence and topic decay in online discussions has extended the investigation of topic development into the domain of digital discourse.

One of the recent and succinct models of topic development, summarised in Fig. 1, was proposed by Watson Todd (2016, see also Schubert & Renkema, 2018: 118). It makes an important distinction between two types of discourses: continuous and discontinuous. Within a continuous discourse, a topic can either be sustained or undergo drift, achieved through the provision of additional details or by adopting a more generalised perspective. In contrast, the discontinuous discourse is marked by coherent or non-coherent changes of topic. Thus, the development of a topic can be understood as a continuum with three benchmarks: topic maintenance, topic drift and topic shift.

In this view, continuous discourse captures narrative that explores the topic within its semantic field and can be the expression of one's emotional attitude towards a discourse, its elements, and participants (Example 1), or the articulation of one's positions and beliefs on a range of issues.

(1).

Comment A: *I've created a new documentary 'The Sumerians – Fall of the First Cities'.*

Comment B: *This was incredible! I am 81 and was completely into this program. Thank you for a job well done!!*

Furthermore, continuous discourse can involve a gradual shift from one aspect of the topic to another, without a clear break and while maintaining semantic continuity.[1] There is a connection to the previously mentioned elements within a discourse through a logical transition[2] that guides the introduction of new discourse elements to the topic. These newly introduced elements are naturally integrated into a semantic field of the discussed topic (Example 2).

(2).

Comment A: *In a way, grammar is a reflection of our experience.*

Comment B: *Cannot agree more! As Wittgenstein once said, 'Like everything metaphysical, the harmony between thought and reality is to be found in the grammar of the language'.*

By contrast, discontinuous discourse is marked by an abrupt change of the topic to a new one, exhibiting the lowest degree of topic cohesion. In discontinuous discourse, an utterance may be anchored in the previous context but makes a significant leap from the initial point to a new topic. This is illustrated in Example 3, where the second comment, though semantically rooted in the previous discussion about fixing a medical issue, shifts the focus from a shared health problem—a condition currently incurable in modern medicine—to a political context.

(3).

Comment A: *My vision has been severely impaired due to a detached retina. I hope to have it fixed someday.*

Comment B: *It's time to fix people's lives is their political beliefs and to vote for leaders who will actually reform our broken medical system.*

These definitions of continuous and discontinuous discourse enable the conceptualisation of a topic development continuum. However, this requires identifying discourse aspects to map as dimensions of discourse space.

2.2. Discourse aspects

Aspects of discourse are widely recognised foundational elements upon which discourse analysis is built. These include cohesion (e.g., Halliday and Hasan, 1976), *attitude* (Gee, 1999)*, logical quality* (Grice, 1975) and *coherence* (van Dijk, 1980; Bublitz, 2011: 38; Sinclair, 1991: 102). Drawing on our extensive review of the literature, we identify these four dimensions as fundamental to understanding the essential characteristics of effective discourse across a range of theoretical frameworks. Our contribution lies in the synthesis of these traditionally distinct analytical dimensions into a unified framework that captures both structural features (coherence and cohesion) and evaluative aspects (attitude and logical quality). Whereas previous models have tended to focus on isolated dimensions or specific types of discourse, our integrated approach offers a multidimensional account of discourse, captured by Fig. 2.

While other discourse aspects can also be identified (e.g., intentionality and communication modalities), this study focuses on these four key aspects, aligning with its practical aim of analysing YouTube comment sections. These aspects represent the most frequently studied and relevant parts of discourse for our purposes. There are several other aspects of discourse, but they are either absent from the current online context (e.g. intonation, gesture and turn-taking), or invariant in the context (e.g. mode, medium, genre and situationality), minimised by the constraints of the system (e.g. power), not reliably accessible (e.g. identity) or emergent from the aspects we explore (e.g. thematic progression).

In this study, the aspects of cohesion, attitude and logical quality are measured objectively with the help of *thematic strategies*—general logico-semantic-discourse categories of how discourse is construed in two reciprocal utterances. These thematic strategies (listed, explained and exemplified in Table 1 in Appendix) were distilled from the typology

[1] If we envision the semantic domain of a topic as a space, semantic continuity refers to the coherent navigation of meaning within that domain, with relatively small distances between semantic concepts for each transition. However, it should be acknowledged that the semantic continuity of a topic may display different characteristics depending on various factors (e.g., genre, mode of communication and environment). An illustration of semantic continuity in the context of the topic 'war' can be observed through the following sequence of transitions: *war* > *war strategies* > *weapons* > *troops and forces* > *impact on civilians* > *international response*. On the other hand, the following is an example of disrupted semantic continuity: *war* > *war strategies* > *war games* > *tabletop gaming* > *board games*.

[2] By 'logical transition' we mean a phase in the progression of discourse which logically follows from the previous one. For example: *He is an expert in this field* → *Therefore, we should trust his judgments*. Here, the conclusion regarding a person's trustworthiness is derived from the earlier statement about their professionalism.

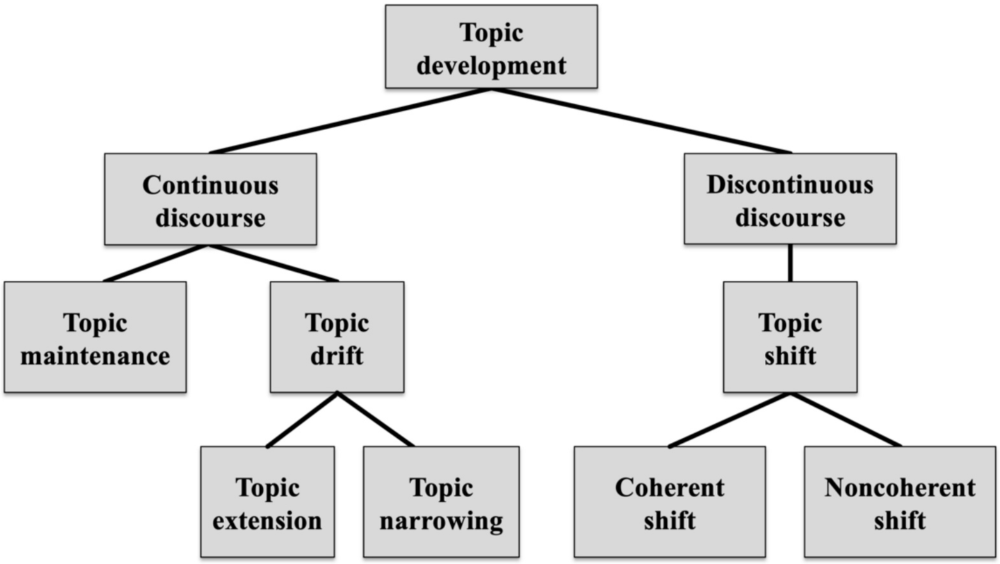

Fig. 1. Typology of topic development (adopted from Watson Todd, 2016: 71).

of metadiscourse resources (Hyland, 2005), which were reanalysed to form broader categories. They encompass the most prevalent methods for structuring information coherently on the higher levels of discourse. However, this list of thematic strategies is not exhaustive and can be further fine-grained.

Cohesion

Cohesion is a general quality of a text, representing how semantically consistent it is (Halliday and Hasan, 1976). It connects the meaning within a text and helps establish context and reflects the semantic relationships between a given item and others that come before or after it, through words or grammatical structures. Within this aspect, 11 thematic strategies[3] have been identified which describe the organisational structure of comments to enhance cohesion (see supporting materials). Specifically, 'Temporal exploration' examines temporal aspect of the topic (e.g., historical events, chronological order or time as a scientific concept); 'Spatial exploration' identifies the concepts relevant to a space, and 'Qualitative exploration' the qualitative discussion of discourse elements. Further, 'Didactic exploration' is a relationship enabling the commenter to derive a lesson from the topic.

Next, 'Endophoric reference' determines whether the discourse includes references to an entity within the semantic fields, established by the previous comment, and 'Self-reference' whether the commenter makes a reference to themselves. Moreover, 'Comparison and contrast' signals the presence of comparison, 'Consequence' includes a discussion of consequences and results, where 'consequence' refers to the causal outcome of an event. Furthermore, 'Hypothetical scenario' describes imagined, counterfactual or speculative situations; and if a reply engages with the grouping of information into specific categories or taxonomic structures, it goes under the thematic strategy of 'Categorisation'. Finally, 'Support' describes a topic development, where a commenter provides explicit support for their claim.

Attitude

Attitude reflects the emotional load of a reply, as mapped against the original comment. Seven thematic strategies were identified to evaluate users' attitude towards the topic. 'Sympathy and condolences' conveys a sense of compassion and support; and 'Wishes and wants' captures the articulation of desires, hopes, or aspirations. 'Irony' implies the expression of one's intended meaning through language which, taken literally, appears on the surface to express the opposite; 'Humour' is a strategy to classify comments that entertain and make others laugh. 'Hyperbole' captures exaggeration in a comment to emphasise certain aspects. Finally, 'Agreement' implies the acceptance of the viewpoints expressed by other discourse participants and 'Disagreement' characterises the discourse when a participant challenges the stance put forth in the main topic or by another user.

Logical quality

Logical quality refers to the degree to which a statement adheres to

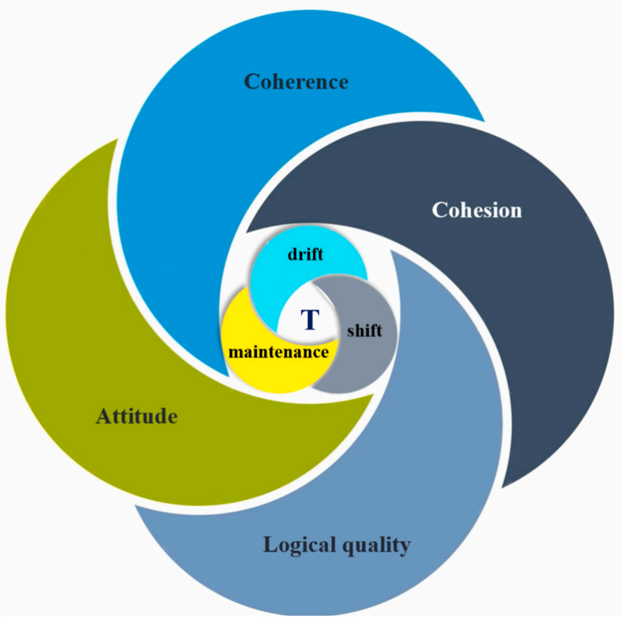

Fig. 2. Discourse aspects of topic development ('T' stands for 'topic').

[3] Taking inspiration from Hyland (2005), we asked ourselves what other metadiscourse, high-level features might be present in the text. Therefore, these 11 thematic strategies were empirically derived through a creative and inductive exploration of the comments in the YouTube comment section. While a larger set of thematic strategies was identified, these 11 emerged as the most frequent. A similar approach was used to identify thematic strategies for the other two aspects.

Table 1
Definitions of the human derailment rating.

Rating	Definition	Example
1	Discourse fully maintained: discourse is fully focused and cohesive, and strictly stays within the topic of the previous discourse.	**Comment A**: *Bananas are a rich source of potassium, and they're often used in fruit salads.* **Comment B**: *Definitely, they are great for potassium. I often add them to my own fruit salads.*
2	Discourse mostly maintained: discourse may have some minor deviations from the previously discussed topic and may introduce some new aspects, but does not elaborate on these aspects	**Comment A**: *Bananas are a rich source of potassium, and they're often used in fruit salads..* **Comment B**: *That's true, you can buy bananas cheaply at certain times of year.*
3	Topically balanced and neutral discourse: a balanced representation of topic maintenance and topic drift or further drift from the main topic, with greater elaboration on newly introduced aspects of the topic. The pure expression of attitude (e.g., sympathy, irony.) falls in this category. Frequently accompanied by smileys.	**Comment A**: *Bananas are a rich source of potassium, and they're often used in fruit salads..* **Comment B**: *Speaking of bananas, my cat goes crazy for them!*
4	Discourse partially derailed: The discourse is off track, but the gist of a main topic is maintained such that the discourse can be brought back.	***Comment A:*** *Bananas are a rich source of potassium, and they're often used in fruit salads.* ***Comment B:*** *The word 'potassium' originates from the Arabic word 'qali'.*
5	Discourse fully derailed: where an entirely new topic is introduced, and there is no relation to the original topic. This also includes completely random or nonsensical comments.	**Comment A**: *Bananas are a rich source of potassium, and they're often used in fruit salads.* **Comment B**: *The ice caps are melting at an alarming rate.*

the principles of logic.[4] Seven thematic strategies describe this category in association with the most common logical fallacies. 'Non-sequitur' describes a reply which does not logically follow from its previous discourse; and 'Red-herring' indicates a diversion of the topic to less significant aspects or tangential points. In contrast to the 'Straw Man' fallacy, which involves distorting an opponent's argument, a Red-Herring introduces an unrelated issue that diverts attention away from the actual topic of discussion (Tindale, 2007: 28). Then, 'Genre shift' represents a change of a genre of the comment, and 'Loaded question' implies a biased and controversial assumption. 'Ad-hominem' captures personal attacks against discourse participants, and 'Ad verecundiam' is an appeal to authority. Finally, 'False dilemma' features arguments presenting a premise that erroneously considers only two exclusive options.

Coherence as an inverse function of discourse derailment

Coherence is generally perceived as a "cognitive category that depends on the language user's interpretation and is not an invariant property of discourse or text" (Bublitz, 2011: 38). Thus, it is a measure that captures how a text is perceived as coherent.

Whereas the dimensions of cohesion, attitude and logical quality are measured objectively using thematic strategies, coherence is assessed subjectively, as the degree to which a pair of comments appears interconnected. Thus, we define coherence as how well a text maintains and develops a central topic, as perceived by a human annotator. It is not based on specific textual or linguistic clues, but rather on the human annotator's perception of topic development. In addition, when inverted, coherence can also be understood as the measure of discourse derailment, reflecting how closely a reply message adheres to or deviates from the topic established in the previous message. Thus, this measure is operationalised as 'human derailment rating' in the subsequent sections. The evaluation of the human derailment rating was performed using the Likert scale, ranging from 1 to 5, with 1 indicating fully maintained (coherent) discourse and 5 indicating fully derailed (incoherent) discourse. This scale is defined and exemplified in Table 1.

The units of discourse analysis and the principles of labelling

The unit of our discourse analysis is a pair of comments, and a thematic strategy captures the discourse landscape that emerges in the reply to this comment. We evaluate the discourse of the response to the comment in relation to the original comment, rather than assessing the response in isolation.

Given numerous layers of meaning in language, determining the presence of a specific thematic strategy in a pair of comments is challenging. Thus, the analysis was performed from the 'most obvious presence' principle: if a thematic strategy was deemed the most obvious, it was marked as present (1); otherwise, it was annotated as absent (0).

In this analysis, the second message in each pair was systematically coded for the presence of thematic strategies across three discourse dimensions. Each second message in the thread initiates the next pair, forming a continuous sequence that maps all messages. Crucially, each second message was evaluated in relation to, and within the context of, the preceding first message. This means that a single comment could exhibit multiple thematic strategies from various dimensions. Re-examining example (3), we observe how 'Comment B' is assessed against 'Comment A' using the established criteria of the coding scheme. Based on this framework, the following thematic strategies have been identified in 'Comment B' (each assigned a value of 1 in the coding scheme; all other strategies are coded as 0).

- **Cohesion**: 'Temporal exploration'—the comment explicitly mentions the timing for a specific action;'Didactic exploration'—the comment draws a moral conclusion and seeks to instruct the audience.
- **Attitude**: 'Irony'—the comment conveys a bitter tone regarding the state of the medical system, highlighting the irony in the notion that individuals fail to make 'right' choices.
- **Logical quality**: 'Non-sequitur'—the comment's argument does not logically follow from the preceding statement in Comment A; 'Red herring'—the comment shifts attention to a broader political issue, implying that systemic reform is a more urgent concern.

Therefore, there are two thematic strategies related to cohesion, one strategy pertaining to attitude and two thematic strategies associated with logical quality. This example shows that a single comment can encompass multiple thematic strategies.

2.3. A discourse space

Treating cohesion, attitude, logical quality and coherence as distinct dimensions enables the construction of a discourse space. In such a space, the first three dimensions serve as coordinate axes, while the fourth dimension represents vectors within this space (metaphorically, coherence can be envisaged as a 'time' vector within this space). The concept of *discourse space* emerges as the *conceptual* landscape within which discourse evolves, encompassing all conceivable states or conditions that discourse can occupy or transition between. Individual messages (i.e., the second message in a pair) occupy a position in this space and message threads constitute a path through the space.

In practical applications, mapping discourse spaces across different texts can yield valuable insights into the structural and stylistic properties that characterise various text types, registers and genres. In this study, we construct a discourse space specifically to investigate discourse derailment, a tactic frequently used in disinformation campaigns.

Malicious attempts to influence public perception can take many

[4] Logical principles are basic rules that govern consistent reasoning: e.g., the principle of non-contradiction (a contradictory statement cannot be true and false at the same time) and the principle of sufficient reason (everything has a reason and cause).

forms, relying on a range of communicative strategies such as misinformation, disinformation and fake news. These efforts may also include selective framing (presenting a news item in a deliberately biased light), demagoguery (appeals to prejudice and emotion), or distraction tactics that shift attention away from key issues. It is crucial to recognise that all these efforts are carried out through language in context—that is, through discourse.

By applying the concept of discourse space, this study aims to enhance our understanding of the discourse strategies underlying discourse derailment. Specifically, we examine how these strategies manifest within the discourse space and how they correlate with the derailed discourse. This approach allows us to move beyond surface-level linguistic features and engage with the broader contextual dimensions of language. In addition, the concept of discourse space. Exploring the discourse space also offers insight into the process of mediation, revealing platform-specific metadiscourse patterns that emerge on YouTube and how topic development in the comment section is shaped by the platform's unique affordances and interactional dynamics. In the following subsection, we conceptualise the discourse space framework through mathematical formalisation.

2.3.1. Discourse space, topic development, discourse (dis)continuity and discourse derailment: Formalisation of the relationships

The relations between the above-discussed concepts can be formalised as follows. Let $D \subset \mathbb{R}^3$ be the discourse space, which is defined as a three-dimensional Euclidean coordinate system (Corral, 2023: 1), with axes corresponding to the discourse dimensions discussed above: x represents 'Cohesion'; y 'Attitude' and $z' = \frac{1}{z}$ 'Logical quality'. Each message m_i is represented as a data point $D_i \in D$ in this discourse space such that

$$m_i \mapsto D_i = \left(x_i, y_i, z_i'\right) \in D$$

where $x_i \subset \mathbb{R}^3$ denotes a degree of cohesion in each message m_i; $y_i \subset \mathbb{R}^3$ reflects a degree of attitude; and $z_i' \subset \mathbb{R}^3$ represents a logical quality of a given message.

Note that for the x and y axes ('Cohesion' and 'Attitude'), larger values indicate stronger presence of the given property, whereas, for the $z' = \frac{1}{z}$ axis ('Logical quality'), larger values indicate lower logical quality. This asymmetry arises from the nature of logical quality: sound reasoning is typically exemplified by a single principle, the *ad rem* argument (i.e., 'to the point'),[5] whereas there exists a wide variety of logical fallacies that can undermine the logic of an argument. Accordingly, the axis labels x ,y and $z' = \frac{1}{z}$ should be understood as value-neutral categories (that is to say, they simply label the dimensions: for example, the third dimension is referred to as 'logical quality' rather than 'logical fallacy', the latter of which would imply a negative connotation for that coordinate). These coordinates define a locally three-dimensional space and reflect the degree of expression of discourse features. The 'Cohesion' and 'Attitude' axes represent a continuum from weak to strong expression, whereas the 'Logical quality' axis increases from logical coherence (high-quality logic) to logical incoherence (low-quality logic).

Topic development can then be represented as the vector between two messages m_i and m_j (Corral, 2023: 6):

$$\overrightarrow{v}_{ij} = \left(D_j^{(x)} - D_i^{(x)}; D_j^{(y)} - D_i^{(y)}; D_j^{(z)} - D_i^{(z)}\right)$$

This vector has two key parameters: (i) discourse distance and (ii) discourse directionality. *Discourse distance* within the discourse space corresponds to the perceived measure of coherence as described in subsection 2.2. Formally, discourse distance between two messages is defined as follows:

$$\text{Discourse distance } (m_i,\ m_j) =. \left\| \overrightarrow{v}_{ij} \right\| = \left\| \left(D_j^{(x)} - D_i^{(x)}; D_j^{(y)} - D_i^{(y)}; D_j^{(z)} - D_i^{(z)}\right) \right\|$$

Discourse directionality, in turn, shows which discourse features are changing and how. It is captured by the orientation of the vector $\overrightarrow{v}_{ij}$ in the discourse space and its angle. The orientation of the vector $\overrightarrow{v}_{ij}$ is defined relative to the axes and determines whether the topic is shifting toward lower or higher cohesion (the x axis), weaker or stronger expression of attitude (the y axis) or better or worse logical quality (the z' axis). The orientation can be evaluated by the angle θ between the vector $\overrightarrow{v}_{ij}$ and coordinate axes (i.e., x, y and z'). For example, a cosine of the angle between the vector $\overrightarrow{v}_{ij}$ and the unit vector $\widehat{x}$ along the x-axis can be calculated as follows:

$$\text{Discourse directionality (relative to } x - \text{axis)} = \cos(\theta_x) = \frac{\overrightarrow{v}_{ij} \bullet \widehat{x}}{\left\| \overrightarrow{v}_{ij} \right\|}$$

where $\overrightarrow{v}_{ij}$ is the topic development vector from message i to message j; $\widehat{x}$ is the unit vector in the direction of the x-axis; $\overrightarrow{v}_{ij} \bullet \widehat{x}$ is the dot product which measures a degree of alignment of the vector $\overrightarrow{v}_{ij}$ with the x-axis; and $\left\| \overrightarrow{v}_{ij} \right\|$ is the length of the vector.

Since the axes in the discourse space represent abstract discourse dimensions, each region of the space can be associated with specific discourse characteristics. As demonstrated later in this paper, certain subspaces within the discourse space correlate strongly with certain discourse strategies, including disinformation and bad-faith rhetorical practices. Depending on the nature of the texts analysed, specific subspaces within the discourse space may also correspond to distinct genres or registers.

There are various ways to advance a topic. Some thematic strategies, introduced above, keep the topic on track, such as agreement and encouragement, and some are less cooperative such as ad hominem or red-herring. A single message might combine several thematic strategies, and it is likely that some strategies appear together more often than others, creating clusters of frequently observed points in the total possible space. This suggests that continuous and discontinuous discourse form distinct patterns within this space.

Therefore, we define *topic continuity* either as a single vector $\overrightarrow{v}_{ij}$ between two discourse units (e.g., messages, sentences, paragraphs, etc.), or as a sequence (i.e., sum) of vectors $\sum \overrightarrow{v}_{k(k+1)}$ that represent the progression of discourse across multiple turns. In both cases, topic continuity is characterised by relatively short discourse distances and by the positioning of subsequent discourse units within regions that exhibit similar discourse features to those observed in the area of the initiating discourse unit D_i.

In contrast, *topic discontinuity* emerges when a single vector $\overrightarrow{v}_{ij}$ or as a sequence (i.e., sum) of vectors $\sum \overrightarrow{v}_{k(k+1)}$ show a large discourse distance and a sharp angular deviation from the initiating discourse unit D_i. *Discourse derailment* is then a specific type of topic discontinuity. It occurs when a discourse unit D_j that moves a discourse thread forward occurs in a region associated with disinformation or bad-faith rhetoric practices—i.e., $D_j \in \mathscr{R}_{disinfo}$—and/or when the length of the vector $\left\| \overrightarrow{v}_{ij} \right\|$ is large, and/or the angle of deviation θ is high.

In this study, we demonstrate that the discourse space D can be segmented into regions reflecting different discourse strategies, which,

[5] Theoretically, there is potential to develop an approach that aligns the 'Logical quality' axis more closely with the other two dimensions. This could be achieved, for instance, by expanding the representation of high-quality reasoning to include a broader range of argument structures, such as Toulmin's (2003) model of argumentation or the framework of informal logic proposed by Johnson and Blair (2002). However, pursuing this alignment would require a substantial conceptual and methodological effort, which falls outside the scope of the present study.

in turn, may align with identifiable genres, registers or rhetorical intents. In a companion paper (Authors, 2024), we examine the extent to which discourse space can be approximated by a semantic space constructed from word embeddings generated by large language models (LLMs), using cosine similarity to measure semantic distances between messages. Comparing semantic and discourse spaces offers a promising avenue for assessing the extent to which LLMs capture and understand discourse structure.

2.3.2. *Discourse space as a manifold*

With the formalisation we have developed so far, we now take a step further by proposing that the discourse space can be conceptualised as a manifold. A discourse space emerges as a dynamic system of possible discourse states, evolving over time according to sequences of thematic strategies (e.g., a loaded question followed by an ad hominem or disagreement, but unlikely followed by encouragement). Furthermore, thematic strategies exhibit variable topological influence: some operate globally across the discourse (e.g., qualitative exploration), while others affect only local regions (e.g., sympathy). Consequently, we hypothesise that the structure of the discourse space can be understood as a manifold: locally Euclidean and smooth (as suggested by our Euclidean coordinate formalisation in the previous subsection), yet globally complex and potentially non-linear.

Let $D\subset\mathbb{R}^3$ denote the set of data points, where each point represents discourse information of a message. The discourse space can be formally defined as a manifold $\mathscr{M}\subset\mathbb{R}^3$, if the following condition applies: for every $D_i \in D$, there exists an open neighbourhood[6] $U\subset D$ that contains D_i and homeomorphism[7] $\varphi : U\rightarrow\mathbb{R}^k$ (Pinchuk et al., 2023: 5–6), such that the change of coordinates between overlapping regions of the space is gradual and differentiable. This makes the discourse space amenable to analysis using manifold learning techniques and differential geometry (Munkres, 2018; for application in linguistic fields, see e.g. Dinnage, 2023).

In these sections, we have outlined the foundational structure of the discourse space framework. However, we acknowledge that future research will refine and expand this framework, adding greater detail and complexity.

3. Data collection and coding

Validating the method requires a test dataset. In the linguistic study of discourse, although there is a wealth of corpora available for extracting semantic and syntactic information from texts, only a scant few resources are equipped with an integrated discourse tag-set. For example, one of these few corpora is the Switchboard Dialog Act Corpus which is tagged with a shallow discourse tag-set of approximately 60 basic dialog act tags and combinations ("SWBD-DAMSL" labels; Godfrey and Holliman, 1993), which incorporates both traditional sociolinguistic and discourse-theoretic rhetorical relations/adjacency-pairs, as well as some more form-based models. However, this corpus was developed for spoken discourse. Thus, it does not provide measures of cohesion across the course of a discussion.

In this study, we chose to focus on YouTube comments. YouTube is one of the most influential social media platforms today and the locus of much online discourse participation via its text comments section which represents a complex multimodal text (Benson, 2016). Because of YouTube's widespread popularity, ease of accessibility and ability to amplify comments by attaching them to popular videos, it offers a fertile ground for actors to influence individual's perceptions and beliefs. Thus, YouTube, as a powerful communication medium, plays a central role in shaping contemporary online discourse. Previous studies looked at various aspects of YouTube comment discourse, such as conflict management (Bou-Franch & Blitvich, 2014), polarisation (Gupta et al., 2023), hate speech (Döring and Mohseni, 2020) and hope speech (Chakravarthi, 2022). YouTube comments sections have also been identified as the locus of malicious disinformation campaigns (Hussein et al., 2020; Golovchenko et al., 2020; Shao et al., 2018). Therefore, it provides an opportunity to gain insights into the malicious and harmful discourse derailment.

News channels in particular serve as hotbeds for intense discussions (e.g. Inwood and Zappavigna's (2023) study of YouTube comments on RT (formerly Russia Today) about the Skripal poisoning). Hence, this study focuses on comments from videos on the YouTube channel of BBC World News, one of the most influential and trustworthy news organisations (Benton, 2018).

This section describes the collection and human tagging of the dataset. Messages were collected from the comment sections of BBC News YouTube videos and then tagged by a human for a range of thematic strategies. The following subsections give detailed accounts of these steps. Subsection 3.1 elucidates the criteria and procedures for which YouTube video comments were sampled. Subsection 3.2 explains the structure of YouTube comment data, and subsection 3.3 the adopted perspective on discourse during data tagging. Finally, subsection 3.4 details the process of the human tagging of the data and reliability.

3.1. *Video sampling*

Videos from the BBC News YouTube channel from 2023 were sampled using a constructed week sampling method (Luke et al., 2011). This was done by selecting days diagonally from each day of the working week (i.e., Monday from week 1, Tuesday from week 2, Wednesday from week 3, etc.), starting in the first week of January 2023.

For each date, we identified all videos uploaded on this date that were examples of 'hard news'—time-constrained political events and the societal consequences reported in a thematic, impersonal style (Reinemann et al., 2012). The total number of comments and the number of replies to the first 10 top level comments (excluding non-English comments) were recorded. For each date, a video was sampled at random that had at least 5 top level comments with 5 replies.

The comments for 50 videos were obtained using the YouTube API, taking the top 7 top-level comments and up to 7 replies to each top-level comment. Data from the API is ordered according to the order user would see them at the point of data collection. This order is determined by a closed-source algorithm but depends on a combination of the number of 'likes', the date posted and the number of replies. For our purposes, taking the initial messages that a user would see seems like a representative sample of posts that might actually be read, those being the most likely to influence users. Data was anonymised on download.

Non-English messages were excluded automatically using the lingua language detector (https://pypi.org/project/lingua-language-detector/) because discourse strategies might vary between languages. Message pairs were automatically extracted and copied into an Excel spreadsheet for manual data tagging.

3.2. *Comments and their replies: Types of message pairs*

The units of the analysis in this research are pairs of YouTube messages. YouTube users can either post a 'top-level comment' as a response to a video, or a 'reply' to a top-level comment. Unlike some other social media sites, there is currently only one level of embedding allowed in YouTube comment sections. While there is a button to reply to a 'reply', this only creates a new reply to the top-level comment and inserts the first author's name to the suggested text of the second reply (e.g., "@username").

[6] In simple terms, an open neighbourhood refers to a nearby region around a point that behaves like a regular space.

[7] A homeomorphism is a way to stretch or bend one shape into another without cutting or gluing, as long as both shapes are the same dimension. A classic example is how a mug can be reshaped into a donut: they are topologically the same.

To capture the textual discourse within a specific video, we have categorised the messages into a pair structure shown in Fig. 3. Each top-level comment was paired with the video title (as a proxy for the video content). Each reply was paired with the relevant previous message within its top-level thread: either the message immediately before it or, if the message referenced a commenter handle, the most recent message by that commenter within the thread. In this structure, three types of message pairs can be distinguished. In the first pair type ('*Pair 1*' and '*Pair 5*' in Fig. 3 and labelled as 'videoLevel' in the dataset), the first message represents the discourse in the video using the title of the video and the second message is a 'top-level' comment on this video. The second type ('*Pair 2*' and '*Pair 6*', labelled as 'topLevel'), includes a top-level comment from the first pair and the first reply to this comment. Finally, the third type ('*Pair 3*' and '*Pair 4*', labelled as 'reply') includes two sequential replies. By adopting this structure, we aimed to establish how the discourse changes from message to message.

Usernames were anonymised on download, meaning that no usernames were stored at rest. User metadata was removed and usernames in the metadata and usernames within comment texts were replaced with irreversibly hashed pseudonyms. To facilitate human understanding of the relations between users, hashing was done using the Python library 'humanhash3', which creates human-readable hashes which combine four arbitrary words (e.g., "friend-april-winner-two").

3.3. Perspective

The aim is to tag the discourse from the perspective of the average user of YouTube. We assume that users are familiar with Internet culture and conventions. Some discourse can look very different from this perspective, especially when users make analogies between the current discourse and popular culture references. Take Example 4, from a video about Al Gore making predictions about climate change.

(4).

Comment 1: *Every prediction he's ever made has been wrong.*

Comment 2: *He was right about manbearpig though.*

The second comment seems very distant from the first, except if you know that the user is referencing an episode of the cult satire *South Park* where Al Gore keeps making predictions about a mythical creature, 'manbearpig', which nobody believes but turns out to be right. This is the source for many Internet memes. Thus, the comment is in fact cohesive and might represent topic extension, if one is familiar with Internet culture. We aim to capture this kind of information precisely because it might be the kind of relationship that an automatic system cannot easily detect.

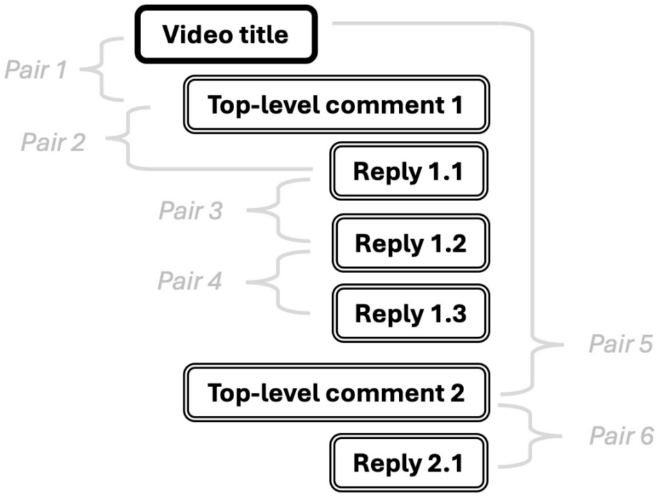

Fig. 3. The structure of message pairs.

3.4. Human data tagging and reliability

The data tagging proceeded in three phases. In the first phase, three researchers independently tagged a small dataset for 40 binary variables representing the presence or absence of different discourse strategies. For each variable, any disagreements and/or ambiguities were identified and discussed, which led to the update of the thematic strategies. In the second phase, three researchers independently tagged a new dataset of data from 200 message pairs from 10 videos. Formal reliability tests were carried out, using several metrics. The standard reliability metric in linguistics is Cohen's kappa, which effectively computes the extent to which the agreement is above chance as a proportion of the distance between chance levels and perfect agreement (e.g., if chance levels were 50 %, and coders were showing 60 % agreement, this would be 10 points above chance, or 0.2 or 20 % of the way from chance to perfect). This metric depends on various aspects such as the number of datapoints and the frequency of each type. Ultimately, it produces a number that ranges from 0 to 1, where values above 0.4 are widely considered good agreement in linguistics and psychology (e.g., Landis & Koch, 1977). When comparing the agreement of three taggers, the intraclass correlation coefficient (ICC) was used, which is a generalisation of the principles of the kappa statistics. Overall, the reliability improved by 28 % between phase one and two.

The results of the second phase revealed some variables that were either exceedingly rare to be useful in the main analysis, had low reliability, or had good reliability but were redundant given other variables. These were removed to leave 25 binary variables representing thematic strategies and one ordinal variable representing the coherence rate. The average agreement between taggers for these variables was 88 %. All variables had ICC values significantly above chance, and 14 variables had ICC above 0.4 (see Supporting Information 2). While this is not ideal, given the inherent subjectivity of variables like 'humour', achieving perfect agreement is not realistically anticipated. Nevertheless, to improve the coding further, in a third phase, the disagreements between taggers from phase 2 were reviewed by the taggers and small revisions were made to the coding scheme, accordingly. The final dataset was coded by one human tagger, providing internal consistency.

Tagging was completed for 50 videos. There were some replies where the original message had been deleted (by the author or by YouTube employees for unknown reasons) between the reply being sent and the data being collected (less than 1 % of the data). These replies were not tagged by the tagger. One message was in a language other than English (and not detected by the automatic language detector) and this was also removed.

The final human-tagged dataset included 1,684 message pairs from 50 videos tagged for 25 thematic strategies and one coherence rate variable.

4. Methodology

To model a discourse space from our dataset, we used Principal Component Analysis (PCA). PCA attempts to represent the variation between multivariate observations using a smaller number of dimensions called principal components. In this study, PCA first measures the Euclidean distance between thematic strategies of cohesion, attitude and logic, and, secondly, examines correlations among them. The former is visualised through a distance biplot of data points scattered across the first two principal components (2D biplot). The latter is illustrated through a biplot of a correlation circle, with vector arrows representing each variable. The PCA analysis was implemented in R (R Core Team, 2022), using packages 'FactorMineR' (Lê et al., 2008), 'ggplot2' (Wickham, 2016) and 'factoextra' (Kassambara & Mundt, 2020).

We also used t-distributed stochastic neighbour embedding (t-SNE; e.g., van der Maaten and Hinton, 2008) to visualise clusters of datapoints. It is considered a superior approach for capturing non-linear structures within high-dimensional data. The method of t-SNE prioritises two key aspects: (i) modelling dissimilar datapoints by means of large pairwise distances, and (ii) modelling similar datapoints by means of small pairwise distances (van der Maaten & Hinton, 2008: 2587). This method was applied using the 'Rtsne' package (Krijthe, 2015).

5. Analyses and results

In this section, we present the analyses of our YouTube dataset. Subsection 5.1 outlines the results of the PCA and t-SNE analyses; and subsection 5.2 considers them in a wider context. Finally, subsection 5.3 discusses the inherent limitations of this study.

5.1. Clustering analyses

The PCA model fitted to the dataset of the human-tagged thematic strategies identifies general trends in the relationship between them and coherence rate. Its first two dimensions that maximise the difference between message pairs based on their discourse strategies are visualised in Fig. 4. These first two PCA dimensions account for 17.3 % of the variability observed in our dataset. The points in the plot are colour-coded according to the measure of human-tagged discourse derailment—i.e., inverse cohesion. It is important to note that the discourse derailment measure was solely used to colour the datapoints and was not included in the actual model. The arrows representing the variables (i.e., thematic strategies) demonstrate the influence of the data points within the discourse space. The biplot shows that the derailing messages cluster together within the discourse space.

Non-derailing messages mostly group together due to the presence of endophoric reference, self-reference, qualitative exploration and sympathy (i.e., these variables have the longest arrows indicating a greater pull). In contrast, messages with high derailment cluster around the presence of variables such as non-sequitur, red-herring, hyperbole, ad hominem and humour. This biplot suggests regions in the discourse space correlated with derailment. This suggests that there is systematic discourse structure in the data.

The biplot in Fig. 4 also enables the identification of relationships between thematic strategies, particularly in terms of their positive or negative correlations. A small angle between vectors indicates a strong positive correlation; an angle near 90 degrees suggests little to no correlation; and an angle greater than 90 degrees (approaching 180 degrees) reflects a negative correlation between variables. Four general patterns of positive correlations among thematic strategies can be observed in this biplot, indicating that the strategies within each group tend to co-occur more frequently: (i) agreement, disagreement, sympathy, wish, self-reference and endophoric reference; (ii) self-reference, endophoric reference, consequence, comparison, spatial exploration, qualitative exploration, support, categorisation, temporal exploration and hypothetical scenario; (iii) didactic exploration, hyperbole, false dilemma, irony and red herring; and (iv) loaded question, humour, non-sequitur and ad hominem. In contrast, negatively correlated strategies—those that tend not to co-occur—include, for example: agreement and red herring; endophoric reference and non-sequitur; sympathy and non-sequitur; humour and wish; agreement and ad hominem; as well as agreement and didactic exploration. However, the PCA model provides a coarse approximation of the datapoints. The method of t-SNE allows us to capture more subtle relationships between variables. Fig. 5 (a) and (b) illustrate the clusters with enhanced clarity. The plots serve as a visual map of the various combinations of thematic strategies (the model's input variables) and their relationship to discourse derailment (represented by the colour gradient of the data points). More broadly, they offer a representation of YouTube-mediated discourse dynamics.

The patterns emerging from the t-SNE analysis are captured in Table 2. The two most frequently used strategies—qualitative exploration and endophoric reference—often co-occur with a range of other thematic strategies and are strongly associated with messages that remain on topic. This is expected, as coherent engagement typically involves referencing previously discussed concepts and elaborating on their qualitative dimensions. Such patterns suggest that when a message explores the qualities of a topic, it is less likely to contribute to discourse derailment. More unexpectedly, however, endophoric reference

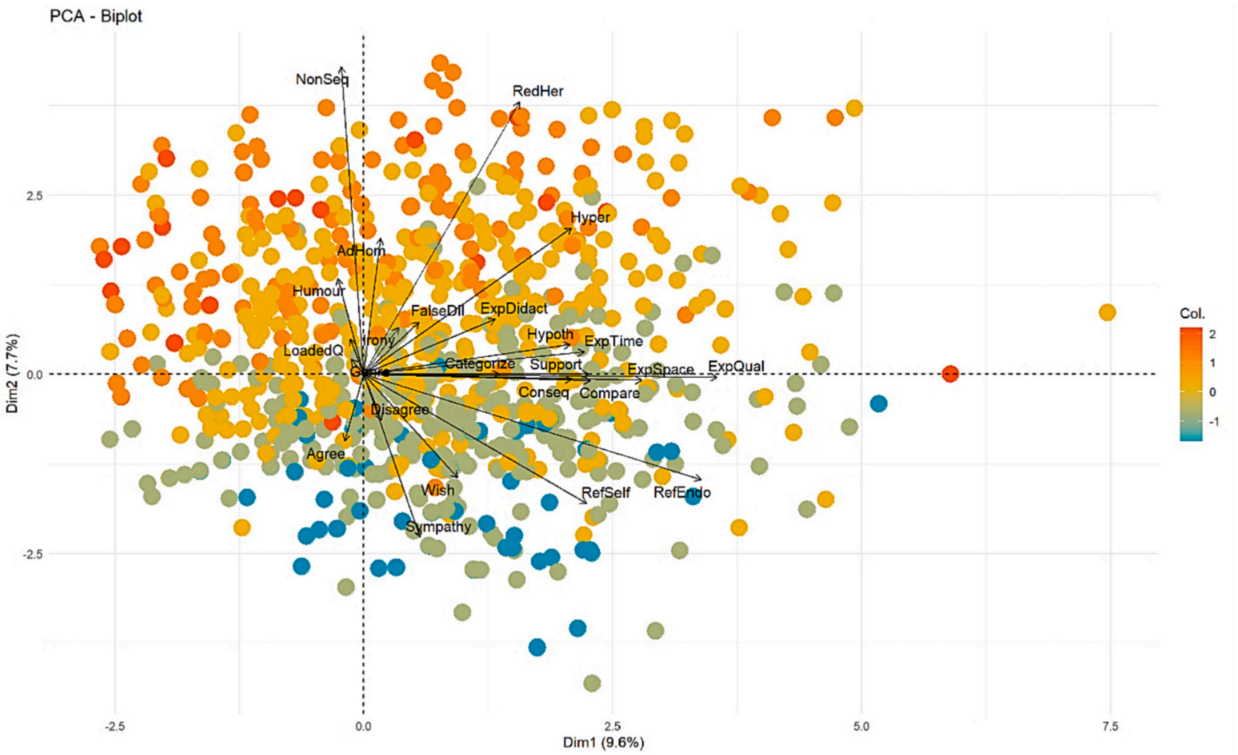

Fig. 4. The biplot of the first two principal components based on thematic strategies. Warmer colours indicate higher levels of discourse derailment. Arrows represent thematic strategies.

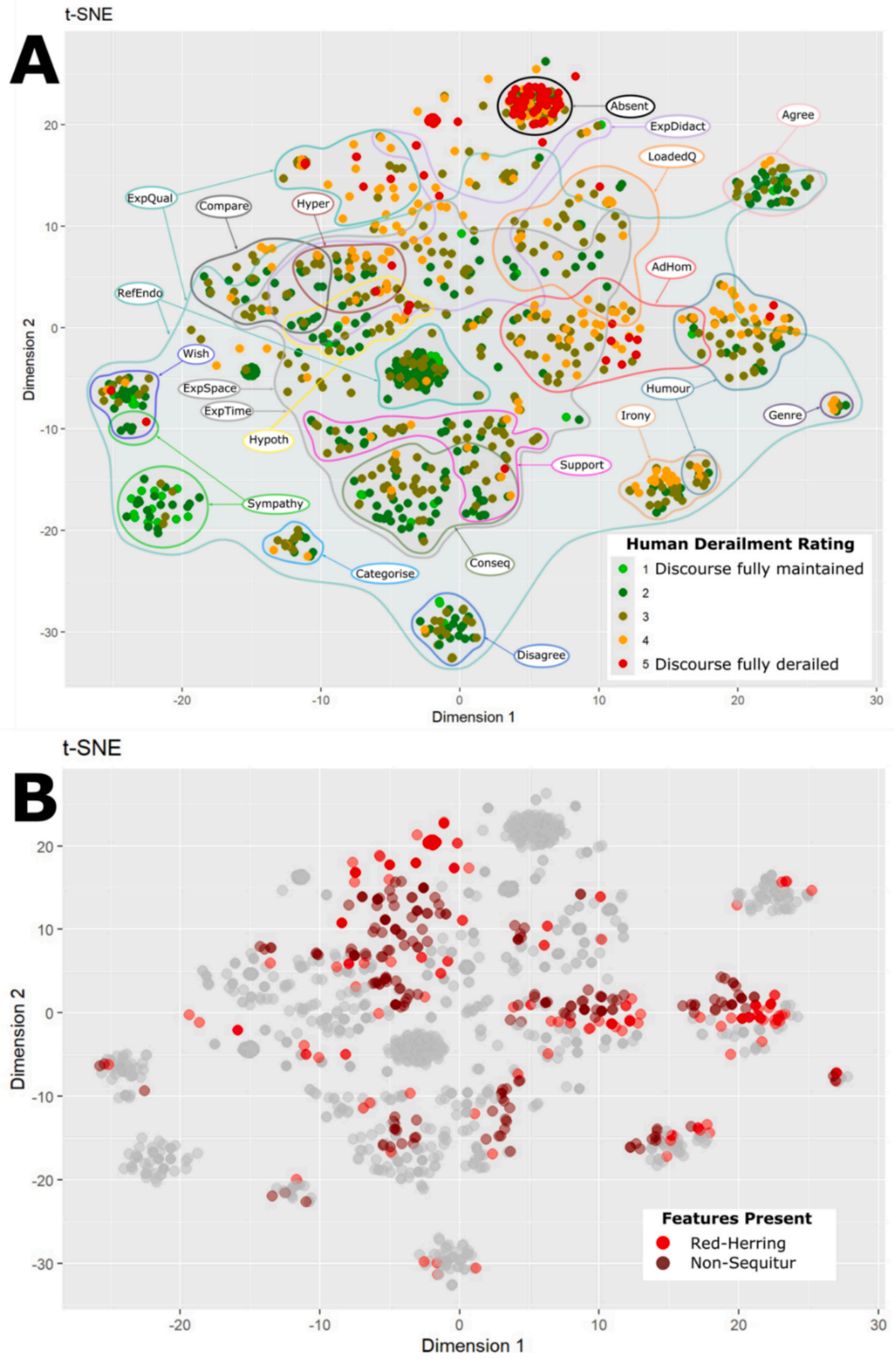

Fig. 5. The t-SNE plot of thematic strategies: (a) coloured by human derailment ratings, with clusters approximately highlighted by encircled lines; (b) coloured by the thematic strategies of red-herring (light red) and non-sequitur (dark red), two prominent derailing strategies. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

occasionally appears alongside strategies typically linked to discourse derailment, such as red herring, non-sequitur and ad hominem. Likewise, qualitative exploration sometimes co-occurs with derailment strategies like loaded question, didactic exploration, ad hominem and humour. This indicates that even in derailed discourse, a superficial linguistic link to the main topic may be preserved. Through the use of endophoric reference and qualitative exploration, such messages can create the illusion of coherence, maintaining a surface-level connection to the discourse thread despite a shift away from its substance.

Exploration of time and space appears across all core clusters in the plot, with spatial exploration occurring more frequently. These strategies are most commonly associated with hypothetical scenarios, support and consequence. This suggests that when individuals engage with hypothetical situations, they naturally need to situate these events within specific temporal and spatial contexts. Similarly, providing support for an argument or discussing consequences often involves anchoring ideas in time and space. In a less expected pattern, spatial exploration also features prominently in clusters related to sympathy and categorisation. This pattern implies that sympathy and categorisation may be more grounded in spatial cognitive frames or shaped by metaphorical spatial representations.

In contrast, messages involving agreement and disagreement show relatively little use of temporal or spatial exploration. Arguably, this may be seen as indirect evidence of the need for cognitive closure (Webster & Kruglanski, 1994), which is the human tendency to seek quick decisions, resolve uncertainty or impose structure on beliefs. When a person expresses agreement or disagreement, it suggests they hold a firm opinion and have reached a sense of closure, leaving little room for considerations of *when* or *where* something holds true.

Despite their clusters being similar in size, agreement and disagreement are far apart in the discourse space—with the agreement cluster, surprisingly, located closer to derailment-related clusters. This suggests that disagreement is not necessarily the prevalent strategy in discourse derailment (only 3.3 % of messages with disagreement have high derailment scores). Whereas both share several strategies (e.g., spatial and qualitative exploration, support and humour), disagreement uniquely includes temporal, consequential and sympathetic aspects. In contrast, agreement more often co-occurs with self-reference and hypothetical scenarios. This suggests that agreement and disagreement play different roles in shaping discourse space.

The sympathy cluster emerges as one of the most non-derailing clusters. It occasionally includes strategies like didactic exploration, comparison, consequence, hyperbole and even red herring or loaded questions, but these are relatively rare. In particular, the overlap with the wish thematic strategy increases the likelihood of derailing discourse.

Perhaps less intuitively, the wish cluster incorporates nearly all thematic strategies (except genre shift), including those more closely tied to derailment. Specifically, its overlap with hypothetical scenarios, when combined with hyperbole, red herring and non-sequitur, often contribute to derailing the discourse.

Categorisation and genre shift form isolated clusters. The categorisation cluster sits near sympathy, consequence and disagreement, suggesting it tends to appear in those discourse contexts. The cluster incorporates nearly all thematic strategies (except genre shift) and is often paired with irony. This pairing hints that irony may be used to challenge categorisation, creating a space where derailment can take hold. The genre shift cluster, meanwhile, is located between humour and irony, pointing to a functional similarity. Unlike categorisation, genre shift is closely linked with discourse derailment. It frequently combines with strategies like temporal, qualitative and didactic exploration, comparison, hypothetical scenarios, humour, and endophoric and self-reference. It also serves as a key hub for the classic markers of derailed discourse—red herrings and non-sequiturs.

The consequence and support clusters often overlap, reflecting a common tendency among YouTube users to strengthen arguments by emphasising causal outcomes—a combination that is largely non-derailing. Similarly, discourse tends to stay on track when comparison, hypothetical scenarios and didactic exploration intersect. However, when red herring or non-sequitur strategies enter these overlaps, derailment becomes more likely. The highest levels of derailment occur in intersections that combine hyperbole, hypothetical scenarios, didactic exploration, red herring and non-sequitur—or, in some cases, false dilemma instead of hyperbole. More broadly, derailed discourse—whether logically structured or not—often involves imaginary or mutually exclusive scenarios, presented in a didactic and exaggerated tone.

The most derailing clusters are those centred on ad hominem and loaded questions in Fig. 5(a), and red herring and non-sequitur in Fig. 5 (b). Interestingly, loaded questions show a higher tendency to derail the discourse—especially when paired with ad hominem attacks or when lacking qualitative exploration of the topic. In contrast, when loaded questions are framed with a didactic tone and engage with the topic's qualitative aspects, they no longer signal derailment.

Finally, humour and irony, fascinating subjects of study in their own right, can play very different roles in discourse. Their impact depends on the company they keep: when paired with strategies like red herring or non-sequitur, they tend to signal derailment. However, on their own, they do not necessarily derail discourse.

Thus, the main insight gleaned from this t-SNE analysis is that discourse derailment is multifaceted and that an approach that considers strategies in isolation may miss important aspects. Most messages exhibit multiple thematic strategies, and the level of discourse derailment may be predicted by specific combinations of thematic strategies rather than the presence of specific individual strategies. Additionally, the highest category of derailment often has an **absence** of thematic

Table 2
A typology of common combinations of strategies in the derailing and non-derailing discourse.

Discourse	Combination of strategies
Derailing	The absence of all 25 strategies: this is perhaps the area of the random shift of the topic
Derailing	Ad hominem in the absence of the exploration of time and space, and endophoric reference
Derailing	Genre shift, red-herring and non-sequitur
Derailing	Comparison, hypothetical scenario, didactic exploration, red-herring and non-sequitur
Derailing	Hyperbole, hypothetical scenario, didactic exploration, red-herring and non-sequitur
Derailing	Hypothetical scenario, didactic exploration, false dilemma, red-herring and non-sequitur
Derailing	Ad hominem, loaded question in the absence of qualitative exploration
Derailing	Support and red-herring
Derailing	Consequence and non-sequitur
Derailing	Irony, red-herring and non-sequitur
Derailing	Humour, red-herring and non-sequitur in the absence of endophoric reference
Partially derailing	Humour, ad hominem, exploration of space and time and qualitative exploration
Partially derailing	Sympathy and wish
Partially derailing	Wish, hypothetical scenario, hyperbole, red-herring and non-sequitur
Partially derailing	Categorisation and irony
Partially derailing	Support and consequence
Partially derailing	Disagreement and red-herring
Partially derailing	Ad hominem and loaded question in the presence of qualitative exploration
Non-derailing	Comparison, hypothetical scenario and didactic exploration
Non-derailing	Sympathy, qualitative exploration and endophoric reference
Non-derailing	Endophoric reference and qualitative exploration
Non-derailing	Humour, qualitative exploration and agreement

strategies, as evident from the one of the densest clusters in the plot at the top.

5.2. Discussion

The objective of this article has been to present a new methodology that facilitates the mapping of discourse features at a metadiscourse level. We demonstrated that discourse information, contained in pairs of messages, can be mapped onto a discourse space. This consistent mapping allows us to explore how different features relate to one another and identify regions within the space associated with discontinuous discourse—regions that include messages which could be categorised as part of disinformation efforts.

More specifically, we showed that a discourse space can be modelled using thematic strategies. Certain regions within this space are associated with reduced cohesion which corresponds to instances of discourse derailment. These regions are characterised by the **presence or/and absence** of specific discourse features, and to the best of our knowledge, our study is the first to bring this fact to light. For example, the absence of endophoric reference and the presence of non-sequitur and red-herring are the most significant indicators of the derailing discourse. Similarly, the absence of sympathy or empathy in a message increases the likelihood of derailed discourse, or if sympathy is combined with the expression of wish. The derailing discourse is also linked to other attitudinal variables such as 'humour', 'irony' and 'hyperbole', alongside such cohesion variables as 'didactic exploration', 'comparison', 'hypothetical scenario', 'consequence' and 'support'.

Crucially, the complete absence of the discourse features can act as a filter for identifying random messages. We observed various types of derailing messages, noting that those disrupting discourse in a meaningful manner tend to fall within a coherence rate range of above 3 and below 5. On the other hand, many messages with the highest derailing score (i.e., 5) are random and devoid of context, suggesting random discourse shift and no intent to foster meaningful discussion. Therefore, it is advisable that such messages do not become the primary focus of the detection system, and they should be filtered out. Because these random messages are characterised by the complete absence of thematic strategies, they can be easily identified during the early stages of message processing.

Further, our study highlights the vital role of top-down (contextual) mechanisms in shaping discourse space, indicating their complex nature that makes them less susceptible to conforming to easily identifiable patterns. The agreement and disagreement thematic strategies from our analysis illustrate this point. Despite some similarities, the agreement and disagreement strategies show different dynamics: within derailing discourse, agreement often intertwines with red-herring and loaded questions, while disagreement frequently accompanies ad-hominem attacks. Thus, derailing discourse constitutes a complex phenomenon, and the concept of discourse derailment appears adept at capturing nuanced relations between the elements of discourse.

In light of YouTube's mediating effect on online discourse, it can be inferred that the platform fosters a highly dynamic communicative environment, characterised by the expression of multiple thematic strategies within brief, often isolated, messages. Due to the typically short nature of YouTube comments, users tend to distil their thoughts to the essential points they wish to convey. The discussed combinations of thematic strategies thus reflect how users articulate their perspectives within the constraints of the medium. This results in a notable *condensation* and *intensification* of discourse strategies over a limited textual span—features less characteristic of academic or literary writing, where ideas are typically developed more coherently and extensively over longer stretches of text.

Our analysis reveals that nearly half of the discourse space is occupied by derailing or partially derailing comments. This high incidence of topic discontinuity appears to be shaped by two primary factors. First, many users leverage the YouTube comment section space to advance personal or others' agendas—an aspect of YouTube discourse that warrants closer attention in the context of disinformation studies. Second, the platform's inherently low level of reciprocity (Wattenhofer et al., 2021: 357), contributes significantly to the fragmentation of dialogue. Conversations are often temporally disjointed and unreciprocated; users seldom attempt to engage with others or establish coherent conversational threads. Instead, commenting frequently serves as a self-contained act of expression or reaction, directed more toward the video content or other users' comments than toward sustained dialogue. This behaviour positions the YouTube comment section more as a space for individual self-expression than as a site of interactive conversation.

A further contributing factor to the prevalence of derailing discourse may be the heightened level of viewer engagement observed on politically oriented YouTube channels (Heydari et al. 2019). The emotionally charged nature of political topics often prompts users to express their beliefs with urgency. The affordances of anonymity (Brown, 2018) and the expectation of non-reciprocal interaction may, in turn, encourage more direct and confrontational modes of expression.

Overall, the developed methodology has a broad range of potential applications. First, it can be used to represent various texts, registers and modes of language at a metadiscourse level. By mapping these linguistic units onto a discourse space, it provides an opportunity to explore the relationships between different features within a text, or to compare the distribution of features across different texts or registers at a higher metadiscourse level. Second, different regions within the discourse space can be associated with varying levels of text quality. As demonstrated in this study, some regions are linked to bad-faith practices and rhetoric. Identifying these regions allows for the detection of malicious interventions in discourse. Third, the methodology offers a novel way to approach the concept of topic development at the discourse level, facilitating the identification of how discourse evolves over time. Finally, this approach provides a foundation for evaluating the performance of LLMs and other automated systems, whose outputs primarily rely on semantic and low-level lexical features, by comparing them against human interpretations of discourse, as demonstrated in a related study (Authors, 2024, preprint).

5.3. Limitations

The first constraint concerns the asymmetry of discourse dimensions of the discourse space as construed in this research. Thematic strategies have different levels of significance and influence within the discourse space. Whereas cohesion and logical quality strategies display a more pronounced global impact, attitudinal strategies tend to be more locally oriented. The difference in the nature of thematic strategies may result in their asymmetric representations in the discourse space. One potential approach to capitalise on this limitation could be selecting an equal number of thematic strategies for each level of significance across the three dimensions of the discourse space.

The second constraint is related to the completeness of the discourse space representation. We selected a restricted set of thematic strategies—those that had a reasonable agreement between human coders during the reliability test and that were adequately represented in our dataset. However, there are multiple other thematic strategies that could be potentially considered such as problem–solution, cause-effect, procedural exploration, encouragement, frustration, enthusiasm, ad rem and faulty analogy, etc.

Finally, the proposed approach represents one of the first attempts to visualise texts using purely discourse-level features, serving primarily as a proof of concept that such a level of representation is indeed feasible. Given that discourse is an inherently complex, sophisticated and qualitative phenomenon—one that resists full reduction to semantic or lexical levels—the method we developed is necessarily extensive. However, with continued application and exploration, this approach holds significant potential for further refinement, methodological development and the condensation of its current abstractions into more streamlined forms.

6. Conclusion

This paper has introduced the concept of discourse space as a framework for identifying discourse patterns and examining topic development at a higher level of abstraction. The proposed model functions as a pattern-recognition system at the metadiscourse level that enables the detection of how higher-order discourse features interact: i.e., which thematic strategies tend to co-occur and which are mutually exclusive. Our analysis reveals, for example, that in cases of discourse derailment, a superficial linguistic connection to the main topic is often maintained. However, derailment may emerge unexpectedly when certain strategies intersect, such as when categorisation is paired with irony or sympathy is expressed alongside the wish strategy. The most pronounced levels of derailment are associated either with the absence of strategies or with the combination of multiple thematic strategies, including hyperbole, hypothetical scenarios, didactic exposition, red herring and non-sequitur elements. Another notable finding is the divergent patterning of agreement and disagreement within discourse. Disagreement tends to align with temporal, consequential and sympathetic elements, while agreement more frequently co-occurs with self-referential statements and hypothetical scenarios.

The study also suggests the value of conceptualising discourse patterns as a hierarchical structure, ranging from low-level lexical patterns to metadiscourse thematic strategies that exert a global influence on discourse. Further exploration of this multi-level pattern architecture may yield insights into the ways in which metadiscourse strategies correlate with semantic, lexical and grammatical features at lower levels of analysis. Such insights hold potential for improving automated detection systems and advancing the discourse-processing capabilities of LLMs, by enabling them to align discourse structures more closely with human-coded interpretive dimensions.

Furthermore, due to the methodological breadth and metadiscourse focus, the discourse space framework offers a wide range of analytical possibilities. Its applications extend beyond discourse pattern recognition to include the interpretation and comparison of meaning across texts and registers, the computational detection of topic derailment and the evaluation of how closely LLMs' discourse representations approximate human understanding. It also offers a productive framework for cognitive inquiry, where the distribution and interaction of discourse patterns may provide insight into how individuals process language and structure information in context.

Finally, this method offers valuable tools for comparing the mediating effects of different social or media platforms and for identifying discourse features that are specific to a given medium. Because it enables the construction of a generative model of discourse behaviour within mediated environments, it also allows for the analysis of how such discourse evolves over time. From this perspective, one can begin to examine how distinct discourse spaces emerge in response to different forms of technological mediation.

CRediT authorship contribution statement

Kateryna Krykoniuk: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Cleo Hopkin-King:** Writing – review & editing, Writing – original draft, Project administration, Investigation, Formal analysis, Conceptualization. **Seán G. Roberts:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the AI and autonomy for intelligence, surveillance and reconnaissance (A2ISR) project at the Defence Science and Technology Laboratory (Dstl) through the Defence and Security Accelerator (DASA), which is part of the UK Ministry of Defence, United Kingdom.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dcm.2025.100929.

Data availability

The authors do not have permission to share data.

References

Bublitz, W., 2011. Cohesion and coherence. Discursive Pragmatics 8, 37–49.

Benson, P., 2016. The discourse of YouTube: Multimodal text in a global context. Routledge.

Benton, J.(2018). Here's how much Americans trust 38 major news organizations (hint: not all that much. Nieman Lab. Archived from the original on 8 December 2020. Retrieved 29 December 2022. https://www.niemanlab.org/2018/10/heres-how-much-americans-trust-38-major-news-organizations-hint-not-all-that-much/.

Bou-Franch, P., Blitvich, P.G., 2014. Conflict management in massive polylogues: a case study from YouTube. J. Pragmat. 73, 19–36.

Brown, A., 2018. What is so special about online (as compared to offline) hate speech? Ethnicities 18 (3), 297–326. https://www.jstor.org/stable/26497929.

Brown, G., Yule, G., 1983. Discourse analysis. Cambridge University Press, Cambridge.

Chakravarthi, B.R., 2022. Hope speech detection in YouTube comments. Soc. Netw. Anal. Min. 12 (1), 75.

Corral, M. (2023). Vectors in Euclidean Space. Available at: https://www.mecmath.net/VectorCalculus.pdf.

Dinnage, R. (2023). Surfing the grammar manifold: Ancestral state estimation and evolutionary rate estimation of Grambank characters on a phylogeny using latent generative models. Grambank Workshop, 13 - 14 September 2023, MPI-EVA, Leipzig, Germany.

Döring, N., Mohseni, M.R., 2020. Gendered hate speech in YouTube and YouNow comments: results of two content analyses. SCM Studies in Communication and Media 9 (1), 62–88.

Gee, J.P., 1999. An introduction to discourse analysis: Theory and Method. Routledge, New York and London.

Godfrey, J., and Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium.

Golovchenko, Y., Buntain, C., Eady, G., Brown, M.A., Tucker, J.A., 2020. Cross-platform state propaganda: Russian trolls on twitter and YouTube during the 2016 US Presidential Election. Int. J. Press/politics 25 (3), 357–389.

Govers, J., Feldman, P., Dant, A., Patros, P., 2023. Down the rabbit hole: detecting online extremism, radicalisation, and politicised hate speech. ACM Comput. Surv. 55 (14s), 1–35.

Grice, P. (1975). Logic and conversation. In: Cole, P.; Morgan, J. (Eds.). *Syntax and semantics*. Vol. 3: Speech acts. New York: Academic Press. pp. 41–58.

Gupta, S., Jain, G., Tiwari, A.A., 2023. Polarised social media discourse during COVID-19 pandemic: evidence from YouTube. Behav. Inform. Technol. 42 (2), 227–248.

Halliday, M.A.K., Hasan, R., 1976. Cohesion in English. Longman.

Heydari, A., Zhang, J., Appel, S., Wu, X., & Ranade, G. (2019). YouTube chatter: Understanding online comments discourse on misinformative and political YouTube videos. arXiv preprint arXiv:1907.00435.

Herring, S. C. (1999). Interactional coherence in CMC. In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. HICSS-32. Abstracts and CD-ROM of Full Papers, pp. 13–pp. IEEE.

Hussein, E., Juneja, P., Mitra, T., 2020. Measuring misinformation in video search platforms: an audit study on YouTube. Proc. ACM. Hum. Comput. Interact. 4 (CSCW1), 1–27.

Hyland, K., 2005. Metadiscourse: Exploring Interaction in Writing. A&C Black.

Igwebuike, E.E., Chimuanya, L., 2021. Legitimating falsehood in social media: a discourse analysis of political fake news. Discourse Commun. 15 (1), 42–58.

Inwood, O., Zappavigna, M., 2023. Attitudes about propaganda and Disinformation: Identifying discursive personae in YouTube comment sections. In: The Routledge Handbook of Discourse and Disinformation. Routledge, pp. 239–257.

Johnson, R. H., & Blair, J. A. (2002). Informal logic and the reconfiguration of logic. In Studies in Logic and Practical Reasoning (Vol. 1), pp. 339–396.

Kassambara, A., and Mundt, F. (2020)._factoextra: Extract and Visualize the Results of Multivariate Data Analyses_. R package version 1.0.7, <https://CRAN.R-project.org/package=factoextra>.

Krijthe, J., H. (2015). Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation, URL: https://github.com/jkrijthe/Rtsne.

Authors, (2024).

Landis, J.R., Koch, G.G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 363–374.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: an R package for multivariate analysis. J. Stat. Softw. 25 (1), 1–18. https://doi.org/10.18637/jss.v025.i01.

Luke, D.A., Caburnay, C.A., Cohen, E.L., 2011. How much is enough? New recommendations for using constructed week sampling in newspaper content analysis of health stories. Commun. Methods Meas. 5 (1), 76–91.

Miltsakaki, E., Kukich, K., 2004. Evaluation of text coherence for electronic essay scoring systems. Nat. Lang. Eng. 10 (1), 25–55.

Munkres, J.R., 2018. Analysis on manifolds. CRC Press.

Pinchuk, S., Shafikov, R., and Sukhov, A. (2023). *Geometry of holomorphic mappings*, pp. xi+-213. Birkhäuser.

Reinemann, C., Stanyer, J., Scherr, S., Legnante, G., 2012. Hard and soft news: a review of concepts, operationalizations and key findings. Journalism 13 (2), 221–239.

Sabbah, F., 2024. Critical discourse analysis approaches to investigating fake news and disinformation. In the Routledge Handbook of Discourse and Disinformation 33–51.

Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.

Sagredos, C., Nikolova, E., 2022. 'Slut I hate you' a critical discourse analysis of gendered conflict on YouTube. J. Language Aggression and Conflict 10 (1), 169–196.

Schubert, C., Renkema, J., 2018. Introduction to discourse studies. Introduction to Discourse Studies 1–469.

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., Menczer, F., 2018. The spread of low-credibility content by social bots. Nat. Commun. 9 (1), 1–9.

Sinclair, J., 1991. Corpus, concordance, collocation. Oxford University Press.

Tindale, C.W., 2007. Fallacies and argument appraisal. Cambridge University Press.

Toulmin, S.E., 2003. The uses of argument. Cambridge University Press.

van der Maaten, L.J.P., Hinton, G.E., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

van Dijk, T.A., 1980. The semantics and pragmatics of functional coherence in discourse. J. Pragmat. 4, 233–252.

Wattenhofer, M., Wattenhofer, R., Zhu, Z., 2021. The YouTube social network. Proce. Int. AAAI Conference on Web and Social Media 6 (1), 354–361. https://doi.org/10.1609/icwsm.v6i1.14243.

Watson Todd, R., 2016. Discourse topics. Discourse Topics 1–320.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wignell, P., Chai, K., Tan, S., O'Halloran, K., Lange, R., 2021. Natural language understanding and multimodal discourse analysis for interpreting extremist communications and the re-use of these materials online. Terrorism and Political Violence 33 (1), 71–95.

Williams, T.J.V., Tzani, C., 2024. How does language influence the radicalisation process? A systematic review of research exploring online extremist communication and discussion. Behav. Sci. Terror. Political Aggress. 16 (3), 310–330.

Webster, D.M., Kruglanski, A.W., 1994. Individual differences in need for cognitive closure. J. Pers. Soc. Psychol. 67 (6), 1049.

Potter, A., 2008. Interactional coherence in asynchronous learning networks: A rhetorical approach. The Internet and Higher Education 11 (2), 87–97.

R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.