Final version of the accepted paper. Cite as: *Asimul Haque and Muhammad Abulaish, CARES: A commonsense knowledge-enriched and graph-based contextual learning approach for rumor detection on social media, Expert Systems with Applications, Elsevier, Volume 266, 125965, 2025. DOI: https://doi.org/10.1016/j.eswa.2024.125965*

CARES: A Commonsense Knowledge-Enriched and Graph-Based Contextual Learning Approach for Rumor Detection on Social Media

Asimul Haque[a], Muhammad Abulaish[a,*]

[a]*Department of Computer Science, South Asian University, New Delhi, India*

Abstract

With the popularity of social media, rumors have become more prevalent than ever. The impact of rumors is adverse, making their detection a growing concern and an indispensable requirement. It necessitates a refined rumor detection approach that ensures credible information to users and maintains a trustworthy social media ecosystem. This paper presents a rumor detection approach, CARES, as a fusion of contextual learning-based rumor-relevant information, commonsense-based background knowledge, and embedding models. It utilizes the knowledge associated with the social media posts and reactions to leverage the textual and latent information and capture their inherent semantic affinity, the dynamics of social interaction such as responsive and reactive behavior of social media users, and commonsense-based background knowledge. Considering that social media entities are intertwined and rumors correlate with topics and contextual entities, CARES identifies and employs them to encode underlying rumor-based patterns and provide commonsense-based understanding. It uses a graph-based representation for contextual learning that recognizes two prevalent categories of words that form a building block of rumor-based patterns. The utilization of external lexicon-based knowledge in pattern selection leverages the reactive behavior of social media users in the expression of emotion, sentiment, skepticism and inquisitiveness, which makes the patterns more enriched and generic towards rumors. These patterns are finally ranked, and only the top-*k* check-worthy patterns are used for detecting rumors. CARES leverages commonsense understanding using ConceptNet, which provides facts as semantic evidence and enriches the posts with structural and background knowledge surrounding the contextual utterances and related topics. In order to preserve the semantic relations and commonsense understanding, it uses two pre-trained embedding models, GloVe and ConceptNet Numberbatch. The proposed approach is evaluated over three publicly available 𝕏 (formerly known as Twitter) datasets. The experimental results are promising, and they remarkably outperform the nine baselines and state-of-the-art approaches, demonstrating substantial effectiveness for detecting rumors on social media.

Keywords: Social Network Analysis, Rumor Detection, Contextual Learning, ConceptNet, Semantic Knowledge

1. Introduction

Social media networks and online communication tools provide convenient platforms to users for composing, exchanging, and propagating news, opinions, and ideas (Zubiaga et al., 2018; Bondielli & Marcelloni, 2019; Zhou & Zafarani, 2021). Such attractive features have increased the popularity of online social media drastically. Although these platforms provide convenience, they also facilitate malicious users to spread misinformation (Zubiaga et al., 2018; Bondielli & Marcelloni, 2019) that fuels and distorts issues rapidly. Misinformation acts like a virus in the sense that it exploits our weaknesses, biases, prejudice, and emotions and stymies our efforts to address them rationally. Rumor is one type of misinformation consisting of controversial news, including factually incorrect information about celebrities, politics, crises, and social affairs, which quickly diffuses across social media platforms with unchecked veracity. In due course, its veracity may be determined as true or false, or the same may remain unchecked (Zubiaga et al., 2018).

*Corresponding author
Email addresses: asimulhaque@students.sau.ac.in (Asimul Haque), abulaish@sau.ac.in (Muhammad Abulaish)

Rumors disseminate faster than authentic news and the negative consequences are aggravated (Vosoughi et al., 2018). This diffusion of rumors is impacted and facilitated by intrinsic human tendencies such as confirmation bias that leads users to ignore contrary information (Nickerson, 1998). The nature of people to usually read and share information if it resonates with their preconceived view without caring about their credibility drastically impacts individuals, societies, and nations through their faster dissemination and more undetectable disguises (Vosoughi et al., 2018). Therefore, automatic rumor detection is a challenging undertaking for the research community and an absolute necessity in today's culture.

Extensive studies have revealed that rumors and similar false information, such as misinformation and disinformation span almost every realm of human activity across diverse fields and pose a consequential threat to our social system, including financial markets, democracy, political affairs, and the health system (Allcott & Gentzkow, 2017; Sicilia et al., 2018; Azzimonti & Fernandes, 2023). For example, CHEQ and the University of Baltimore have researched and reported that the monetary damage caused by false information propagating mainly on social networks costs the global economy $78 billion yearly (Brown, 2019). Similarly, with the advent of COVID-19, many epidemic-related rumors appeared on social media worldwide, from the origin of coronavirus to anti-vaccine arguments, leading to a terrible impact on epidemic prevention, social stability, and health welfare efforts (Frenkel et al., 2020). The World Health Organization termed misinformation as an *"infodemic"*[1] and the World Economic Forum has ranked misinformation and disinformation as the biggest global risks in their Global Risks Report 2021 (McLennan, 2021) and 2024[2]. The severe implications of misinformation have motivated many governments, regional bodies, and international organizations to adopt multiple combating strategies. For example, the United Nations urged social media users to *"pause take care before share"*. Similarly, `verified` is one of the UN initiatives launched in response to the COVID-19 crisis and has reached over a billion people worldwide with reliable and accurate information. In addition, numerous fact-checking sites like `Snopes`, `PolitiFact`, `RealityCheck`, `FactCheck`, and `NewsGuard` are available to evaluate the accuracy of claims and news websites. Apart from that, social media firms employ experts and encourage users to flag or report non-credible posts. They have started removing the content and blocking the users who propagate misinformation and hate speech. Nevertheless, these processes are slow and involve much human effort (Graves, 2018). At a time when technological advancements such as `ChatGPT` and `Generative-AI` have further complicated the evolving hazards and threats of false information[3], more than these approaches and human-dependable manual fact-checking are needed to combat them in order to reduce their impact on social media users and society as a whole. It requires adaptability on multiple fronts, including social media platform regulations, media literacy programs, the responsible adoption of advanced emerging tools, and the devising of robust rumor detection models for addressing the issue on a larger scale.

On social media, crowd intelligence is a valuable source for debunking rumors since, on very few occasions, posts alone provide clues of their non-credibility (Zubiaga et al., 2018; Li et al., 2023a). If a post seems unreliable, users express opinions on its veracity and insightful clues, enabling a chain reaction that spreads over the network and produces distinctive bursty retweets and replies within very few minutes, considerably higher than generated by reliable ones (Ma et al., 2017, 2018). Like-minded social media users having the same interests and preferences are likelier to share similar latent information about an event, which circulates an analogous and comparable discourse and conversation when it unfolds (Nguyen et al., 2020; Haque & Abulaish, 2022). The circulation of original posts enables users to provide semantic information and reveals emotion and skepticism in their reactions, convincing us to incorporate them together while determining their credibility. We relate the social media phenomenon of echo chambers, where social entities are closely connected and often behave similarly and interact with other entities of common interest to reinforce and promote their narratives (Vicario et al., 2016; Nguyen et al., 2020). We focus on leveraging the knowledge from social media posts and reactions to extract and understand both explicit textual content and implicit latent information. The underlying semantic relationships between posts and their reactions can significantly enhance the understanding of how information spreads and influences behavior. This inherent semantic affinity of posts and reactions can lead to a more nuanced understanding of public discourse, ultimately contributing to better decision-making. The motivation behind our work is to capture knowledge from social media content and their interactions involving user engagements and opinions, which provide insights into public sentiments, trends, and

[1] `https://www.who.int/health-topics/infodemic`

[2] `https://www.weforum.org/press/2024/01/global-risks-report-2024-press-release/`

[3] `http://bit.ly/ai-powered-misinformation-is-the-worlds-biggest-short-term-threat-davos-report`

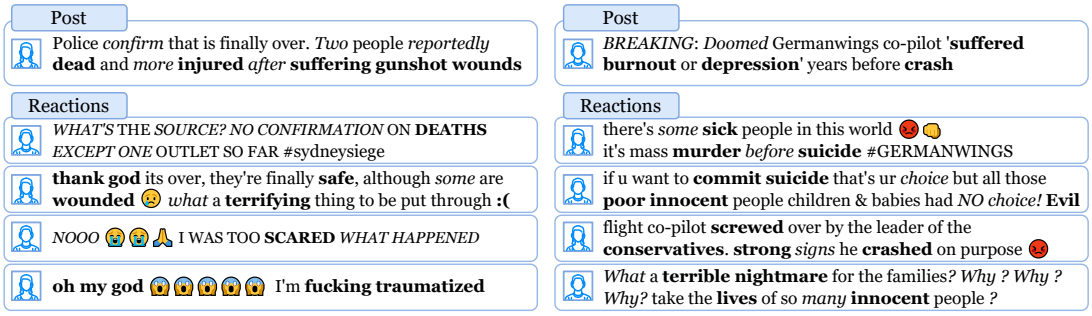


Figure 1: Two examples of rumor with their relevant social context. The italic text signifies the skepticism of social media users and the semantic link between posts and reactions, whereas the bold text captures the emotional resonance with the users

impacts (Vicario et al., 2016; Nguyen et al., 2020). Social media content is often informal, abbreviated, and filled with slang or regional dialects, leading to significant semantic ambiguity. Reaction helps to handle the challenge of correctly interpreting the content, especially when the same words or phrases can have different meanings in different contexts. Incorporating reactions can provide additional information and context, which are used to explore more latent information, such as sarcasm, implicit opinions, implied meanings, and cultural references. We focus on capturing these contexts more in-depth, leading to more accurate and contextually aware content analysis. Thus, the discourse on the topics and conversation provoked by the posts provides context and latent information, facilitating the detection of their veracity by utilizing and considering relevant and contextual information underlying the posts and associated reactions. Figure 1 displays two examples of rumor from the datasets, illustrating the relevance of social context and semantic and emotional correlations between posts and reactions in credibility assessment. Some reactions are aimed at verifying, discussing, or reinforcing the factual aspects of the post, while others align with the post's emotional tone, such as anger, fear, sadness, or astonishment.

In this paper, we present a graph-based rumor detection approach, `CARES`, to harness the context better and to focus on the underlying interactions between important linguistic components, which efficiently captures the global mutual use of linguistic variations in textual data, especially on social media, wherein linguistic behavior is socially and culturally influenced. It augments contextual learning using graph-based methods and related commonsense knowledge-based representation to determine the combinations of feature sets that yield the best predictive capabilities for rumor detection on social media. The graph-based method extracts contextual entities and relevant information for constructing pattern-based rumourous features that capture latent rumor-relevant information. The patterns are further enriched by social media users' emotional, inquisitive, and skeptical behavior using external lexicon-based knowledge. The proposed approach not only relies on linguistic, contextual, and semantic features from social media posts and reactions but also leverages knowledge other than the implicit information inherent in the text and the knowledge surrounding the contextual utterances and semantics, which enables the analysis of a human conversation that drives contextual knowledge from commonsense (Zhong et al., 2019; Sharma et al., 2023; Zhang et al., 2024). We also refer that the seeming plausibility of posts creates the impression of veracity since humans believe depending on how much a story fits their prior knowledge (Connell & Keane, 2004). This prior and background knowledge also acts as fact and semantically supporting evidence associated with the contextual utterances that benefit the rumor detection task. Arguing that commonsense understanding is fundamental for conversations and an essential factor in human believability, the proposed approach leverages `ConceptNet` (Speer et al., 2017) as a source of an external commonsense knowledge representation that uses contextual utterance, entities, and relevant topics to address the challenge of lacking prior and background knowledge. `ConceptNet` is a semantic network that assists machines in understanding the meanings and semantic relationships of words and specific facts used by humans.

The proposed approach analyzes the dynamics of social media posts and associated reactions. Firstly, the textual data of social media is modeled as a word co-occurrence graph, which produces two prevalent categories of words – *substantial words* and *bridge words*. These words are extracted using two node-related metrics: clustering coefficient and eigenvector centrality. These graph-based metrics are used to know the correlation and allow the exploitation of contextual and relevant information related to rumors (Haque & Abulaish, 2022). The *substantial words* are usually subject-related and topical words clustered together, specified by the clustering coefficient. The *bridge words*, on the

other hand, are identified using eigenvector centrality, which connects the *substantial words*. These two families of words collaborate to construct contextually rich patterns. Secondly, the proposed approach utilizes prior and background knowledge as supporting evidence and leverages a commonsense understanding encompassing contextual utterances, related topics, and semantics. These two different preferences of textual clues play complementary roles in detecting rumors on social media. Furthermore, we use two pre-trained word embedding models to preserve semantic-related knowledge: `GloVe` (Pennington et al., 2014) to learn the potential similarity between rumor-based patterns and incoming posts, and the `ConceptNet Numberbatch` (Speer et al., 2017) to provide semantically commonsense-based background knowledge. Finally, the rumor patterns-based features, commonsense-based background knowledge representation, and embedding models are fused to generate feature set for rumor classification. A preliminary version of this work has been published in a conference proceedings (Haque & Abulaish, 2022). However, this paper substantially enhances the prior work. It leverages the posts and reactions for better representation and provides prior and background knowledge as fact and evidence. It improves the learning ability to detect rumors by fusing rumor-based structural patterns, text semantics, lexicon-based knowledge, and commonsense-based background knowledge. In summary, the main contributions of this paper can be briefly described as follows:

- Developing a commonsense knowledge-enriched graph-based contextual learning approach, `CARES`, which integrates contextualized rumor-based patterns, commonsense-based background knowledge, and word embedding models for rumor detection.
- Considering the usefulness of posts and reactions and the dynamics of social interaction, such as responsive and reactive behavior of social media users for effectively detecting the rumors on social media.
- Extracting the features for rumor classification in two aspects. Firstly, it learns the representation of contextual patterns using a graph-based approach that produces rich structural descriptors for rumors. It utilizes external lexicon-based knowledge to make these patterns enriched and generic. Secondly, it leverages the commonsense-based understanding surrounding contextual utterances, related topics, and structural descriptors.
- Performing an empirical study on three benchmark datasets that demonstrates the efficacy of the proposed approach. It also highlights the effectiveness of incorporating information from users' reactions and commonsense-based background knowledge in rumor detection.

The rest of the paper is organized as follows. Section 2 presents a brief review of the related literature. Section 3 presents the functional details of `CARES`. Section 4 presents the details of the datasets and experimental setup. It also presents the evaluation results and comparative analysis of `CARES`. Finally, Section 5 concludes the paper and provides direction of future work.

2. Related Works

There are several ways to categorize rumor detection methods. This study exclusively examines text-based approaches, which are categorized based on the type and source of features, including content-based approaches, social context-based approaches, and external knowledge-based approaches.

2.1. Content-Based Approaches

Social media contents contain clues, mainly from textual and visual aspects, to differentiate rumor from non-rumor. Numerous approaches have focused on investigating characteristics that are based on hand-crafted features. In the first work on evaluating information credibility on social media, (Castillo et al., 2011) listed post-based features, such as word length, exclamation or question marks. (Zhao et al., 2015) proposed the detection of emerging rumors based on correction and verification signals extracted using regular expressions that exploit the inquiry mind of users. (Vaibhav et al., 2019) modeled a document as a sentence graph that captured the sentence interactions and applied graph attention networks to learn document representation. (Haque & Abulaish, 2022) proposed an approach that modeled latent textual representations and underlying information using a graph-based approach for extracting rich rumorous patterns that achieved good performance in detecting rumors. Kim & Ko (2021) presented a graph-based method using a summarization technique that uses document information for fake news detection. Recent methods

investigated emotional and sentimental aspects for rumor detection (Liu et al., 2023). (Abulaish et al., 2019) proposed a graph-based approach for rumor detection using anxious and doubtful terms present in microblogging posts. Ajao et al. (2019) determined the sentimental relationship between authentic and fake content in social media posts. Rezaei et al. (2023) investigated fake news detection using the most useful linguistic features, such as textual, sentiment, semantic, and readability. However, using content only as features without exploring social interactions, external evidence, or explanations provides limited context and knowledge for rumor detection.

2.2. Social Context-Based Approaches

The social context-based approaches use social media users' profiles and reactions, including replying, retweeting, and mentioning. Multiple studies have shown that social context against rumors encompasses different and distinct characteristics and factors, such as the timing of their reactions, the speed of dissemination, and the quantity and nature of their responses (Ma et al., 2017, 2018; Kochkina et al., 2018). Social context-based features mainly include user-based, post-based, and network-based information containing user profiles and social relationships (Tredici & Fernández, 2020). It also includes stance-based and propagation-based approaches that utilize users' opinions to infer authenticity, assuming that the credibility of social media posts is highly related to credible propagation networks. The social context-based approaches are based on the wisdom of the crowd, which is not limited to extracting representative features from the textual content but makes more use of the social contextual information of posts and reactions used as an auxiliary judgment and effectively exploits users' sentiments and emotional aspects (Haque & Abulaish, 2022, 2023). Zubiaga et al. (2017) leveraged context learning with the assumption that context is crucial in determining the veracity of rumors. This approach has further extended to exploiting the homophily nature of social media users (Lathiya et al., 2020). Nguyen et al. (2020) presented a graph learning framework for fake news detection using the social structural engagements of users. Ma et al. (2017) created a kernel-based method that captures high-order patterns for differentiating rumors. Ma et al. (2018) developed a tree-structured recursive neural network to learn the embedding of rumor propagation structures. Shu et al. (2019) and (Lu & Li, 2020) learned the sequential effect of users' responses and proposed the co-attention model to leverage content and social context for detecting credible information. Bai et al. (2021) used a graph-based relationship between source and reply and proposed an ensemble graph convolutional neural network with a node proportion allocation mechanism. Tu et al. (2021) proposed a CNN-based model using text and propagation structure. Recently, some studies based on robust mathematical modeling have been proposed for rumor diffusion. Ghosh et al. (2023) presented mathematical model to investigate the rumor propagation mechanism with the interference of inhibitors. This model not only signifies the dynamics of rumor propagation on social media but also has the potential to significantly impact the understanding and regulation of social media dynamics, particularly with the influence of counter-rumor spreaders in regulating the transmission process and controlling its harmful effects. Guo & Yan (2023) proposed a rumor propagation model with the double refutation mechanism. This model considered the combined effect of the external refutation of media reports and the internal refutation of the counter individuals. It has used the thresholds for rumor propagation, obtained by mathematical analysis of the model using Routh-Hurwitz judgment and Jordan matrix theory. Li et al. (2023b) proposed a novel source location method based on infection potential energy. This approach used a network reconstruction method and a network pruning method to fully use diffusion information in the network and reduce the search space of candidate source nodes since most existing methods ignore the direction of infected nodes. This approach developed a novel source location method considering rumor propagation centrality and diffusion tendency.

2.3. External Knowledge-Based Approaches

The external knowledge-based rumor detection approaches can be classified in two ways. One line of work focuses on fact-checking, which commonly uses structure triplets extracted from knowledge graphs comparable to the post contents (Fionda & Pirrò, 2018). Sheng et al. (2021) integrated fact-based and features-based models using external fact-checking sources into one framework via model preference learning. Sun et al. (2022) modeled the dynamics of propagation and knowledge-entity structures from a knowledge graph with a unified framework. Hu et al. (2021) used the entity linking method to capture entity descriptions and leverage the knowledge base for fake news detection. Another parallel line of work, such as (Yang et al., 2023) focused on news content and social media replies and used external knowledge to capture critical entities. Moreover, several studies have focused on incorporating commonsense knowledge graph databases, such as ConceptNet (Speer et al., 2017), which use graphical structures to express intricate

relations between concepts to boost performance on various tasks (Xie & Pu, 2021). These tasks vary from generative conversation, such as open-domain dialog and question-answering, to sentiment analysis and stance detection (Bauer et al., 2018; Cambria et al., 2022; Zhang et al., 2024). Likewise, some studies demonstrated the usefulness of commonsense knowledge in text representations for satire, meme, and emotion detection (Zhong et al., 2019; Sharma et al., 2023). Additionally, the advancement of visual technology has facilitated the augmentation of images with text, enabling models to extract cross-modal semantic relationships, thus enhancing multimodal rumor detection. Wang et al. (2023) proposed a cross-modal contrastive learning framework for multimodal fake news detection. This approach focused on image-text contrastive learning objectives to achieve more accurate image-text alignment. It introduced an attention guidance module to effectively and interpretably aggregate features from different modalities. Luvembe et al. (2024) proposed a unified complementary attention fusion with an optimized deep neural network that captures subtle cross-modal relationships for multimodal fake news detection. They utilized image captions to represent them in a shared semantic space with text to capture subtle relationships. A complimentary attention method based on scaled dot product was applied to fuse captions and text bi-directionally to learn fine-grained correlations. This technique has enhanced semantic significance and cross-modal interaction, mitigating feature variations and noisy features. The optimized deep neural network has leveraged compositional learning for higher-level representations.

It can be observed from the above discussions that researchers have made significant strides in exploring various facets of rumor detection. However, none considered the combined influence of affective, contextual, and commonsense knowledge of social media users for detecting rumors. It may be noted that the commonsense-based background knowledge is also a standard paradigm used by human judgment and can significantly assist in judging the credibility of a social media post. Moreover, existing works have not fully embraced the contextual learning approach, which is crucial for understanding and inferring the context and latent information embedded within social media posts and reactions. These post-reaction interactions provide rich data for identifying sentiment, skepticism, sarcasm, irony, and emotion patterns, which are prominent characteristics of social media users and potentially valuable features for credibility analysis. Moreover, existing works are limited in addressing the challenge of lacking prior and background knowledge, which is fundamental for conversations, an essential factor in human believability, and crucial for real-world applications where topics evolve constantly.

3. Proposed Approach

This section presents functional details of the proposed rumor detection approach, `CARES`, as visualized in Figure 2. It primarily focuses on a graph-based contextual learning approach for extracting descriptive rumor-based patterns. The proposed approach incorporates the posts and reactions to better infer underlying information and social interaction dynamics for constructing rumor-based structural patterns that comprise the affective information through emotion and sentiment knowledge, and semantic information from word embeddings. In social media context, the term *reactions* typically refer to the response or feedback from users to a particular post, including but not limited to comments, likes, shares, retweets, emojis, etc. However, we have considered only comments (reply text) and emojis to model user reactions. The proposed approach leverages a commonsense knowledge graph to provide semantically commonsense-based background knowledge. It highlights the learning ability to detect rumors by fusing rumor-based structural patterns, text semantics, lexicon-based knowledge, and commonsense-based background knowledge that are described in the following sub-sections.

3.1. Graph Generation

In order to utilize the knowledge associated with the posts and reactions and leverage the textual and latent information to capture their inherent semantic affinity, we incorporate them using graph-based methods, which can capture the global mutual use of linguistic variations and contextual information in textual data. These methods emphasize the importance of considering and harnessing the context of social media posts and focus on underlying interactions between important linguistic components, which are socially and culturally influenced. Considering posts and reactions are highly correlated, posts contain linguistic cues such as opinionated and sensational language styles, and reactions contain supplementary knowledge and useful semantic information in the form of emotions or sentiments, opinions, skepticism, and inquisitiveness, which help explore more latent information and identify the underlying rumor-based patterns. Their inclusiveness is maintained by combining them in a single input text. For each social

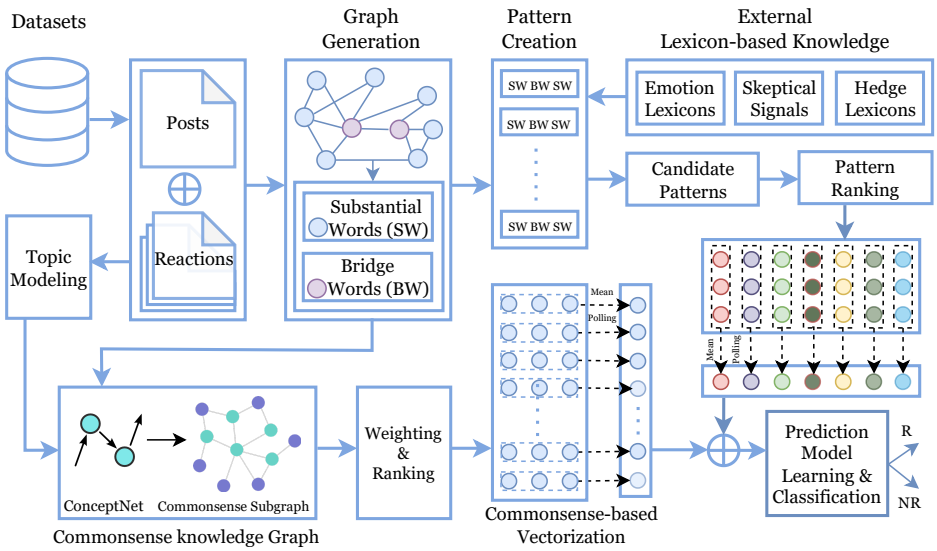


Figure 2: The work-flow of the proposed rumor detection approach, CARES

media source post t_i, its associated reactions $R_{t_i} = \{r_1^{t_i}, r_2^{t_i}, ..., r_{|R_{t_i}|}^{t_i}\}$ are used to create a single input text by concatenating them using equation 1.

$$\mathcal{X}_i = t_i \oplus \{R_{t_i}\}_{r_1}^{|R_{t_i}|} \tag{1}$$

The input texts are modeled as a word co-occurrence graph $G(V, E)$, where words are taken as a finite set of nodes V, and E represents the correlation between words. The graph G is modified by pruning the edges using two subgraphs, G_r and G_{nr}, which are also word co-occurrence graph of input texts labeled as rumor and non-rumor in training data, respectively. For the pruning purpose, point-wise mutual information (PMI) is employed on both G_r and G_{nr} to calculate the weights of edges that preserve the global word co-occurrence information. PMI scores can reflect the correlation between words based on the global corpus rather than specific post content. We only preserve edges with positive point-wise mutual information (PPMI) scores, which implies high semantic correlations. The PPMI of word pairs is calculated using equations 2, 3, and 4, where $|W|$, $W(w_i)$, and $|W(w_i, w_j)|$ are the total number of sliding windows, the number of sliding windows containing the word w_i, and the number of sliding windows containing both words w_i and w_j, respectively.

$$p(w_i) = \frac{|W(w_i)|}{|W|} \tag{2}$$

$$p(w_i, w_j) = \frac{|W(w_i, w_j)|}{|W|} \tag{3}$$

$$PPMI(w_i, w_j) = \max\left(\log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, 0\right) \tag{4}$$

After applying the above equations on both the subgraphs G_r and G_{nr}, the adjacency matrix $A(i, j)$ of graph G is updated as shown in equation 5, where $PPMI_r(w_i, w_j)$ and $PPMI_{nr}(w_i, w_j)$ are PPMI scores of the word pairs (w_i, w_j) in G_r and G_{nr}, respectively.

$$A(i, j) = \begin{cases} 1, & PPMI_r(w_i, w_j) \geq PPMI_{nr}(w_i, w_j) \\ 0, & \text{Otherwise.} \end{cases} \tag{5}$$

The purpose of this step is to construct a graph that recognizes two prevalent categories of words comprising contextual and semantic relation and relevant to constructing rumor-based patterns. These two categories are *substantial* and *bridge* words collected by performing node-level statistical measures (i.e., clustering coefficient and eigenvector centrality) on the graph.

3.1.1. Substantial Words

Social media entities and utterances are intertwined and correlated with topics and events, which contain similar types of discourse and conversation, possessing contextual and latent information when an event starts unfolding. These entities also comprise users' responsive and reactive behavior that clustered psychological, sentimental, and skeptical words, imposing sentiments, emotions, inquisitiveness, and suspicions about the event. We refer to such topical and behavioral words that are usually clustered together as substantial words, and they are identified by calculating the clustering coefficient of each node $w_i \in V$ using equation 6, where N_i and k_i are the neighbors and degree of node w_i, respectively.

$$C_{c(w_i)} = \frac{2|e_{jk} : w_j, w_k \in N_i, e_{jk} \in E|}{k_i(k_i - 1)} \tag{6}$$

The average clustering coefficient of node $w_i \in V$ is defined as $C_{c(w_i)}/|V|$, and the average clustering score of graph G is defined as $\sum_{w \in V} C_{c_w}/|V|$. Finally, the list of substantial words is compiled with words having a clustering coefficient value greater than the empirically determined threshold θ_{C_c}.

3.1.2. Bridge Words

The substantial words identified above are connected with the words that we recognize as bridge words. As the name signifies, they make a bridge between two substantial words for their meaningful deliberation and contribution to creating rumor-based patterns. The bridge words mainly acknowledge the casual writing style of social media users, who usually represent the words informally. The eigenvector centrality measure is applied for identifying bridge words. This centrality assesses the influence of a node in a network with the hypothesis of considering the frequent as well as focal nodes in the graph and also keeping in mind that a node is more influential if it is connected to influential nodes. Also, the benefit of this centrality measure is to avoid frequent but irrelevant words in the graph. The eigenvector centrality of a node $w_i \in V$ is defined in equation 7 as a proportionality relation, i.e., the centrality of a node w_i is proportional to the combined centrality values of its neighbor $w_j \in N_i$, where N_i is the neighbor of a node w_i, and $A(i, j)$ is the adjacency matrix of graph G.

$$C_{e(w_i)} \propto \sum_{w_j \in N_i} A(i, j) C_e(w_j) \tag{7}$$

$$\lambda x = Ax \tag{8}$$

Equation 8 is the matrix form representation that exactly looks like the eigenvector equation, where λ is a proportionality constant. The centrality vector x is the eigenvector of the adjacency matrix $A(i, j)$ associated with the eigenvalue λ. Under the Perron-Frobenius theorem, it takes the largest value of λ. We find the corresponding eigenvector that is positive and unique, which gives us the eigenvector centrality of each node $w_i \in V$. The list of bridge words is compiled with the words having a centrality value greater than the empirically determined threshold θ_{C_e}.

3.2. Pattern Creation

The objective of this step is to create informative patterns with the use of two prevalent categories of words identified above. People follow syntactic rules in their communication to convey the proper meaning, which social media users also pursue. The order of words in a language is critical for delivering meaningful information. In this step, this syntactic rule and contextual information are regarded for creating patterns, which are more relevant for detecting rumors. For this purpose, we combine *substantial* (sw) and *bridge* (bw) words to create contextually and syntactically rich patterns that also incorporate their meaningful interaction. We first locate the words that satisfy the threshold condition from both categories in graph G. After that, the connection of one category of words to another is recognized for creating interactive and meaningful patterns, which are achieved by connecting the nodes with their immediate neighbors such that not all three nodes are from the same category. The connection of words in a way (e.g., $<$ *sw bw sw* $>$ and $<$ *sw bw bw* $>$) that extracts 3-word patterns that retain the contextual and syntactical meaning. All these created patterns are stored in a list P.

3.3. Pattern Selection

In this section, we are interested in selecting more informative, contextually and semantically rich, rumor-related patterns. The dynamics of social interaction, such as the responsive and reactive natures of social media users, are also regarded to leverage their emotional, inquisitive, and skeptical behavioral aspects. These aspects are incorporated using external lexicon-based knowledge, making patterns enriched and generic towards rumors. The following steps focus on selecting check-worthy patterns that are highly associated with rumors.

3.3.1. Candidate Patterns

The pattern selection in this step is based on the assumption that the rumor contents are related to a specific event, i.e., as events unfold, users begin to circulate similar sets of responses and generate comparable types of statements. For selecting event-specific rumor-based patterns, we match the created patterns P in section 3.2 with the portion of training data labeled as rumor only. For the purpose of matching, patterns are enriched using continuous word representations to preserve semantic relatedness with rumors. We utilize the publicly available pre-trained word embedding model `GloVe`, which gives a semantic vector representation of a pattern p_i as vector $[v_1, v_2, v_3]$, where $v_i \in \mathcal{R}^d$ and $p_i \in P$; $\mathcal{R}^d$ is a d-dimensional word embedding vector for a i^{th} word in a pattern p_i. We split the training rumor corpus into 3-grams and apply the same embedding model to preserve the semantic similarity with the patterns. The mean-pooling of embeddings is used to aggregate the embedded information. We are using the cosine similarity and carrying out the patterns that are similar to some extent with a similarity score above the determined threshold. The higher the similarity score, the more rumor-relevant the patterns are. The cosine similarity between patterns and 3-grams is calculated using equation 9, where p_i is a pattern, and n_j is a 3-grams from the training rumor corpus.

$$Cosine(p_i, n_j) = \frac{p_i \cdot n_j}{||p_i|| \cdot ||n_j||} \tag{9}$$

3.3.2. External Knowledge-Enriched Patterns

There are few occasions where posts provide clues of their non-credibility, whereas sometimes, users express uncertainty, emotion, skepticism, and hedges in their reactions (Liu et al., 2023; Haque & Abulaish, 2023). Users mainly re-circulate original posts with the negative sentiment that makes content appalling. Within a short time, denying and questioning the validity of these posts are observed in the reaction. In contrast, credible posts invoke moderate engagement, mainly comprised of a supportive stance with neutral sentiment that stabilizes quickly. Such temporal shifts in user perception serve as important signals for distinguishing rumors from non-rumors. We consider the dynamics of social interaction in the form of the responsive and reactive nature of social media users since the writing style and users' behavior are primarily similar for unverified or rumorous content. This step incorporates external lexicon-based knowledge for enriching pattern selections that leverage the emotional, sentimental, and hedged aspects of social media posts. It also employs the inquisitive nature of social media users, such as correcting and verifying posts. The incorporation of this knowledge for pattern selections is explained as follows:

Emotion lexicon-based knowledge leverages social media users' emotional and sentimental aspects since messages in human communications inherently convey emotions and sentiments, specifically in rumorous posts, which are presented to take advantage of arousal and sensitive users' behaviors. When reacting to rumorous posts, social media users usually have negative sentiments and express fear, anxiety, astonishment, surprise, and disgust Vosoughi et al. (2018). The capturing of patterns having emotions and sentimental words is incorporated using external lexicons – NRC Emotion Lexicon, NRC Affect Intensity Lexicon (NRC-AIL), SenticNet, and EmoSenticNet. The NRC Emotion Lexicon (Mohammad & Turney, 2013) consisting emotion words based on Plutchik's wheel; it also has negative and positive sentiments, while NRC-AIL (Mohammad, 2018) contains the same emotion categories with a real-valued intensity score for the words. EmoSenticNet (Poria et al., 2014) consisting words with associative emotion categories based on Ekman's six basic emotion taxonomy, while SenticNet (Cambria et al., 2016) consists of word sentiment scores. The NRC-AIL and EmoSenticNet lexicons are used to identify emotion words related to fear, anxiety, astonishment, anger, and disgust. We match list P with lexicons to retrieve the patterns with these emotion words. Further, NRC Emotion Lexicon and SenticNet are used to identify associations with sentiments by matching them with list P to retrieve the patterns with negative sentiments. All these pattern selections are based on n-gram matching. EmoSenticNet and SenticNet are based on commonsense knowledge, which enriches selected patterns in those aspects. Thus, the patterns with negative sentiments and emotions are chosen for feature enrichment.

Skeptical-based knowledge leverages social media users' skepticism, doubtfulness, and inquisitive behaviors evoked by mistrust toward online information. It is observed that when the posts get published, people show skepticism and start questioning and verifying them in the form of support, rebut, and denial. The correction and verification inquiries in the replies are also observed in (Zhao et al., 2015). Inspired by this work, (Haque & Abulaish, 2022) constructed regular expressions, such as (*true*|*not*|*true*), (*real*?|*really*?), (*rumor*|*debunk*), and ($false|fake$), for extracting skeptical patterns. We follow the same procedure and utilize these regular expressions to match the patterns in list P to select the skeptical patterns.

Hedge lexicon-based knowledge leverages indications of exaggerated words, vagueness, and mystifying and obscuring language, which people use to cultivate rumors. People use them to dramatize or sensationalize the social media post. Lexicons from (Islam et al., 2020) are used to identify hedge words in the patterns using n-gram based matching. Thus, those patterns that have hedge words are chosen from list P.

Incorporating external knowledge to utilize these dynamics of social media users drives the patterns to be enriched and generic since any type of rumourous post experiences these critical signals. Finally, the resulting check-worthy patterns are a combination of patterns selected in sections 3.3.1 and 3.3.2 and stored in list CP. These check-worthy patterns comprise the affective information provided by emotion and sentiment knowledge, semantic information provided by word embeddings, and contextual information using graph-based methods; hence, these are contextualized affective patterns.

3.4. Pattern Weighting and Ranking

This step determines the top-k check-worthy patterns from the selected patterns list CP. We are interested in finding highly relevant patterns that are more indicative of rumors; therefore, a weight value is assigned to each check-worthy pattern to define its ranking. It is achieved using the $tf - idf$ weight score, where each pattern is assigned a weighted value associated with the training corpus data labeled as rumor only.

In this step, the standard formula of $tf - idf$ is modified to calculate the weight score for mapping the patterns close to being rumorous. The modified $tf - idf$ is applied on the check-worthy selected patterns list CP to assign weight scores to the patterns. Each pattern of the list CP is considered as a term, and its frequency value is calculated, named pattern frequency pf. For this purpose, we split the input texts of the rumor training corpus into 3-grams and matched the patterns with them at a 3-gram level. The pattern frequency of a pattern pf_{cp_i} is defined in equation 10, where $cp_i \in CP$, tc is the rumor training corpus, $f(cp_i, tc)$ is the frequency of a pattern cp_i in tc, and $|cp|$ is the total number of patterns present in tc.

$$pf_{cp_i} = \log_2\left(1 + \frac{f(cp_i, tc)}{|cp|}\right) \tag{10}$$

The inverse document frequency idf for a pattern cp_i is computed to provide the relevance of pattern cp_i with the training corpus using equation 11, where $cp_i \in CP$, $n(X)$ is the total number of input texts in the rumor training corpus tc and $f(x, cp_i)$ represents the total number of input texts where pattern cp_i occurs.

$$idf_{cp_i} = \log_2\left(1 + \frac{n(X) \in tc}{f(x, cp_i)}\right) \tag{11}$$

Thus, the pattern frequency-inverse document frequency $pf - idf$ is calculated by scalar multiplication of equations 10 and 11, as shown in equation 12, where $pf - idf_{cp_i}$ is the final weight score for a pattern cp_i.

$$pf - idf_{cp_i} = pf_{cp_i} \times idf_{cp_i} \tag{12}$$

The pattern having greater the $tf - idf$ weight score, the more indicative rumor-based patterns are; we rank the patterns corresponding to their weight score for selecting top ranked check-worthy patterns. Finally, the top-k ranked patterns are used for detecting rumors on social media effectively. We experimentally demonstrate how the top-k patterns produce better results, where the value of k is determined empirically.

3.5. Commonsense Knowledg-Based Enrichement

This section leverages external commonsense knowledge with the assumption that human conversation fundamentally drives contextual knowledge from commonsense understanding, and their reasoning depends on how much

a story fits their prior knowledge. The above sections consider linguistic, contextual, and semantic features from social media posts and reactions. Now, we utilize prior and background knowledge that leverages a commonsense understanding encompassing contextual utterances, related topics, and semantics. Inspired by the successful utilization of commonsense knowledge representation in addressing the challenge of lacking background knowledge in several tasks (Zhong et al., 2019; Sharma et al., 2023; Zhang et al., 2024), we employ `ConceptNet` (Speer et al., 2017) as a source of external commonsense-based knowledge, which also addresses the challenge of lacking prior and contextual background knowledge. In our approach, subjective knowledge is gained using reactions and objective knowledge using external knowledge. Both types of knowledge are useful since the objective information is consistent, has high purity, and can be used as evidence. At the same time, subjective information contains certain clues for debunking rumors. Therefore, the prior and background knowledge is also considered to provide fact and semantically supporting evidence associated with the contextual utterances and related topics that benefit the rumor detection task.

3.5.1. Commonsense Subgraph Construction

`ConceptNet` is a large-scale multilingual and widely used commonsense knowledge semantic graph containing over 5.9 million assertions, 3.1 million concepts, and 38 symmmetric and asymmetric relations for English. The nodes are concepts, and the edges are relations, providing a textual description of node relationship types. Each triplet is an assertion, represented as $R = (c_1, r, c_2)$, where c_1 is the head concept, c_2 is the tail concept, and r denotes the relationship between these two concepts. Each assertion is associated with a confidence score [1; 10]. An example assertion is *(fear, Causes, hate)* with a confidence score of 2.32. Due to the enormous size and the large amount of unnecessary information, we need an effective way of selecting relations with the assumption that important concepts are those that relate to the contextual utterance and related topics. The aim is to capture commonsense-based structural and semantic information related to social media posts that facilitates the detection task. For this purpose, we utilize the identified prevalent categories of words in section 3.1 and topical words from the text content of the dataset to match them with concept nodes in `ConceptNet`. These topical words are extracted by applying latent dirichlet allocation (`LDA`) (Blei et al., 2003) to mine the latent topics T from all the documents. It is defined using equation 13 as joint probability distribution considering a topic mixture φ with a set of M event topics $T = \{t_1, t_2, ..., t_{|T|}\}$ and words $W = \{w_i, w_2, ..., w_{|W|}\}$, where η, and α are model parameters.

$$p(\varphi, T, W|\eta, \alpha) = p(\varphi|\eta) \prod_{i=1}^{M} p(t_i|\varphi) p(w_i|t_i, \alpha) \tag{13}$$

The words extracted from the top-100 topics are supplied to identify relevant commonsense knowledge. Subsequently, we establish relationships within the one-hop paths of all matched concepts to reconstruct the commonsense subgraph $CN = (C, Y)$, where $c_i \in C$ represents the concept set composed of all matched words, and $y_i \in Y$ represents the edges between these concepts. Finally, we reteive the relation in form of assertions with their confidence scores and stored in list CS. Relations with specific semantics, such as *UsedFor* and *HasPrerequisite*, tend to connect common words and topics, while more general relations, such as *Synonym* and *RelatedTo*, connect rarer words. The assertions with confidence scores less than 1 are removed to reduce the ineffective relations. In the cases where relation types often co-occur, such as *HasPrerequisite, Has-Subevent*, and *HasFirstSubevent*, which indicate ambiguity (Becker et al., 2019), we consider the relation associated with the highest confidence score.

3.5.2. Relevant Commonsense-Based Knowledge Extraction

Moreover, the selection of adequate and relevant commonsense-enriched assertions that provide contextually and semantically rich knowledge surrounding the check-worthy patterns list CP to the model is carried out in twofold: (i) collect all potentially relevant and semantically similar assertions from list CS. For this purpose, we enrich patterns in CP and assertions using the commonsense-based, publicly available pre-trained word embedding model `ConceptNet Numberbatch`, a continuous word representation to preserve semantic relatedness with the surrounding context. A pattern cp_i and an assertion cs_j are represented as vectors $[nb_1^{cp_i}, nb_2^{cp_i}, nb_3^{cp_i}]$ and $[nb_1^{cs_j}, nb_2^{cs_j}, nb_3^{cs_j}]$, respectively, where $nb_i^{cp_i}, nb_i^{cs_j} \in \mathcal{S}^{d'}$, $cp_i \in CP$, and $cs_j \in CS$; $\mathcal{S}^{d'}$ is a d'-dimensional word embedding vector. The cosine similarity is used to measure the relatedness of the assertion. Each assertion is assigned by a scalar value $a(cs_j)$, the largest similarity score of that assertion corresponding to the patterns defined in equation 14.

$$a(cs_j) = \max_{1 \le i \le |CP|} \left\{ \frac{cp_i \cdot cs_j}{\|cp_i\| \cdot \|cs_j\|} \right\} \tag{14}$$

(ii) Selecting the most relevant and effective assertions is carried out by measuring their strength of the relatedness $rel(cs_j)$ using equation 15, where $a(cs_j)$ is an assigned similarity score, and cf is the given confidence score of the assertion. The assertions having a larger relatedness score, the more relevant and effective commonsense-based knowledge they provide to the model; therefore, we sort the assertions with their relatedness score and use them for detecting tasks.

$$rel(cs_j) = a(cs_j) \times cf \tag{15}$$

In this way, we leverage the commonsense-enriched and contextual learned representation in our approach, which considers the commonsense meaning of the words and utilizes semantic learning to cover the common theme between patterns and assertions using the `ConceptNet Numberbatch` embedding, which includes lexical and world knowledge from different sources and utilize semantic spaces on intrinsic evaluations of word relatedness and increases the commonsense understanding effectiveness.

3.6. Fusion of Features

This section considers generating the features derived from posts using the top-k rumor-based patterns, relevant commonsense-based background knowledge, and embedding models. We apply two embedding models to preserve semantic richness in two different aspects: `Glove` trained over $\mathbb{X}$ (formerly `Twitter`) data and `ConceptNet Numberbatch` for commonsense-based understanding. The fusion of these features is carried out as follows:

In the first step, the semantic vectorizations of the top-k patterns and 3-grams of input texts are performed using a pre-trained word embedding model, `Glove`. The semantic vector of a pattern cp_i is represented as $[v_1^{cp_i}, v_2^{cp_i}, v_3^{cp_i}]$, where $v_i^{cp_i} \in \mathcal{R}^d$ and $cp_i \in CP$; $\mathcal{R}^d$ is a d-dimensional word embedding vector. We employ a mean-pooling operator to aggregate information from the embedded representation. It is formulated as $\overline{v}_{cp_i} = mean(v_1^{cp_i}, v_2^{cp_i}, v_3^{cp_i})$, and a further semantic vector of a pattern cp_i is represented in a matrix of $[\overline{v}_{cp_i} \times d]$. Moreover, we split each input text X_j into 3-grams and apply the same embedding model to preserve the semantic relatedness with the patterns. For example, if a text has m numbers of 3-grams, its semantic vector is represented as a matrix of $[\overline{v}_{n_{j_1}} \times d \quad \overline{v}_{n_{j_2}} \times d \quad \dots \quad \overline{v}_{n_{j_m}} \times d]_{1 \times m}$, where $n_{j_m} \in X_j$ and d is the dimension of the word embedding vector. A pattern cp_i is matched with all the n-grams of X_j to calculate mean similarity scores f_{cp_i} using equation 16, where $Cosine(cp_i, X_{j_m})$ is a similarity score of a pattern cp_i and an n-gram.

$$f_{cp_i} = \frac{\sum_1^m Cosine(cp_i, n_{j_m})}{m} \tag{16}$$

This step is repeated for all the k patterns to obtain the similarity score with n-grams of an input text. Thus, the rumor-based patterns feature vector $\mathcal{F}_{cp}$ for input texts X_j is obtained as defined in equation 17.

$$\mathcal{F}_{cp} = [f_{cp_1}, f_{cp_2}, ..., f_{cp_k}] \tag{17}$$

The second step generates the commonsense-based feature vector by providing semantic enrichment to the top-l relevant assertions and 3-grams of input texts using a commonsense-based d'-dimensional pre-trained word embedding model `ConceptNet Numberbatch`. The top-l relevant assertions have higher relatedness scores and are considered by the empirically determined threshold. Similarly to the above step, we semantic vectorize the the top-l assertions and 3-grams of input texts to calculate their similarity scores. These relevant commonsense assertions help to bring the commonsense knowledge close to contextually and semantically commonsense understanding related to utterances and topics. An assertion cs_k is matched with all the n-grams of X_j to calculate mean similarity scores f_{cs_k} by following the same step described above, where $Cosine(cs_k, X_{j_m})$ is a similarity score of an assertion cs_k and an n-gram; further, this step is repeated for all l assertions. Therefore, the relevant commonsense-based knowledge feature vector $\mathcal{F}_{cs}$ for input texts X_j is obtained as defined in equation 18.

$$\mathcal{F}_{cs} = [f_{cs_1}, f_{cs_2}, ..., f_{cs_l}] \tag{18}$$

The third step represents the commonsense score, which measures the commonsense knowledge of X_j by fetching it towards the knowledge of relevant commonsense-based assertion. For this purpose, semantic vectors of the top-l

assertions and input texts are generated using a pre-trained word embedding model `ConceptNet Numberbatch`. After that, the mean embedding vectors of X_j and l assertions, represented as $g_{(X_j)} = [nb]_{1\times d'}$ and $g_{(cs)} = [g_{\overline{cs}}]_{l\times d'}$, respectively, are obtained; $\mathcal{S}^{d'}$ is a d'-dimensional word embedding vector. Their vector multiplication yields the commonsense score-based feature vector $\mathcal{F}_s$ for input texts X_j, as defined in equation 19.

$$\mathcal{F}_s = g_{(X_j)} \cdot g_{(cs)}^{\mathsf{T}} \tag{19}$$

The final feature vector is obtained by concatenating the feature vectors of rumor-based patterns $\mathcal{F}_{cp}$, commonsense-based knowledge $\mathcal{F}_{cs}$, and commonsense score-based $\mathcal{F}_s$, as defined in equation 20.

$$\mathcal{FV} = \mathcal{F}_{cp} \oplus \mathcal{F}_{cs} \oplus \mathcal{F}_s \tag{20}$$

4. Experimental Setup and Results

This section presents a detailed description of the datasets, experimental setup, evaluation and comparative analysis results, and sensitivity analysis of the proposed approach. We begin by detailing three 𝕏 datasets and then discuss the evaluation metrics. Thereafter, we analyze the experimental results over the datasets and compare the performance of the proposed approach with the baselines and state-of-the-art approaches. Finally, an ablation study, sensitivity analysis, and a case study are presented for a deeper analysis of the proposed approach. The experiments aims to answer the following questions: (1) Is `CARES` able to achieve satisfactory performance for rumor detection compared to state-of-the-art approaches? (2) How effective is leveraging posts and reactions jointly in improving the detection performance of `CARES`? (3) How do external lexicon-based and commonsense-based knowledge contribute to the performance of `CARES`?

4.1. Datasets

We have used three publicly available real-world datasets: Pheme, Twitter15, and Twitter16 (Zubiaga et al., 2017; Ma et al., 2017). The Pheme dataset contains 5802 source posts with their reactions and metadata of five breaking news events, where source posts are annotated as rumors or non-rumors. Twitter15 and Twitter16 datasets have both a political and a more general context, annotated with finer-grained classes from which we consider true and false labels containing 742 and 412 source posts, respectively, as considered in the studies (Lu & Li, 2020; Pelrine et al., 2021). Since the original Twitter15 and Twitter16 datasets do not have publicly available reaction texts, we retrieve them by crawling the given reaction IDs via the `Twitter` API[4]. We have ignored the deleted and missing reactions during the crawling process due to their negligible amount and removed reactions, which do not have textual data in preprocessing stage. The reasons for selecting these two datasets are: (i) their labels are accurate since they are manually checked and annotated by experts, and (ii) they include social context, which helps to create contextual and semantics representation. They involve a post-and-reaction structure that is conducive to the model's design, as the interactions between posts and their associated reactions provide rich data for identifying sentiment, skepticism, and emotion patterns, which are central aspects of the proposed approach for detecting rumors. A detailed statistics of the aforesaid datasets are presented in Tables 1 and 2.

4.2. Evaluation Metrics

We have used three standard evaluation metrics: *precision*, *recall*, and *f1-score* as defined in equations 21, 22, and 23, respectively to assess the performance of the proposed approach. These metrics are defined in terms of true-positive (tp), i.e., the total number of rumors correctly predicted as rumors; false-positive (fp), i.e., the total number of non-rumors predicted as rumors; and false-negative (fn), i.e., the total number of rumors incorrectly predicted as non-rumors. *Precision* assesses the correctness, *recall* assesses the completeness of coverage of the classifier, and the *f1-score* provides a way to combine the contribution of both *precision* and *recall* evenly using harmonic mean.

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

[4] `https://dev.twitter.com/rest/public`

Table 1: Statistics of the Pheme dataset

Events Name	Rumors	Non-Rumors	Source Posts	Reactions	Total
Charlie Hebdo	458	1,621	2,079	36,189	38,268
Sydney Siege	522	699	1,221	22,775	23,996
Ferguson	284	859	1,143	23,032	24,175
Ottawa Shooting	470	420	890	11,394	12,284
Germanwings Crash	238	231	469	4,020	4,489
Total	1,972	3,830	5,802	97,410	103,212

Table 2: Statistics of the Twitter15 and Twitter16 datasets

Details	True Rumors	False Rumors	Source Posts	Reactions	Total
Twitter15	372	370	742	16,324	17,066
Twitte16	207	205	412	6,592	7,004

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{23}$$

4.3. Experimental Setup

All textual contents of the datasets undergo preprocessing phase, including noise filtering, cleaning, contraction, tokenization, and normalization. We employ `pycontractions`[5], a `Python` library for expanding contractions, to improve text clarity, maintaining the original meaning of the text. It is crucial in our research as it ensures that the text is accurately represented which is essential for the subsequent analysis. Non-alphabetic words, including numbers and symbols, and one-lettered words are removed from the text. The punctuation and emoticons such as ?, !, and : (′ are included since they play a significant role in examining the writing style of users. Stop-words and capital letters are essential to understanding the context of posts. The capital letters show emphasis, whereas stop-words have intact syntactic sense and are retained to form rumor-based patterns. Likewise, emojis are also frequently used to convey emotion in informal texts; therefore, we have included them by converting them into text using the `Python emoji` module. We also normalized datasets to avoid any event-specific biasedness and Twitter-names. For this purpose, we replace the Twitter-specific tokens, such as *hashtags, urls, retweets*, and *mentions*, with the tags $< hashtags >$, $< urls >$, $< retweet >$, and $< mentions >$, respectively. Furthermore, we generated graphs for all three datasets in which the number of nodes and edges (V, E) for Pheme, Twitter15, and Twitter16 are $(3076, 219153)$, $(2147, 70450)$, and $(1472, 41227)$, respectively. We use the `NetworkX` package of `Python` to generate graphs and implement its methods.

The evaluation is underpinned by an efficient hold-out validation strategy. The datasets are split into training and testing sets in a proportion of $80 : 20$, ensuring that the model is trained on a sufficiently large portion of the data while still having a substantial amount of data for testing. We used `Python` programming language to implement the proposed approach, baselines, and state-of-the-art methods for empirical study. Referring to studies (Sicilia et al., 2018; Ajao et al., 2019), our evaluation is a comprehensive exploration of classification algorithms, each representing a different learning paradigm. These include a support vector machine (`SVM`) as a kernel machine, a gradient boosting (`GB`) as an ensemble technique, conditional random fields (`CRF`) as a conditional sequential classifier, and logistic regression (`LR`) as a non-sequential linear model. In our experiment for all three datasets, these classifiers are implemented at their default hyper-parameter settings in `Scikit-learn`[6], a widely used machine learning library in `Python`. The hyper-parameter settings of the baselines and state-of-the-art methods are followed as recommended in their

[5] `https://pypi.org/project/pycontractions/`
[6] `https://scikit-learn.org/stable/`

respective papers. CRF classifier is implemented using the PyStruct[7] (Sutton & McCallum, 2012; Müller & Behnke, 2014). We have employed two publicly available embedding models; Glove used a 100-dimensional word-level pre-trained embedding over the Twitter dataset with 27 billion tokens, whereas ConceptNet Numberbatch used a 300-dimensional embedding vector. Those words that do not exist in the pre-trained word vectors are ignored. The feature vector of the proposed approach is $k + l + l$, where k is the size of the rumor-based pattern features, and l is the size of the commonsense-based assertions features. It may be noted that 3-word patterns are used in this work since this setting experimentally produced the best results.

4.4. Evaluation Results and Comparative Analysis

This section aims to address the first previously-mentioned evaluation question by presenting comparative performance evaluation results of CARES with the following baselines and state-of-the-art (SOTA) approaches for rumor detection.

DTC(Castillo et al., 2011): This approach analyzed microblog postings on trending topics to classify them as credible or non-credible. It combined hand-crafted statistical features from the posts and reactions for credibility assessment.

PTK(Ma et al., 2017): This approach proposed a kernel-based method called propagation tree kernel for rumor detection. It used high-order patterns of rumors by evaluating the similarities between their propagation tree structures. It captured the salient substructures in the propagation trees, considering different propagation paths from source posts to the roots of subtrees, which represent the context of transmission.

Zubiaga et al. (2017): This approach used linear-chain conditional random fields to leverage the sequential dynamics of context preceding a social media post. The content and user social features in this approach are intended to capture textual feature and the role of user behavior in rumor detection.

Abulaish et al. (2019): This approach used a graph-based approach to model user-generated content on social media for rumor detection. It used sentimental terms in social media posts, such as anxiety and doubt, to detect rumors.

Ajao et al. (2019): This approach proposed sentiment-aware misinformation detection, considering the sentiments of social media posts. It used a ratio of negative and positive emotions to detect rumors.

Lathiya et al. (2020): This approach considered the sequential dynamics of social media posts and the usefulness of users' homophilic connections as a feature for rumor detection. The content and user social features, such as rumor-ratio and homophily features, are used with linear-chain conditional random fields.

Haque & Abulaish (2022): This approach used a graph-based approach that leverages posts and reactions to extract underlying rumorous patterns for rumor detection. It learned contextual and semantic features using a word embedding model to enhance the detection task.

Rezaei et al. (2023): This approach considered the most useful linguistic features, such as textual, sentiment, semantic, and readability for detecting rumors.

Haque & Abulaish (2023): This approach introduced an emotion-enhanced and psycholinguistic features-based approach for detecting rumors on social media that combines emotion-related aspects, psycholinguistic features, and a word embedding model. Emotional and psycholinguistic features are extracted from posts and comments to enhance the approach and detect rumors more effectively.

A summary of comparative results in terms of *precision* (P), *recall* (R), and *f1-score* (F1) is presented in Tables 3, 4, and 5 over three datasets. It can be observed from the results that CARES achieves the best performance compared to all SOTA approaches, validating its effectiveness in detecting rumors. Its best result is obtained using GB, which scored the highest value of precision and f1-score, whereas SVM achieved the highest recall value over all datasets. The experimental results also demonstrate that methods (Abulaish et al., 2019) and (Zubiaga et al., 2017) are underperforming

[7] https://pystruct.github.io/

Table 3: Comparative performance evaluation results of CARES with state-of-the-art approaches on the Pheme dataset

Approach	GB			SVM			CRF			LR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DTC	73.48	70.44	71.92	70.37	68.72	69.53	77.81	67.68	72.39	76.27	58.56	66.25
PTK	74.09	67.59	70.69	73.94	66.38	69.95	70.79	68.62	69.68	69.00	65.05	66.96
Zubiaga et al. (2017)	52.45	54.18	53.30	36.61	44.79	40.28	69.21	54.60	61.04	33.13	45.68	38.40
Abulaish et al. (2019)	56.27	55.41	55.83	41.28	45.61	43.33	64.64	60.10	62.29	40.78	41.91	41.34
Ajao et al. (2019)	84.82	85.15	84.98	86.67	85.81	86.23	86.43	84.63	85.52	83.85	85.12	84.48
Lathiya et al. (2020)	65.23	61.35	63.23	60.87	52.65	56.46	66.53	64.29	65.39	48.97	51.27	50.09
Haque & Abulaish (2022)	92.75	94.22	93.47	93.56	91.75	92.64	91.36	88.91	90.11	91.15	85.94	88.46
Rezaei et al. (2023)	91.62	87.85	89.70	88.21	87.02	87.60	86.28	85.74	86.00	88.34	84.18	86.19
Haque & Abulaish (2023)	87.47	93.26	90.27	88.56	92.42	90.44	89.08	90.57	89.81	90.35	86.52	88.39
CARES	98.56	95.92	97.22	95.88	96.58	96.30	95.64	93.87	94.74	95.39	92.90	94.12

Table 4: Comparative performance evaluation results of CARES with state-of-the-art approaches on Twitter15 dataset

Approach	GB			SVM			CRF			LR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DTC	70.23	72.71	71.44	69.34	71.63	70.46	72.16	66.27	69.09	66.33	70.46	68.35
PTK	73.54	76.20	74.84	76.45	73.05	74.71	71.86	72.00	71.92	74.52	69.26	71.80
Zubiaga et al. (2017)	64.77	50.62	56.82	42.66	34.18	38.13	45.24	59.53	51.42	36.61	50.12	42.31
Abulaish et al. (2019)	60.11	48.08	53.43	47.13	40.70	43.67	62.39	57.06	59.60	49.77	38.42	43.36
Ajao et al. (2019)	81.56	73.04	77.06	79.25	71.36	75.10	79.64	73.17	76.26	80.41	68.73	74.11
Lathiya et al. (2020)	67.18	60.42	63.62	58.63	51.54	54.85	64.25	59.72	61.90	49.54	52.26	50.86
Haque & Abulaish (2022)	87.28	85.05	86.15	83.09	84.26	83.67	82.10	82.77	82.43	84.38	79.84	82.04
Rezaei et al. (2023)	79.54	83.26	81.35	80.71	80.00	80.35	79.36	81.22	80.28	81.56	77.68	79.57
Haque & Abulaish (2023)	88.37	81.25	84.66	85.17	79.66	82.32	81.50	79.81	80.64	82.91	77.63	80.18
CARES	92.84	89.30	91.04	90.56	89.34	89.94	90.29	86.63	88.42	85.32	87.59	86.44

over all datasets, suggesting their inability to extract underlying information and higher-order representations from rumor data. When (Lathiya et al., 2020) used homophilic features and rumor-ratio, the performance improved. Among these baselines, PTK gained relatively good results with linguistic and structural features from a time series structure in the propagation tree. Ajao et al. (2019) used emotion as a main feature that achieved better performance, indicating the significance of incorporating emotions. Haque & Abulaish (2022) leveraged posts' and reactions' structural and underlying information based on contextual and semantic learning and achieved better results. Among the SOTA approaches that use structure and emotion-related information, CARES has the ability to learn extensive knowledge from external emotion-based and commonsense-based knowledge sources. In addition, the semantic embedding methods allow CARES to learn the semantic context better and leverage the commonsense understanding to outperform the statistical and contextual learning-based method for rumor detection. Compared with the best results of the SOTA approaches, CARES shows improvements in terms of *f1-score* on the Pheme, Twitter15, and Twitter16 datasets by a margin of 3.75%, 4.89%, and 8.16%, respectively. It indicates that CARES incorporates better-enhancing processes in check-worthy pattern selection and effectively leverages commonsense-based knowledge to expand the features that help boost detection tasks.

Table 5: Comparative performance evaluation results of CARES with state-of-the-art approaches on Twitter16 dataset

Approach	GB			SVM			CRF			LR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DTC	73.22	67.67	70.33	70.37	67.95	69.13	67.46	72.85	70.05	69.76	62.15	65.73
PTK	71.48	75.72	73.53	71.66	75.03	73.30	73.59	66.78	70.02	65.45	69.04	67.19
Zubiaga et al. (2017)	63.26	49.00	55.22	38.14	44.37	41.02	50.26	58.44	54.03	44.38	51.30	47.58
Abulaish et al. (2019)	58.71	46.05	51.61	51.22	41.20	45.66	61.64	56.38	58.89	44.27	52.32	47.95
Ajao et al. (2019)	79.25	73.20	76.10	78.17	72.68	75.32	77.70	65.80	71.25	78.11	65.74	71.39
Lathiya et al. (2020)	65.33	60.00	62.55	62.66	53.59	57.77	60.55	63.98	62.21	50.85	52.11	51.47
Haque & Abulaish (2022)	84.70	77.61	81.00	80.24	76.55	78.35	79.21	74.43	76.74	73.32	74.17	73.74
Rezaei et al. (2023)	77.71	76.37	77.03	75.42	76.80	76.11	75.27	70.59	72.85	74.39	69.82	72.02
Haque & Abulaish (2023)	82.10	77.76	79.85	80.45	76.38	78.36	78.17	71.49	74.68	77.00	73.44	75.17
CARES	90.52	87.84	89.16	89.81	88.06	88.92	87.35	82.00	84.59	83.33	84.37	83.84

4.4.1. Ablation Study

In this section, we conduct the ablation study to systematically validate the relative importance of every module of CARES and to address the previously-mentioned questions (2) and (3). The ablation study is performed by excluding each component from the model to check its associated effectiveness, as described below:

w/o reactions: In this method, reactions are removed, and only source posts are considered to design the model and evaluate the performance of the proposed approach for rumor detection.

w/o emo-lex-based knowledge: In this method, both source posts and reactions are considered to design the models without providing emotion lexicons-based knowledge for pattern selection.

w/o skeptical-based knowledge: In this method, source posts and reactions are considered to design the models without providing skeptical and inquiry signal-based knowledge for pattern selection.

w/o hedge lexicon-based knowledge: In this method, both source posts and reactions are considered to design the models without providing hedge lexicon-based knowledge for pattern selection.

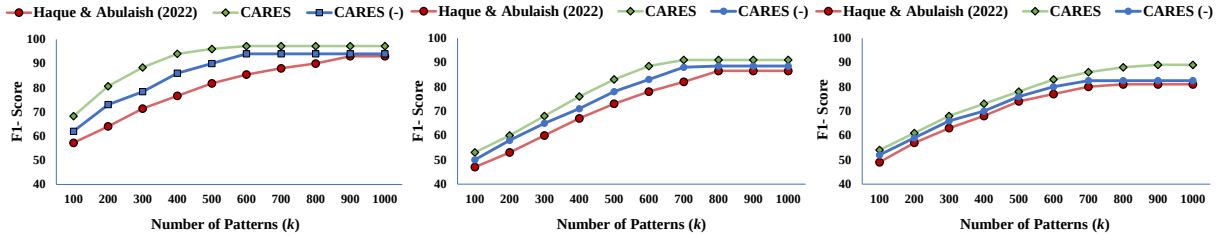
w/o lexicons-based knowledge: In this method, both source posts and reactions are considered to design the models without providing any lexicons-based knowledge during pattern selection.

w/o commonsense-based knowledge: In this method, both source posts and reactions are considered to design the models without providing commonsense-based knowledge features to classifiers.

The experimental results are given in Table 6. It can be observed that removing the reactions affects performance significantly, and in the best case, the *f1-score* drops 22.85%, 24.40%, and 21.75% on the Pheme, Twitter15, and Twitter16 datasets, respectively. CARES performance enhances and generally performs better when reactions are used since they can effectively reflect the process and capture the dynamics of social interactions on social media. Thus, they more accurately depict the characteristics of responsive and reactive users. Next, we evaluate the influence of the external lexicon-based knowledge. Table 6 also shows that not utilizing lexicon-based knowledge during pattern selection brings declines in performance, which exhibits the impact and significance of responsive and reactive aspects in the form of emotional reactions and inquisitive responses of social media users, providing complementary effects for determining the veracity of rumors. Moreover, incorporating the inquiry aspects, skeptical behavior, and use of hedge words boosts the performance marginally, whereas incorporating emotion-based aspects improves the performance significantly. The commonsense-based module also shows its significance; removing it affects performance, and in the best case, the *f1-score* drops 3.24%, 2.74%, and 6.85% on the Pheme, Twitter15, and Twitter16 datasets, respectively. Therefore, each module of the proposed approach is indispensable, enabling CARES to outperform baseline and state-of-the-art approaches remarkably.

Table 6: Ablation study results in terms of f1-score with different settings of CARES on Pheme, Twitter15, and Twitter16 datasets

Approach	Pheme				Twitter15				Twitter16			
	GB	SVM	CRF	LR	GB	SVM	CRF	LR	GB	SVM	CRF	LR
w/o reactions	74.37	72.63	66.51	68.17	66.64	65.02	61.12	65.55	67.41	65.73	62.00	65.28
w/o emo-lex-based knowledge	79.08	77.31	75.79	74.45	73.70	71.84	69.61	71.44	72.26	73.93	69.02	71.81
w/o skeptical-based knowledge	88.57	89.22	85.62	86.98	79.60	78.00	75.95	74.16	77.39	77.18	75.83	74.51
w/o hedge lexicon-based knowledge	94.66	93.23	92.00	90.90	90.53	87.21	85.46	86.50	83.39	81.18	78.83	79.51
w/o lexicons-based knowledge	77.40	74.05	74.41	72.17	71.61	70.11	67.05	69.10	70.68	70.49	66.36	69.24
w/o commonsense-based knowledge	93.98	93.16	91.80	90.58	88.30	85.54	84.55	84.09	82.31	80.28	75.78	78.00



(a) Pheme Dataset (b) Twitter15 Dataset (c) Twitter16 Dataset

Figure 3: Comparative stability analysis of CARES, CARES (−), and (Haque & Abulaish, 2022) at different values of k

4.4.2. *Sensitivity Analysis*

Sensitivity analysis is conducted to quantify the impact of uncertainty on the system's outcome and to know the confidence level of the parameters. In our experiment, we adhere to the default hyper-parameter settings of the classifiers for all three datasets. The performance of CARES is based on rich contextualized-affective top-k patterns where the value of k determines its stability and sensitivity. In this section, we test the sensitivity of CARES and its variants for all three datasets concerning the impact of top-k patterns. Figure 3 illustrates the result of CARES and its variants for all three datasets with different values of k. The performance of CARES increases with an increase in the value of k. It is also evident from Figure 3 that all three methods stabilize with increasing values of k, but the proposed approach stabilizes faster with a lesser number of patterns.

4.4.3. *Coverage and Effectiveness Analysis*

This section examines the effectiveness and coverage of the proposed approach. The patterns described in section 3.3 are contextualized affective patterns that have the advantage of possessing high coverage due to the process involved in their construction. High coverage exhibits stability regarding their usefulness in rumor detection. Figure 3 empirically shows how many top-k patterns are effectively utilized to gain stability in the proposed approach that contributed to achieving performance with and without incorporating commonsense-based knowledge. It can be observed from Figure 3 that CARES (−), a variant of the proposed approach without incorporation of commonsense-based knowledge, gets optimal performance and stabilizes with the minimum number of patterns compared to (Haque & Abulaish, 2022), which performed best among SOTA approaches. CARES outperforms and stabilizes faster, but even CARES (−) achieved better performance and faster stability than (Haque & Abulaish, 2022) because of contextualized affective patterns, which have incorporated rich lexicon-based knowledge. We further analyze the coverage, effectiveness, and consistency of the proposed approach, for which experiments are conducted on the Pheme dataset by taking it event-wise. The Pheme dataset has 5 events, out of which we have used 4 for training and 1 for testing. For example, the result under *Charlie Hebdo* signifies that the model has used this event as testing, and the remaining 4 are used for training. The results are shown in Table 7 in terms of the *f1-score*, that are consistent across events and in line with the overall performance scores. The CARES outperforms all SOTA approaches. Compared to the best result of SOTA, the model improves the f1-scores from a range of 3.44% to 12.03%. The best performance of SOTA is achieved by (Haque

Table 7: Event-wise comparative performance evaluation results of CARES with state-of-the-art approaches on the Pheme dataset

Approach	Germanwings Crash				Ottawa Shooting				Sydney Siege				Charlie Hebdo				Ferguson			
	GB	SVM	CRF	LR	GB	SVM	CRF	LR	GB	SVM	CRF	LR	GB	SVM	CRF	LR	GB	SVM	CRF	LR
DTC	66.32	64.41	64.13	60.07	61.61	60.03	62.81	56.68	64.04	64.09	62.35	59.03	61.30	60.00	60.96	57.23	60.35	61.68	60.34	56.00
PTK	67.66	63.04	62.76	64.32	60.12	59.31	59.62	56.08	63.21	61.20	60.51	57.29	62.70	61.43	60.56	61.55	58.96	60.85	56.51	58.44
Zubiaga et al. (2017)	53.63	34.09	63.92	33.52	47.17	45.83	52.21	42.54	42.23	33.53	46.77	33.11	52.55	45.96	70.09	45.28	64.52	48.00	70.10	47.52
Abulaish et al. (2019)	62.38	61.21	65.00	54.41	55.10	55.63	54.19	50.53	64.28	60.76	59.51	54.29	59.70	61.13	60.35	53.26	55.65	54.34	52.31	50.33
Ajao et al. (2019)	76.07	75.62	74.95	69.27	67.90	66.78	63.39	62.38	70.25	68.88	69.88	68.54	69.11	68.21	64.33	61.00	75.56	74.67	74.25	71.00
Lathiya et al. (2020)	55.33	41.28	60.56	40.25	59.24	56.17	62.91	50.87	48.89	38.78	47.51	36.01	67.20	53.24	70.46	52.34	70.00	58.26	71.28	50.55
Haque & Abulaish (2022)	83.36	82.53	81.63	77.95	76.03	75.35	74.00	69.55	81.13	80.61	79.32	75.22	80.45	79.26	79.10	73.47	72.63	71.56	71.84	66.28
Rezaei et al. (2023)	77.52	76.31	75.64	71.50	70.85	71.11	69.66	64.72	72.48	71.70	73.10	70.48	73.24	72.70	70.49	69.68	74.70	74.00	73.26	72.25
Haque & Abulaish (2023)	78.21	76.84	75.20	73.29	72.58	73.16	72.76	66.36	75.29	74.61	76.58	73.04	75.45	74.61	74.77	74.18	76.82	75.79	75.50	73.35
CARES	93.15	91.41	90.51	88.33	88.06	83.84	81.90	79.62	85.59	85.06	84.22	80.64	83.89	83.19	81.45	79.83	80.75	80.72	78.48	76.71

& Abulaish, 2022) for all events except *Ferguson*. It can be said that the reason for the consistent performance of the proposed approach is that the graph-based approach covers the topical words and the writing style of social media users. Also, the rumor-based patterns contribute to better performance since they are enriched and generic by using external lexicon-based knowledge. Commonsense-based knowledge also contributes to the consistency and effectiveness of the proposed approach by providing general facts as semantic evidence to the detection model. Finally, word embedding models cover the higher dimensional semantic space.

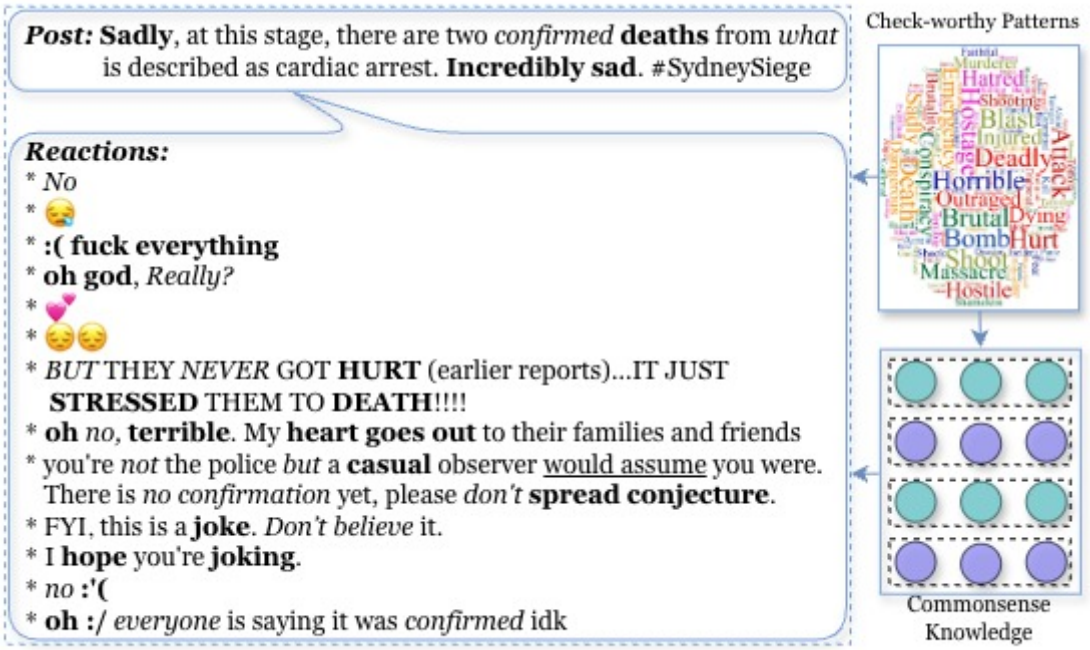


Figure 4: A sample case study of rumor detection using CARES

4.4.4. Case Study

We conduct a case study to illustrate the process of rumor detection by CARES. We select a random sample of a post and associated reactions from the Pheme dataset. After the rumor gets published, users start responding; some

respond with emotion, while others start questioning and verifying. We can observe negative emotions, sentiments, and debunking responses ("FYI, this is a joke. Don't believe it") in the reactions. There are very few positive sentiments (heart emoji). Figure4 shows the patterns of negative emotion (bold text), skepticism (italic text), and hedge words (underlined text) in the post and reactions. Words of top-100 check-worthy rumor patterns are represented in the word-cloud. These patterns are matched with posts and reactions to calculate their semantic similarity. This process involves generating rumor-based feature vectors, which are numerical representations of the semantic similarity between the patterns and the posts/reactions. It also leverages a commonsense understanding encompassing contextual utterances, related topics, and semantics. The visualization in Figure4 shows that contextually enriched rumor-based patterns capture information from posts and reactions to assess their credibility.

5. Conclusion and Future Work

In this paper, we have presented a rumor detection approach, CARES, which integrates graph-based contextual learning, lexicon-based knowledge, commonsense-based background knowledge, and semantically enrichment using word embedding models. It incorporates posts and reactions to capture the context and identify the clues to construct effective rumor-based descriptors. It also leverages the underlying latent information and social media users' responsive and reactive behavior by investigating the inquisitive, skeptical, sentimental, and emotional clues using lexicon-based knowledge that captures implicit and explicit linguistic expressive information, significantly making the detection system more effective. We further leverage a commonsense knowledge graph to provide semantically commonsense-based background knowledge for the detection task. Extensive experiments on three datasets demonstrate that CARES remarkably improves the rumor detection task and outperforms the state-of-the-art approaches. In addition, the ablation study results demonstrate and explain the effectiveness and rationality of each module in CARES. The ablation study also highlights the significance of utilizing post-related reactions. It also reveals the contribution of commonsense-based background knowledge and demonstrates that it substantially enhances rumor detection tasks. We also perform sensitivity analysis and a case study to better understand rumor detection process by CARES. We believe CARES can be used in other social media text classification tasks, such as emotion, sentiment, and hate speech detection. We have considered 3-word patterns because long patterns are infrequent and contribute to diminishing performance gain or negatively impacting some datasets since long contextual information may introduce additional noises, e.g., polysemes expressing different meanings in different words of the same context. A more thorough investigation and generalization of this diminishing return phenomenon is an interesting future work. Furthermore, since rumors often focus on certain events, CARES may be expanded to investigate how to handle event-specific elements in order to improve performance even further.

References

Abulaish, M., Kumari, N., Fazil, M., & Singh, B. (2019). A graph-theoretic embedding-based approach for rumor detection in twitter. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019* (pp. 466–470). ACM. doi:10.1145/3350546.3352569.

Ajao, O., Bhowmik, D., & Zargari, S. (2019). Sentiment aware fake news detection on online social networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019* (pp. 2507–2511). IEEE. doi:10.1109/ICASSP.2019.8683170.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*, 211–236. doi:10.1257/jep.31.2.211.

Azzimonti, M., & Fernandes, M. (2023). Social media networks, fake news, and polarization. *European journal of political economy*, *76*, 102256. doi:10.3386/w24462.

Bai, N., Meng, F., Rui, X., & Wang, Z. (2021). Rumour detection based on graph convolutional neural net. *IEEE Access*, *9*, 21686–21693. doi:10.1109/ACCESS.2021.3050563.

Bauer, L., Wang, Y., & Bansal, M. (2018). Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018* (pp. 4220–4230). Association for Computational Linguistics. doi:10.18653/V1/D18-1454.

Becker, M., Staniek, M., Nastase, V., & Frank, A. (2019). Assessing the difficulty of classifying conceptnet relations in a multi-label classification setting. In *Proceedings of the Workshop on Meaning Relations between Phrases and Sentences, RELATIONS@IWCS 2019, Gothenburg, Sweden, May 23, 2019* (pp. 1–14). Association for Computational Linguistics. doi:10.18653/V1/W19-0801.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55. doi:https://doi.org/10.1016/j.ins.2019.05.035.

Brown, E. (2019). Online fake news is costing us $78 billion globally each year. *ZDNet, December*, *18*.

Cambria, E., Liu, Q., Decherchi, S., Xing, F., & Kwok, K. (2022). Senticnet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022* (pp. 3829–3839). European Language Resources Association.

Cambria, E., Poria, S., Bajpai, R., & Schuller, B. W. (2016). Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan* (pp. 2666–2677). ACL.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011* (pp. 675–684). ACM. doi:10.1145/1963405.1963500.

Connell, L., & Keane, M. T. (2004). What plausibly affects plausibility? concept coherence and distributional word coherence as factors influencing plausibility judgments. *Memory & cognition*, *32*, 185–197. doi:10.3758/BF03196851.

Fionda, V., & Pirrò, G. (2018). Fact checking via evidence patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden* (pp. 3755–3761). ijcai.org. doi:10.24963/IJCAI.2018/522.

Frenkel, S., Alba, D., & Zhong, R. (2020). Surge of virus misinformation stumps facebook and twitter. *The New York Times*, *8*.

Ghosh, M., Das, P., & Das, P. (2023). A comparative study of deterministic and stochastic dynamics of rumor propagation model with counter-rumor spreader. *Nonlinear Dynamics*, *111*, 16875–16894. doi:10.1007/s11071-023-08768-1.

Graves, D. (2018). Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*, . doi:10.60625/risj-nqnx-bg89.

Guo, H., & Yan, X. (2023). Dynamic modeling and simulation of rumor propagation based on the double refutation mechanism. *Information Sciences*, *630*, 385–402. doi:10.1016/J.INS.2022.10.095.

Haque, A., & Abulaish, M. (2022). A graph-based approach leveraging posts and reactions for detecting rumors on online social media. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, PACLIC 2022, Manila, Philippines, October 20-22, 2022* (pp. 533–544). Association for Computational Linguistics.

Haque, A., & Abulaish, M. (2023). An emotion-enriched and psycholinguistics features-based approach for rumor detection on online social media. In *Proceedings of the 11th International Workshop on Natural Language Processing for Social Media (SocialNLP), Co-located with IJCNLP-AACL, Bali, Indonesia, November 1-4, 2023* (pp. 28–37). Bali, Indonesia: Association for Computational Linguistics.

Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., & Zhou, M. (2021). Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021* (pp. 754–763). Association for Computational Linguistics. doi:10.18653/V1/2021.ACL-LONG.62.

Islam, J., Xiao, L., & Mercer, R. E. (2020). A lexicon-based approach for detecting hedges in informal text. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020* (pp. 3109–3113). European Language Resources Association.

Kim, G., & Ko, Y. (2021). Graph-based fake news detection using a summarization technique. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3276–3280).

Kochkina, E., Liakata, M., & Zubiaga, A. (2018). All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018* (pp. 3402–3413). Association for Computational Linguistics.

Lathiya, S., Dhobi, J. S., Zubiaga, A., Liakata, M., & Procter, R. (2020). Birds of a feather check together: Leveraging homophily for sequential rumour detection. *Online Soc. Networks Media*, *19*, 100097. doi:10.1016/J.OSNEM.2020.100097.

Li, W., Guo, C., Deng, Z., Liu, F., Wang, J., Guo, R., Wang, C., & Jin, Q. (2023a). Coevolution modeling of group behavior and opinion based on public opinion perception. *Knowl. Based Syst.*, *270*, 110547. doi:10.1016/J.KNOSYS.2023.110547.

Li, W., Guo, C., Liu, Y., Zhou, X., Jin, Q., & Xin, M. (2023b). Rumor source localization in social networks based on infection potential energy. *Information Sciences*, *634*, 172–188. doi:10.1016/J.INS.2023.03.098.

Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., & Ananiadou, S. (2023). Emotion detection for misinformation: A review. *CoRR*, *abs/2311.00671*. doi:10.48550/ARXIV.2311.00671.

Lu, Y., & Li, C. (2020). GCAN: graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020* (pp. 505–514). Association for Computational Linguistics. doi:10.18653/V1/2020.ACL-MAIN.48.

Luvembe, A. M., Li, W., Li, S., Liu, F., & Wu, X. (2024). CAF-ODNN: complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Inf. Process. Manag.*, *61*, 103653. doi:10.1016/J.IPM.2024.103653.

Ma, J., Gao, W., & Wong, K. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers* (pp. 708–717). Association for Computational Linguistics. doi:10.18653/V1/P17-1066.

Ma, J., Gao, W., & Wong, K. (2018). Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers* (pp. 1980–1989). Association for Computational Linguistics. doi:10.18653/V1/P18-1184.

McLennan, M. (2021). The global risks report 2021 16th edition. World Economic Forum Cologny, Switzerland.

Mohammad, S. M. (2018). Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Comput. Intell.*, *29*, 436–465. doi:10.1111/J.1467-8640.2012.00460.X.

Müller, A. C., & Behnke, S. (2014). Pystruct: learning structured prediction in python. *J. Mach. Learn. Res.*, *15*, 2055–2060. doi:10.5555/2627435.2670316.

Nguyen, V., Sugiyama, K., Nakov, P., & Kan, M. (2020). FANG: leveraging social context for fake news detection using graph representation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020* (pp. 1165–1174). ACM. doi:10.1145/3340531.3412046.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*, 175–220.

Pelrine, K., Danovitch, J., & Rabbany, R. (2021). The surprising performance of simple baselines for misinformation detection. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021* (pp. 3432–3441). ACM / IW3C2. doi:10.1145/3442381.3450111.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL* (pp. 1532–1543). ACL. doi:10.3115/V1/D14-1162.

Poria, S., Gelbukh, A. F., Cambria, E., Hussain, A., & Huang, G. (2014). Emosenticspace: A novel framework for affective common-sense reasoning. *Knowl. Based Syst.*, *69*, 108–123. doi:10.1016/J.KNOSYS.2014.06.011.

Rezaei, S., Kahani, M., Behkamal, B., & Jalayer, A. (2023). Early multi-class ensemble-based fake news detection using content features. *Soc. Netw. Anal. Min.*, *13*, 16. doi:10.1007/S13278-022-01019-Y.

Sharma, S., S, R., Arora, U., Akhtar, M. S., & Chakraborty, T. (2023). MEMEX: detecting explanatory evidence for memes via knowledge-enriched contextualization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023* (pp. 5272–5290). Association for Computational Linguistics. doi:10.18653/V1/2023.ACL-LONG.289.

Sheng, Q., Zhang, X., Cao, J., & Zhong, L. (2021). Integrating pattern- and fact-based fake news detection via model preference learning. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021* (pp. 1640–1650). ACM. doi:10.1145/3459637.3482440.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019* (pp. 395–405). ACM. doi:10.1145/3292500.3330935.

Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. *Expert Syst. Appl.*, *110*, 33–40. doi:10.1016/J.ESWA.2018.05.019.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA* (pp. 4444–4451). AAAI Press. doi:10.1609/AAAI.V31I1.11164.

Sun, M., Zhang, X., Zheng, J., & Ma, G. (2022). DDGCN: dual dynamic graph convolutional networks for rumor detection on social media. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022* (pp. 4611–4619). AAAI Press. doi:10.1609/AAAI.V36I4.20385.

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, *4*, 267–373. doi:10.1561/2200000013.

Tredici, M. D., & Fernández, R. (2020). Words are the window to the soul: Language-based user representations for fake news detection. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020* (pp. 5467–5479). International Committee on Computational Linguistics. doi:10.18653/V1/2020.COLING-MAIN.477.

Tu, K., Chen, C., Hou, C., Yuan, J., Li, J., & Yuan, X. (2021). Rumor2vec: A rumor detection framework with joint text and propagation structure representation learning. *Information Science*, *560*, 137–151. doi:10.1016/J.INS.2020.12.080.

Vaibhav, V., Annasamy, R. M., & Hovy, E. H. (2019). Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing, TextGraphs@EMNLP 2019, Hong Kong, November 4, 2019* (pp. 134–139). Association for Computational Linguistics. doi:10.18653/V1/D19-5316.

Vicario, M. D., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *CoRR*, *abs/1607.01032*.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*, 1146–1151. doi:10.1126/science.aap9559.

Wang, L., Zhang, C., Xu, H., Xu, Y., Xu, X., & Wang, S. (2023). Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023* (pp. 5696–5704). ACM. doi:10.1145/3581783.3613850.

Xie, Y., & Pu, P. (2021). How commonsense knowledge helps with natural language tasks: A survey of recent resources and methodologies. *CoRR*, *abs/2108.04674*.

Yang, S., Chen, C., Huang, H., & Chen, H. (2023). Entity-aware dual co-attention network for fake news detection. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023* (pp. 106–113). Association for Computational Linguistics. doi:10.18653/V1/2023.FINDINGS-EACL.7.

Zhang, H., Li, Y., Zhu, T., & Li, C. (2024). Commonsense-based adversarial learning framework for zero-shot stance detection. *Neurocomputing*, *563*, 126943. doi:10.1016/J.NEUCOM.2023.126943.

Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015* (pp. 1395–1405). ACM. doi:10.1145/2736277.2741637.

Zhong, P., Wang, D., & Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (pp. 165–176). Association for Computational Linguistics. doi:10.18653/V1/D19-1016.

Zhou, X., & Zafarani, R. (2021). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, *53*, 109:1–109:40. doi:10.1145/3395046.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, *51*, 32:1–32:36. doi:10.1145/3161603.

Zubiaga, A., Liakata, M., & Procter, R. (2017). Exploiting context for rumour detection in social media. In *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I* (pp. 109–123). Springer volume 10539 of *Lecture Notes in Computer Science*. doi:10.1007/978-3-319-67217-5_8.