PARALLAX: Leveraging Polarization Knowledge for Misinformation Detection

Demetris Paschalides[0000−0002−4475−4635], George Pallis[0000−0003−1815−5468], and Marios D. Dikaiakos[0000−0002−4350−6058]

University of Cyprus, Computer Science Department, Nicosia, Cyprus
{dpasch01, pallis, mdd}@ucy.ac.cy

Abstract. Recent techniques for the automated detection of online misinformation typically rely on ML models trained with features extracted from content analysis and/or general-purpose Knowledge Graphs (KGs). These techniques often fail to consider the interplay between misinformation and polarization. To bridge this gap, we introduce PARALLAX, a methodology that enhances misinformation detection by infusing polarization knowledge into existing classifiers. Polarization knowledge is represented in terms of Polarization Knowledge Graphs (PKG). PARALLAX constructs PKGs in an unsupervised way, and uses them to enrich articles with polarization knowledge. A Flexible Knowledge-aware Graph Neural Network (FlexKGNN) is trained on these enriched representations. We tested our methodology on three misinformation datasets, demonstrating that it achieves approximately a 15% improvement in performance over baseline classifiers and consistently outperforms other KGs, which typically reach baseline levels only.

Keywords: Polarization · Knowledge Graph · Misinformation Detection

1 Introduction

In recent years, misinformation has posed significant challenges to societies worldwide, with evident influence on events such as presidential elections [1], referendums, and most recently, the COVID-19 pandemic [6,13]. Moreover, the rise of Foreign Information Manipulation and Interference (FIMI) has added a new dimension to this challenge, as evidenced in the context of the Russo-Ukrainian war [20]. The rampant spread of misleading narratives extends beyond sowing confusion; it also exacerbates societal divisions, forging a complex relationship with the phenomenon of polarization [21]. On the one hand, polarization fosters an environment conducive to misinformation spread. The divisive nature of polarization, coupled with the human tendency for "confirmation bias," makes individuals more susceptible to false or misleading information, especially when it aligns with their existing viewpoints [29,36]. On the other hand, the proliferation of misinformation can also contribute to the escalation of polarization. In an environment with distorted information, individuals often retreat into "echo chambers" of similar views, reinforcing their beliefs, and perceiving dissenters as adversaries [29].

Existing methods to mitigate misinformation primarily focus on the development of Machine Learning (ML) models for fake news detection and/or the establishment of fact-checking initiatives [35], often overlooking the immediate connection between misinformation and polarization. To address this, we need to cope with the complexity of modeling, quantifying, and integrating polarization into misinformation detection algorithms [9].

In this paper, we aim at bridging this gap by proposing PARALLAX, a methodology and toolset for integrating polarization knowledge into existing misinformation classifiers, assessing its contribution on their classification performance. To do so, we address the challenges of: i) representing domain-specific polarization knowledge; ii) encoding news articles with their polarization information; and iii) effectively integrating article polarization cues into existing classifiers to enhance their accuracy. The key contributions of our work are:

- **Polarization Knowledge Graph (PKG)**: The definition of the PKG schema and data structure designed to capture polarization knowledge as a semantic graph of entities, fellowships, topics, and attitudes. To construct the PKG, we develop an unsupervised method to extract and model polarization information from news corpora (see Section 3).
- **Article-specific Polarization Encoding**: We introduce an approach to extract polarization knowledge on a single-article level, and encode this knowledge as a **micro-PKG**, a semantic graph that aligns with the PKG. The micro-PKG reflects the key actors and predicates identified in the article's content. To address content limitations at the article level, we enrich the micro-PKGs by i) incorporating additional polarization context from the PKG and ii) integrating PKG-derived embeddings (see Section 4).
- **Flexible Knowledge-aware Graph Neural Network (FlexKGNN)**: We design a Graph Neural Network to incorporate polarization knowledge encoded in micro-PKGs as a feature into existing misinformation classifiers, to enhance their classification performance. This is achieved by concatenating the feature vector of a given classifier with the internal micro-PKG representation to learn the relationship between polarization and misinformation for improved classification (see Section 5).
- **Evaluation Study and Dataset**: We evaluate PARALLAX on a manually curated COVID-19 misinformation dataset, along with two additional datasets used in prior literature [28]. We assess the contribution of polarization knowledge on the performance of two existing misinformation classifiers, comparing the results with alternative article-encoding methods. Our results reveal that our method outperform others, improving the accuracy of existing classifiers by ≈15%, highlighting the importance of incorporating polarization into misinformation detection (see Section 6).

2 Background and Related Work

Misinformation is defined as false information disseminated, either intentionally or inadvertently, from "unreliable" sources [35]. The content of fake news articles, often mimicking credible sources, distorts public perception by exploiting biases [23, 36], and intensifies societal polarization [1, 36].

Polarization refers to the phenomenon where social or political groups are fragmented into opposing factions that hold different and often conflicting beliefs and values [29]. These factions consist of interacting entities that hold diverse attitudes on various topics. Entities sharing similar beliefs tend to form cohesive fellowships, while conflicts emerge from their disagreements, thus forming fellowship dipoles. In such environment, misinformation aligning with in-group beliefs flourishes, while factual contradictory information is met with skepticism [21,36]. Consequently, polarization manifests across entities, groups, and topics, forming a complex multi-level phenomenon. We refer to entities, fellowships, dipoles, topics, and attitudes as "*polarization information*".

Polarization, rooted in the inter-group conflict theory [29], manifests during the process of "social categorization," where individuals align with groups, i.e fellowships, based on shared beliefs. This fosters in-group loyalty, distinguishing "us" from "them." During "social identification," individuals become members of those groups, a process amplified by cognitive biases [36]. In the concluding "social comparison" process, groups favorably contrast against others, often only considering their perspective, stereotyping out-groups negatively [11]. In such environment, misinformation aligning with in-group beliefs flourishes, while factual contradictory information is frequently met with skepticism or rejection [36].

2.1 Content-based Misinformation Detection

Content-based misinformation detection can be broadly categorized into style- and knowledge-based methods.

Style-based Methods focus on identifying distinctive writing styles in fake news through linguistic features [25, 28], such as modal words, punctuation, and casing [23]. They also emphasize the role of hyperpartisanship in the dissemination and distinction of fake news [35, 36]. Recent works have incorporated contextual embeddings from transformer models, which displayed their potency in accurately distinguishing between fake and real news content [24]. **Knowledge-based Methods** leverage external knowledge to enhance the understanding or verification of news articles [8,15,17,32,35]. Typically, they employ an established KG, such as Wikipedia[1], to serve as the factual world knowledge. A key step in these approaches, is encoding each article in the dataset using the pre-defined KG. Some approaches try to mimic human fact-checking, identifying claims within articles as Subject-Predicate-Object (SPO) triples, and verifying them against the KG [35]. More recent works try to enhance the content of each article, by identifying entities mentioned in each article's text, connecting them to the KG, and further enriching them with adjacent KG entities [8,15,17,32]. Certain works also integrate discussion topics [15] and entity relationships [8, 17, 32], extracted via tools like OpenIE [2], thus, forming a comprehensive heterogeneous graph for each article. To ascertain the veracity of articles, these methods

[1] http://wikipedia.org

use Graph Neural Networks (GNNs), classifying each article representation as originating from genuine or misleading information.

These methods, while broadening article context with external knowledge, have yet to integrate polarization knowledge in misinformation detection, despite its influence on misinformation spread and consumption [30, 36]. Furthermore, their outputs consist of GNNs that base their classification on internal article representations, overlooking existing classifiers. In this work, we address these limitations by introducing a method to represent article-level polarization knowledge. We propose a GNN that integrates this knowledge with existing classifiers, enhancing their classification performance and assessing the contribution of polarization on misinformation detection.

2.2 Representing Polarization Knowledge

Existing computational approaches which study, represent, and quantify different aspects of polarization in online social media typically focus on two broad directions: group- or topic-level polarization. **Group-oriented approaches** typically seek to identify polarized groups of users, and model inter-group polarization based on group segregation-level metrics [3, 9, 12]. **Topic-oriented approaches** typically apply NLP, topic modeling, and Deep Learning (DL) techniques to model, measure, and evaluate the polarized stance of distinct ideological user groups (e.g. Democrats and Republicans in the US) towards particular issues [14, 19]. In our work, we explore a different methodology for modeling polarization: we employ content analysis on a wide number of news articles to construct a semantic graph (i.e. the PKG) that represents the polarization landscape. This graph is extracted in an unsupervised manner from the narratives presented inside the articles, identifying key figures, events, and themes pivotal to public discourse, and capturing divisive and unifying attitudes among them. Our method integrates both group and topic aspects of polarization into the PKG for a comprehensive domain knowledge representation. This structured representation enables its integration with existing classifiers, aiding in evaluating its contribution on relevant tasks, including misinformation detection.

3 Polarization Modeling & Knowledge Extraction

3.1 Polarization Knowledge Graph and Schema

To capture domain-specific polarization knowledge effectively, we introduce the concept of Polarization Knowledge Graph (PKG), which is a structured representation of polarization information defined according to a Subject-Predicate-Object (SPO) schema shown in Fig. 1. The PKG schema is grounded on the inter-group conflict theory [29], comprising of four primary actors, namely *Entity*, *Fellowship*, *Dipole*, and *Topic*. An *Entity* is any individual or group that contributes to the polarization observed in a particular domain of interest; groups include organizations, countries, or religions. A *Topic* is a subject on which entities may hold opposing opinions. A *Fellowship* represents a cohesive sub-group

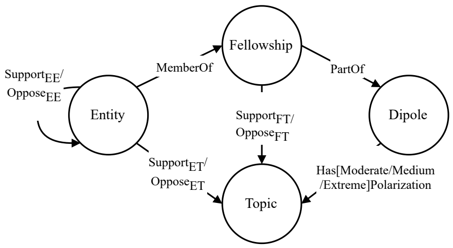


Fig. 1: Diagram of the Polarization Knowledge schema.

of entities with mutual supportive attitudes, while a *Dipole* comprises two fellowships manifesting opposing views or attitudes towards a particular topic. The PKG schema uses predicates to characterize supportive or opposing relationships between its actors: *SupportEE* and *OpposeEE* for entity interactions, *SupportET* and *OpposeET* for entity attitudes on topics, and *MemberOf* for entity associations to fellowships. It defines group attitudes on topics with *SupportFT* and *OpposeFT*, and fellowship conflicts within dipoles with the *PartOf* predicate. Degrees of polarization on topics are indicated by *HasModeratePolarization*, *HasMediumPolarization*, and *HasExtremePolarization*.

To construct a PKG, which reflects the context of a domain under study, we gather a *Supplementary Corpus* comprised of news articles representing the context surrounding this domain. The selection of articles for the corpus is guided by three critical parameters: *theme*, *region*, and *time-frame*. *Theme* refers to a collection of keywords that capture the domain's core concepts; for instance, when focusing on Coronavirus, relevant keywords might include "Coronavirus," "COVID-19," and "SARS-CoV-2." The *Region* parameter helps focus on a specific geographical area, like the United States, and *time-frame* defines a specific period under study through designated start and end dates. We deploy collectors that utilize the aforementioned parameters to filter news articles from the GDELT Project[2] - a large and open database of global news articles. This tailored corpus forms the basis for extracting relevant polarization information. This structured approach ensures that the Supplementary Corpus reflects the domain's knowledge landscape, and can serve as the basis for extracting polarization information and constructing a representative domain-specific PKG. Following, we outline the PKG construction process.

3.2 Extracting the Polarization Information

Given the Supplementary Corpus, we use POLAR [22] to extract domain-specific polarization information. We chose POLAR for its integrated approach that combines both group- and topic-oriented methods, offering a comprehensive extraction of polarization information. Initially, we process the Supplementary Corpus with POLAR to detect and link named entities, denoted as E. Then, we use syntactical dependency parsing and sentiment attitude analysis to identify the supportive or opposing attitudes between entity pairs as a function

[2] https://www.gdeltproject.org/

$r(e_i, e_j) \rightarrow \{Negative, Neutral, Positive\}$. Then, we employ signed network clustering methods to group entities with dense positive attitudes amongst them and discover the entity fellowships F. To identify fellowship dipoles D, we examine the structural balance of all pairs of fellowships; structural balance is a concept tied to polarization in signed networks [3]. Additionally, we perform clustering of semantically similar noun phrases to identify discussion topics T. Sentiment cues are aggregated from relevant sentences and noun phrases, associating an entity's attitude towards a topic. This association is captured by $a(e_i, t_z) \rightarrow [-1, 1]$, where -1 signifies strong opposition and 1 denotes strong support. The output of this process comprises of E, F, D, T, r and a, which collectively represent the polarization information extracted with POLAR. For example:

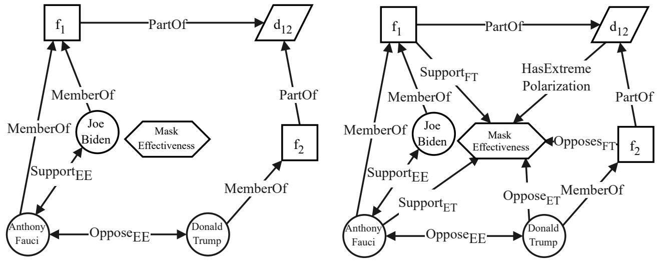
- "*Pres. Trump spent months playing down the effectiveness of masks, ... mocked former V.P. Biden for wearing one.*"
- "*Dr. Fauci ... been begging people to wear masks.*"
- "*Trump ... insulting Fauci for telling the truth.*"
- "*Biden described Fauci as a dedicated public servant ...*"

From these sentences, we discern the entities as $E = \{$"*Joe Biden*", "*Anthony Fauci*", "*Donald Trump*"$\}$, and their relationships: $r($"*Joe Biden*", "*Anthony Fauci*"$) = Positive$, $r($"*Donald Trump*", "*Anthony Fauci*"$) = Negative$, and $r($"*Donald Trump*", "*Joe Biden*"$) = Negative$. These lead to the formation of fellowships $F = \{f_1, f_2\}$ where $f_1 = \{$"*Joe Biden*", "*Anthony Fauci*"$\}$ and $f_2 = \{$"*Donald Trump*"$\}$, establishing the dipole $d_{1,2} \in D$. The discussion centers on the topic $t_1 \in T$ where $t_1 = \{$ *"effectiveness of masks"*, *"masks"*$\}$, labeled as *Mask Effectiveness*. The entity attitudes toward t_1 are quantified as $a($"*Joe Biden*"$, t_1) = 0.0$, $a($"*Anthony Fauci*"$, t_1) = 0.8$, and $a($"*Donald Trump*"$, t_1) = -1.0$.

3.3 Construction of Polarization Knowledge Graph

Although the polarization information extracted with POLAR is valuable, its lack of semantic structure presents challenges for its effective utilization, interpretation, and integration into tasks such as misinformation detection. To address this, we introduce a number of successive transformations designed to transform the identified entities, fellowships, dipoles, topics, and the structural relations thereof, into a PKG. These steps involve the conversion of elements of E, F, D, and T, into actors in the PKG, the derivation of predicates from known structural relationships and the values of functions r and a, and the enrichment of PKG topics with comprehensive descriptions. We initialize the PKG by integrating entities and their relationships, assigning predicates based on the function r. Fig. 2a depicts the initial PKG from the example.

Topical Attitude Predicates: We enrich the PKG by adding identified fellowships and topics, and by computing and integrating the predicates that reflect the attitudes of entities and fellowships towards these topics. We compute these predicates by translating the continuous function a into a categorical form, where a threshold thr determines the relationship type: if $a(e_i, t_j) \geq thr$, we assign a



(a) Initial PKG instance (b) PKG with topic attitudes

Fig. 2: Initial (2a) and the updated (2b) PKG instance.

SupportET predicate between e_i and t_j, otherwise we assign an *OpposeET* predicate. To estimate thr, we examine the sets of positive A_T^+ and negative A_T^- attitudes from every entity toward every topic, setting thr as the average of their median values. This approach captures the inherent division between support and opposition in the Supplementary Corpus, ensuring a balanced threshold for identifying *SupportET* and *OpposeET* predicates. To assign predicates between a fellowship f_i and a topic t_j, the fellowship's aggregated attitude towards t_j is calculated as the average attitude of all of its entity members towards t_j:

$$att_{f_i}^{t_j} = \frac{\sum_{e_k \in f_i} a(e_k, t_j)}{|f_i|}$$

Based on this aggregated attitude, the predicate between f_i and t_j is assigned as *SupportFT* if $att_{f_i}^{t_j} \geq thr$, or *OpposeFT* otherwise. Expanding on our example from Section 3.2, we divide attitudes into $A_T^+ = \{0.8\}$ and $A_T^- = \{-1.0\}$, establishing a threshold $thr = -0.1$. This division results in categorizing the relationships as ("Anthony Fauci", *SupportET*, t_1) and ("Donald Trump", *OpposeET*, t_1). In a similar manner, fellowships' attitudes towards t_1 are encapsulated into the triples $(f_1, \mathit{SupportFT}, t_1)$ and $(f_2, \mathit{OpposeFT}, t_1)$.

Topic Polarization Predicates: To assign the polarization-related predicates, we first measure the degree of disagreement in attitudes between dipole fellowships for each topic. To quantify this, we use the polarization index metric [19]:

$$\mu = (1 - \mathit{\Delta}_A)\delta_A$$

where $\mathit{\Delta}_A = (|A_{t_j}^+| - |A_{t_j}^-|)/(|A_{t_j}^+| + |A_{t_j}^-|)$. This represents the normalized difference between the sizes of positive $A_{t_j}^+$ and negative $A_{t_j}^-$ attitude sets w.r.t. t_j. $\delta_A = |gc^+ - gc^-|/2$ is the difference between the average attitude values gc^+ and gc^- of $A_{t_j}^+$ and $A_{t_j}^-$, respectively. The value of μ ranges from 0 to 1, with 1 indicating extreme polarization and 0 denoting no polarization. This metric aligns with theoretical concepts in political science and sociology that define polarization as both the concentration of opinions at opposing extremes and the distance between those extremes [7]. To assign the polarization predicates, we utilize the following thresholds: *HasModeratePolarization* if $\mu \leq 0.3$, *HasMediumPolarization* if $0.7 \geq \mu > 0.3$, and *HasExtremePolarization* if $\mu > 0.7$. These thresholds were determined through a combination of empirical analysis and theoretical considerations. We conducted a preliminary study on a diverse set of topics across multiple domains, analyzing the distribution of μ values. The results showed that μ values around 0.3 indicate the emergence of moderate polarization, whereas, for μ values above 0.7, extreme polarization occurs. These observations align with the work by Bramson et al. 2016 [5], who suggest that polarization emerges when opposing groups show differences but still have some overlap in their views. However, these thresholds are adaptable, allowing for customization for a variety of datasets. In our example, we calculate the polarization index μ_{t_1} for dipole $d_{1,2}$ towards t_1 based on the distinct positive and negative attitudes towards t_1, $A^+_{t_i} = \{0.8\}$ and $A^-_{t_i} = \{-1.0\}$. These attitudes yield $\Delta_A = 0$, and $\delta_A = (0.8 + 1.0)/2) = 0.9$, leading to $\mu_{t_1} = 0.9$, which translates to the triple $(d_{1,2}, \mathit{HasExtremePolarization}, t_1)$, indicating a notable polarization on t_1. Fig. 2b shows the completed PKG.

4 Article-specific Polarization Encoding

To map an article q, which has not been previously encountered, to a polarization context defined by a relevant PKG, we utilize the methodologies described in Section 3.2. These methods extract polarization knowledge from the contents of q and encode it as a micro-PKG, which is a condensed version of the primary PKG that incorporates select actors and adjusted attitudes to reflect q's narrative. This micro-PKG includes the components E_q, T_q, r_q and a_q, where it is imperative that actors in E_q and T_q correspond with those in the primary PKG. Owing to the inherent limitations of micro-PKGs, which arise from the typically brief length of individual articles, these structures may possess restricted scope and connectivity to broader polarization knowledge. To address this constraint, we enhance the micro-PKGs by integrating supplementary polarization contexts from the primary PKG and applying PKG embeddings. These embeddings serve as node and edge features in a low-dimensional vector space, capturing the structural and semantic properties of actors and predicates.

4.1 Structural Augmentation with Subgroup Dynamics

Our first strategy consists of structurally enriching the micro-PKG by adding context from the primary PKG. Specifically, we identify fellowships in the primary PKG relevant to the entities E_q within the article and include them in the micro-PKG via *MemberOf* predicate. If these fellowships are part of broader dipoles identified in the primary PKG, we integrate these dipoles into the micro-PKG and establish connections using a *PartOf* predicate. Additionally, we link the newly included fellowships and dipoles to topics T_q using relevant *SupportFT*, *OpposeFT*, or polarization-level predicates found in the primary PKG. This method allows us to extend the initial micro-PKG to cover a wider array of conflict dynamics between subgroups, reflecting both explicit mentions in the article and the larger polarization context within the domain.

4.2 Semantic Enhancement through PKG Embeddings

Our second strategy introduces PKG embeddings to enhance the representation of polarization in micro-PKGs. To learn PKG embeddings, we employ the TuckER method, which is very effective in capturing diverse types of actors and predicates [4]. TuckER is trained on known triples from the primary PKG and is evaluated on a triple set with one element (subject, predicate, or object) omitted. It decomposes a tensor into factor matrices for subjects ($\mathbf{S}$), predicates ($\mathbf{P}$), and objects ($\mathbf{O}$), along with a core tensor ($\mathfrak{Z}$) representing the interactions among them. During training, TuckER employs the function $\phi(s,p,o) = \mathfrak{Z} \times_1 \mathbf{s} \times_2 \mathbf{p} \times_3 \mathbf{o}$, where $\mathbf{s}$, $\mathbf{p}$, and $\mathbf{o}$ are the embeddings of a PKG triple's subject, predicate, and object. TuckER applies a logistic sigmoid to each score $\phi(s,p,o)$, predicting the likelihood of a triple's correctness. The training objective is to minimize the binary cross-entropy loss of these predictions, iteratively refining $\mathbf{S}$, $\mathbf{P}$, and $\mathbf{O}$. The resulting $\mathbf{S}$, $\mathbf{O}$ embeddings depict the PKG actor positions in the latent space, whereas the predicate embeddings $\mathbf{P}$ signify their role in linking actors within the PKG. Given a micro-PKG, we calculate each actor's mean embedding from $\mathbf{S}$ and $\mathbf{O}$ to ensure both its subject and object roles are considered in its representation. For each predicate, we compute the mean of its vector from $\mathbf{P}$ and the subject and object associated embeddings, capturing the predicate's context. These embeddings are integrated as nodes and edges features of the micro-PKG.

5 Polarization-Driven Misinformation Detection

To improve the accuracy of existing misinformation classifiers, we aim at enriching their training with PKG-encoded polarization knowledge. To this end, for each article in a Misinformation Dataset (MD) of interest, we construct a PKG for the domain of the MD and compute a micro-PKG for each article in the MD. Subsequently, we label each micro-PKG as either "reliable" or "unreliable," given the credibility of its source article, creating a *micro-PKG Dataset*. Leveraging the graph representations in the micro-PKG Dataset, we re-define misinformation detection as a graph classification, which entails classifying micro-PKGs as originating from "reliable" or "unreliable" news articles. To effectively integrate our approach with existing classifiers, we introduce FlexKGNN, a Graph Neural Network (GNN) designed to assimilate polarization knowledge from micro-PKGs and merge it with features from these classifiers. We base the FlexKGNN core architecture on models from related works that utilize KGs for misinformation detection, employing graph convolution and attention layers [8,17,32]. The novelty of FlexKGNN is in merging the internal micro-PKGs representation with external classifiers prior to the classification, a strategy inspired by ensemble classifiers [31]. We validate the architecture of FlexKGNN, shown in Fig. 3, through hyperparameter tuning across the evaluation datasets.

Initially, an input micro-PKG G passes a Transformer Convolution layer [33], updating the feature vector for each actor v by aggregating neighbor information using self-attention. Feature vector $H_v^{(l+1)}$ at layer $l+1$ is updated as:

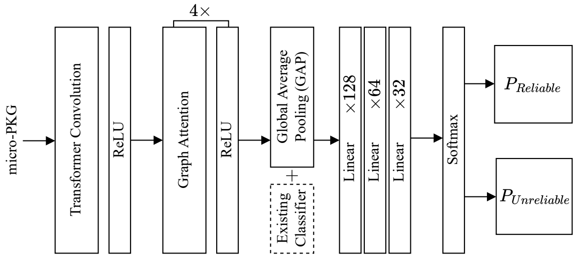


Fig. 3: Overview of the FlexKGNN model architecture.

$$H_v^{(l+1)} = \sigma(B^{(l)} \cdot W^{(l)} \cdot H^{(l)})$$

where $W^{(l)}$ and $B^{(l)}$ are the trainable weight matrix and bias vector for layer l, respectively, and $\sigma(\cdot)$ is a non-linear activation function. Subsequently, four Graph Attention (GAT) layers [33] further leverage self-attention to allow the model to learn the importance of neighbors' information dynamically. After the last attention layer, Global Average Pooling (GAP) aggregates actor and predicate features to form a micro-PKG representation H_G. Finally, H_G is passed through a series of linear layers with a final softmax activation to obtain probabilities for each class $c \in \{\mathit{reliable}, \mathit{unreliable}\}$:

$$P(G) = \text{Softmax}(W^{(c)} \cdot H_G + B^{(c)})$$

where $W^{(c)}$ and $B^{(c)}$ are the weight and bias of the classification layer, and $P(G)$ is the probability distribution over classes ($P_{Reliable}$ and $P_{Unreliable}$).

Incorporating Existing Classifiers: We integrate existing classifiers into FlexKGNN during its training phase. These classifiers, trained on misinformation features from news article content, are combined with polarization cues in the GAP layer of FlexKGNN, where the model learns the micro-PKG representation H_G. For this integration, a feature vector F is defined, representing the point at which the existing classifier merges with FlexKGNN. F can be defined in several ways: i) as a vector of misinformation features extracted using NLP techniques [35], ii) as a vector consisting of the output class probabilities from the existing classifier, or iii) as the penultimate hidden layer of DL models. To merge with FlexKGNN, the vector F is concatenated with the H_G representation, creating an augmented representation $H'_G = [H_G \mid\mid F]$. This approach employs ensemble stacking principles, where multiple models are merged to enhance their capabilities [31]. The H'_G representation is then used to calculate $P(G)$.

6 Experiments and Evaluation

To assess the effectiveness of our approach, we examine a case study focusing on articles related to the COVID-19 pandemic, a period characterized as an "infodemic[3]" exacerbated by political polarization [6,13]. To conduct our evaluation,

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

we compiled a misinformation dataset specific to this context. Our objective is to quantify the impact of integrating polarization knowledge into existing classifiers for misinformation detection. Beyond our primary case study, we apply our approach on two additional datasets to examine its broader applicability. For reproducibility purposes, we make our dataset, results, and code publicly available[4].

COVID-19 Misinformation Dataset: Our dataset comprises articles from GDELT database, from 1/2020 to 12/2021. We retrieved articles with three or more mentions of the keywords *coronavirus*, *COVID*, or *pandemic*. We maintained data integrity by using the Internet Archive[5] as a backup source for removed articles. To annotate the veracity of the collected articles, we employed the Newsguard browser plugin [6], which relies on experts manually scoring the credibility of news websites based on predefined journalistic criteria. We categorized the articles based on their credibility score, identifying 58,888 as "reliable" and 3,523 as "unreliable." The most frequent domains are depicted in Fig. 4.

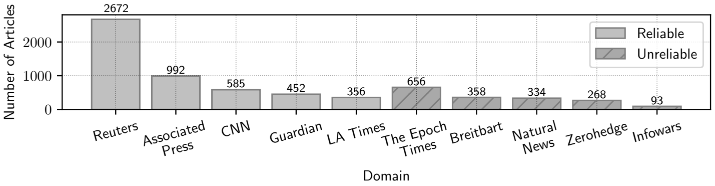


Fig. 4: Number of reliable and unreliable articles by top domains known to generate factual and false information[7].

Additional Datasets: We also evaluate on two additional datasets [28]: i) *Politifact*, consisting of 467 "reliable" and 383 "unreliable" articles, primarily focusing on the US political scene, sourced from the *politifact.com* fact-checking website; and ii) *GossipCop*, consisting of 15,313 "reliable" and 4,781 "unreliable" articles, centered around celebrity news, gathered from *eonline.com* and *gossipcop.com.* To rectify the class imbalance in our datasets, we employed domain stratified undersampling on the majority label.

Dataset	From	To	Top Keywords	# Articles
COVID-19	01/2020	12/2021	*coronavirus, sars, mandate, vaccine*	84,180
Politifact	10/2016	04/2018	*trump, clinton, president, debate*	15,710
GossipCop	06/2017	05/2018	*kardashian, bieber, prince, harry, markle*	12,768

Table 1: Characteristics of the MD Supplementary Corpora.

Supplementary Corpora: To compile the Supplementary Corpora, we automatically extract the parameters of theme, region, and timeframe from each of the MDs, to ensure their relevance with each corpus. To do so, we apply TF-IDF to extract the thematic keywords from the articles, utilize geo-extractors to identify the region, and the publication dates for timeframe alignment. To mitigate overlap, we exclude articles from each Supplementary Corpus that are

[4] https://github.com/dpasch01/PARALLAX

[5] https://archive.org/

[6] https://www.newsguardtech.com/

[7] Reuters digital news report 2020

shared with its corresponding MD. Table 1 outlines these characteristics for the Supplementary Corpora derived from the specified MDs.

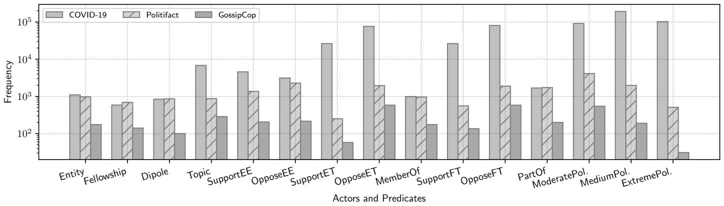


Fig. 5: Frequency of PKG actors and predicates. *ModeratePol.*, *MediumPol.*, and *ExtremePol.* represent *Has*[*Moderate*, *Medium*, *Extreme*]*Polarization* predicates.

6.1 COVID-19 PKG Overview

We construct the PKGs for each Supplementary Corpus. Fig. 5 depicts the actor and predicate frequencies for each of the PKGs. Following, we present an overview of the PKG regarding our primary case study of COVID-19. Given the number of actor and predicate observations, the high number of *OpposeET* (76,425) compared to *SupportET* (26,287) indicates a prevailing negative entity attitude towards various topics. This is also supported by the *SupportFT* (26,287) and *OpposeFT* (80,483) predicates, hinting a significant divide between fellowships. In addition, the 102,193 *HasExtremePolarization*, 192,287 *HasMediumPolarization*, and 91,262 *HasModeratePolarization* instances underscore the highly polarized nature of the topic discussions.

Entity-level Overview: Notable entities and their positive and negative attitudes are illustrated in Fig. 6. These include the *COVID-19 Vaccine* and *Pfizer*, which exhibit positive relationships, indicating their acceptance and favorable image. Regulatory bodies like *FDA* and *CDC* show higher positive attitudes, indicative of public trust [6]. Conversely, geographic entities like *Wuhan*, *Taiwan*, and *China* display mostly negative bias, possibly tied to COVID-19 origin blame. Political figures like, *Donald Trump* and *Joe Biden*, reflect mixed sentiments on their pandemic responses [13].

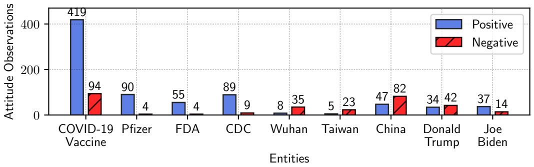


Fig. 6: Number of positive and negative entity attitudes.

Fellowship-level Overview: The PKG reveals several fellowship instances. One fellowship emphasizes the medical response, uniting entities like *COVID-19 Vaccine*, *Pfizer*, *Moderna COVID-19 Vaccine*, and regulatory bodies *FDA* and *CDC*. Another highlights the US public health response, with administrative entities like the *President of the United States* and preventive measures like

Social Distancing and *Face Masks*, along with experts such as *Dr. Anthony Fauci*. A separate fellowship revolves around the *Democratic Party*, including *Joe Biden*, *Barack Obama*, and *Bernie Sanders*. Lastly, a group centered on *Donald Trump*, highlights events like his treatment at the *Walter Reed National Military Medical Center*. Collectively, these fellowships offer comprehensive insights of the pandemic's medical and political dimensions.

Topic-level Overview: Various pandemic-related topics have been identified, which exhibit different degrees of polarization. Topics such as the *COVID-19 Case Numbers*, *Vaccine Efficacy*, *Lockdown Measures*, *COVID-19 Response*, and *COVID-19 Treatments* stand out for their high occurrences of *HasExtremePolarization* predicates. These observations align with findings of significant politicization of the pandemic, the undermining of health authorities, and hesitancy to vaccines [13]. As described in Section 3.2, the PKG topics are identified as semantically similar Noun Phrases (NPs). Initially, this process yielded 6,811 topics, each comprising an average of 1,067 NPs, making their interpretation challenging. To streamline this, we automatically label each topic with a representative title, describing its context. For the labeling, we employ OpenAI's GPT-4 API[8], leveraging the significant results of ChatGPT, including data annotation [10]. Specifically, given the NPs of each topic, ChatGPT was prompted to generate a self-explanatory title in relation to the pandemic. Examples of the annotated topics are shown in Table 2.

Topic Label	**Frequent Noun Phrases**
Vaccine Efficacy	vaccine, immunization, effective vaccine
Lockdown	lockdowns, locked-down people, lock-down
Med. Experts	expert, highly trained expert, medical expert
Resp. Mishandl.	mishandling, horrific handling, improper response handling
Mask Mandate	face mask, face covering, mask mandate
Reopening	reopening, reopening phase, collective reop.
Misinformation	disinformation, misinform, misinformation
Virus Origin	artificial origin, animal origin, man-made

Table 2: Topic examples with their frequent noun phrases.

6.2 Polarization Contribution to Existing Classifiers

Our primary evaluation goal is to measure how the integration of polarization knowledge enhances the performance of existing misinformation classifiers. For this purpose, we integrate two baseline classifiers into FlexKGNN, as detailed in Section 5. These classifiers represent both ML and DL paradigms, combining textual [23] and latent [24] features from ML and DL models, respectively, yielding SOTA results. For our ML baseline classifier, we chose Check-It [23], a feature-based misinformation detection approach. Check-It operates on a set of 256 textual features, to derive its predictions via a logistic regression model. To integrate with FlexKGNN, we concatenate its feature vector with H_G. For our DL baseline, we selected RoBERTa, a well-known pre-trained transformer model, effective in misinformation detection [24]. To integrate it with FlexKGNN, we concatenate its last 768-dimensional hidden layer with H_G.

Experimental Setup: To train the models, we split the MDs into 70% for training and 30% testing, using 3-fold cross-validation. We used a stochastic

[8] https://openai.com/blog/openai-api

gradient descent optimizer with 0.2 momentum and a learning rate of 0.02. To avoid overfitting, we applied a 10 epochs early stopping. We trained for 100 epochs, accelerated by an NVIDIA Tesla T4 GPU.

Model	**COVID-19**		**Politifact**		**GossipCop**	
	Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Check-It (C)	0.717	0.716	0.697	0.697	0.640	0.549
FlexKGNN$_{PKG} + C$	**0.728**	**0.728**	**0.830**	**0.830**	**0.750**	**0.750**
FlexKGNN$_{OpenIE} + C$	0.645	0.633	0.703	0.702	0.639	0.632
FlexKGNN$_{SRL} + C$	0.655	0.651	0.688	0.683	0.699	0.690
FlexKGNN$_{DBPedia} + C$	0.625	0.592	0.646	0.638	0.721	0.717
RoBERTa (R)	0.846	0.845	0.883	0.883	0.795	0.720
FlexKGNN$_{PKG} + R$	**0.917**	**0.915**	**0.935**	**0.935**	0.840	0.840
FlexKGNN$_{OpenIE} + R$	0.814	0.815	0.906	0.898	0.836	0.830
FlexKGNN$_{SRL} + R$	0.863	0.865	0.906	0.906	0.859	0.853
FlexKGNN$_{DBPedia} + R$	0.848	0.848	0.896	0.890	**0.874**	**0.872**

Table 3: Performances scores for baselines classifiers and their integration with PKG, OpenIE, SRL, and DBPedia.

Results: As depicted in Table 3, FlexKGNN$_{PKG}$ exhibits considerable improvement in misinformation detection when integrated with existing classifiers. With the COVID-19 dataset, both Check-It (C) and RoBERTa (R) see enhanced performance through integration with FlexKGNN$_{PKG}$, with RoBERTa F1 score peaking at 0.916, marking an ≈8% enhancement. This trend becomes more notable in the Politifact and GossipCop datasets, yielding performance increases of 19.70% and 26.64% with Check-It, and 4.34% and 12.84% with RoBERTa, respectively. These results highlight the broad applicability of our approach, and the potent contribution of polarization knowledge in misinformation detection.

6.3 Polarization Knowledge Role in Misinformation Detection

Following, we evaluate the impact of the polarization-specific PKG on misinformation detection, contrasting it with broader KGs obtained through knowledge extraction techniques. To establish a comparison, we employ Open Information Extraction (OpenIE) [2], Semantic Role Labeling (SRL) [27], and DBPedia [18] as our foundational knowledge baselines. **OpenIE** is a tool that employs a series of NLP methods and syntactical dependency rules to identify actors and their relations from text. For example, given the sentence: "Anthony Fauci emphasizes the need for a mask mandate", OpenIE discerns the triple of ("Anthony Fauci", "emphasizes", "the need for a mask mandate"). **Semantic Role Labeling (SRL)** is an NLP method that identifies semantic roles in sentences, emphasizing on actors and their actions. While OpenIE derives SPO triples using syntactical rules, SRL captures deeper entity relationships. For example, in "President Trump spent months playing down mask effectiveness", SRL distinguishes the actor "President Trump", the action "spent", and the related activities "months" and "playing down mask effectiveness", yielding two triples: ("President Trump", "spent", "months") and ("President Trump", "spent", "playing down masks effectiveness"). **DBPedia** is a knowledge graph that captures Wikipedia entries in a structured format, facilitating the semantic querying of their relationships and properties. To extract triples from text using DBPedia, we initially apply Named Entity Recognition (NER), where named entities (i.e. actors) within the text are identified. Following this, we use DBPedia Spotlight [18] to link these actors to corresponding DBPedia resources. After these actors are linked, we query DBPedia for possible relationships between them. By applying this process on a sentence such as "Anthony Fauci is the leader of the National Institute of Allergy and Infectious Diseases", the result would be ("Anthony_Fauci", "Leader", "National_Institute_of_Allergy_and_Infectious_Diseases").

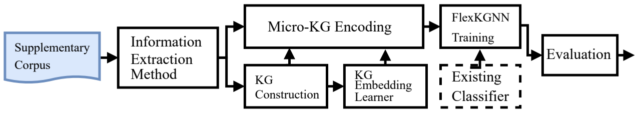


Fig. 7: Methodology for general KG encoding of MD.

Baseline KGs Construction: For each knowledge source, we employ the methodologies of Sections 3 and 4 to construct the primary KG and encode the MDs into micro-KGs (see Fig. 7). The resulting triples from OpenIE and SRL exhibit inconsistencies, as they represent the same actor differently in text (e.g., "Donald Trump" and "President Trump"). To address this, we leverage clustering based on contextualized embeddings for both actors and predicates [26]. By clustering subjects and objects after extracting all triples, we consolidate different textual representations of an entity, like "Donald Trump," into a single, unified representation. This method is similarly applied to predicates, ensuring consistency in our knowledge representation. For DBPedia, triples already have unified representation for each actor and predicate. To construct the primary KG and the micro-KGs, we follow the methodology outlined in [8]. Utilizing the constructed KGs and encoded micro-KGs, we train instances of FlexKGNN$_{OpenIE}$, FlexKGNN$_{SRL}$ and FlexKGNN$_{DBPedia}$.

Results: As shown in Table 3, while the combinations involving FlexKGNN$_{PKG}$ consistently surpass the existing baselines, those that incorporate SRL, OpenIE, and DBPedia, only achieve a performance comparable to that of the baseline classifiers, occasionally decreasing their performance, such as the 11.59% decrease in F1 score observed with FlexKGNN$_{OpenIE}$ on the COVID-19 MD. The only exception occurs with the FlexKGNN$_{SRL}$ and FlexKGNN$_{DBPedia}$ when integrated with RoBERTa on the GossipCop MD, achieving a ≈15% increase, compared to the 12.84% of the FlexKGNN$_{PKG}$. Overall, models integrated with PKG outperform those with KGs due to their fundamental differences. The PKG effectively captures polarization knowledge in the context of the MD, in contrast to general KGs, which often echo information already seen by the existing classifiers. Thus, while general KGs are useful in various settings, the specialized PKG is more effective at detecting misinformation in polarized environments.

6.4 Performance Comparison with External KG Approaches

Additionally, we compare the performance of our methodology with existing approaches that utilize KGs in combination with GNN models for misinformation detection. These models are: i) **KAPALM** [17], a GNN model that fuses coarse- and fine-grained actor KG representations in combination with article content for knowledge-aware misinformation detection, achieving F1 scores of 0.913 on Politifact and 0.717 on GossipCop; ii) **KAN** [8], a knowledge-aware attention GNN which incorporates KG actors to predict the veracity of articles, achieving F1 scores of 0.872 on Politifact and 0.774 on GossipCop; and iii) **KGF** [32], a compositional GNN, which uses OpenIE to extract actors and their relationships from articles, classifying them using graph convolutions, achieving F1 scores of 0.853 on Politifact and 0.723 on GossipCop.

Results: In comparison with the performances in Table 3, FlexKGNN PKG + R outperforms the aforementioned models on both datasets, highlighting the efficacy of integrating polarization knowledge with advanced DL techniques.

6.5 Polarization Knowledge Ablation Study

To understand the individual polarization predicate contributions, we conduct an ablation study, considering the micro-PKGs without: i) polarization predicates, and ii) embeddings. To neutralize the polarization knowledge in the PKG, we first remove the *Dipole* actors and their related predicates, which signify the conflict between fellowships. Specifically, we eliminate the predicates *PartOf*, *HasModeratePolarization*, *HasMediumPolarization*, and *HasExtremePolarization*. To obscure clear signs of opposition or support, we generalize the remaining attitude predicates by merging *SupportEE* and *OpposeEE* into *AttitudeEE*, *SupportET* and *OpposeET* into *AttitudeET*, and *SupportFT* and *OpposeFT* into *AttitudeFT*. This neutralization process is similarly applied to the micro-PKGs. As a result, the modified PKG no longer explicitly captures polarization knowledge.

Ablation	**COVID-19**		**Politifact**		**GossipCop**	
	Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
w/out Polarization Predicates + C	0.694	0.632	0.758	0.758	0.719	0.716
w/out Embeddings + C	0.723	0.723	0.694	0.693	0.698	0.695
w/out Polarization Predicates + R	0.908	0.908	0.903	0.903	0.828	0.825
w/out Embeddings + R	0.906	0.906	0.922	0.921	0.825	0.824

Table 4: Ablation study results on model performance.

Results: As Table 4 indicates, there is a noticeable performance drop across all MDs when these elements are omitted. Specifically, the absence of polarization predicates in the FlexKGNN PKG + C setup leads to ≈6% decrease in effectiveness. Similarly, discarding embeddings results in a ≈8.5% decrease. The performance of FlexKGNN PKG + R without these elements remains robust across the MDs, although a slight reduction of ≈2% is still observed, demonstrating the intrinsic strength of the RoBERTa classifier. These results underscore the added value of polarization and embeddings for misinformation detection.

7 Conclusion and Future Work

In this study, we propose PARALLAX, a methodology that leverages polarization knowledge for improved misinformation detection. Using our FlexKGNN model, augmented with PKG, consistently outperforms methods based on general KGs, achieving an average of ≈15% improvement when integrated with existing classifiers. This demonstrates the effectiveness of incorporating polarization into misinformation detection. While our findings are promising, we acknowledge there are areas for improvement. We plan to extend our evaluation to larger, more diverse datasets to ensure robust assessment and explore the adaptability of our approach to domains with limited or shifting polarization. Additionally, we aim to integrate PARALLAX with other state-of-the-art models, including Convolutional Neural Networks (CNN) and Large Language Models (LLM) [16].To provide deeper justification for polarization contribution, we will employ explainable AI techniques such as GNNExplainer [34]. This will allow us to identify and analyze the PKG triples that contribute most significantly to misinformation classification, potentially revealing new patterns in how polarization relates to misinformation.

Acknowledgments. This research is funded in part by the EU Commission via the ATHENA 101132686 project (HORIZON-CL2-2023-DEMOCRACY-01).

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. JEP (2017)
2. Angeli, G., Johnson, P., Melvin, J., Manning, D.: Leveraging linguistic structure for open information extraction. In: IJCNLP (2015)
3. Aref, S., Neal, Z.: Detecting coalitions by optimally partitioning signed networks of political collaboration. Scientific reports (2020)
4. Balazevic, I., Allen, C., Hospedales, T.: TuckER: Tensor factorization for knowledge graph completion. In: EMNLP-IJCNLP (2019)
5. Bramson, A., Grim, P., Singer, D.J., Berger, W.J., Sack, G., Fisher, S., Flocken, C., Holman, B.: Understanding polarization: Meanings, measures, and model evaluation. Philosophy of Science **84**(1), 115–159 (2017). https://doi.org/10.1086/688938
6. Deane, C., Parker, K., Gramlich, J.: A year of u.s. public opinion on the coronavirus pandemic. Pew Research Center (2021)
7. DiMaggio, P., Evans, J., Bryson, B.: Have american's social attitudes become more polarized? American journal of Sociology (1996)
8. Dun, Y., Tu, K., Chen, C., Hou, C., Yuan, X.: Kan: Knowledge-aware attention network for fake news detection. In: AAAI (2021)
9. Garimella, K., Morales, G.F., Gionis, A., Mathioudakis, M.: Quantifying controversy in social media. Trans. Soc. Comput. (2018)
10. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd workers for text-annotation tasks. PNAS (2023)
11. Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., Roy, D.: Me, my echo chamber, and i: Introspection on social media polarization (2018)
12. Guerra, P., Meira, W., Cardie, C., Kleinberg, R.: A measure of polariz. on soc. media net. based on community boundaries. ICWSM (2013)
13. Hart, P.S., Chinn, S., Soroka, S.: Politicization and polarization in covid-19 news coverage. Science Communication (2020)
14. He, Z., Mokhberian, N., Camara, A., Abeliuk, A., Lerman, K.: Detecting polarized topics using partisanship-aware contextualized topic embeddings. EMNLP (2021)

15. Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., Zhou, M.: Compare to the knowledge: Graph neural fake news detection with external knowledge. In: IJCNLP (2021)
16. Islam, M.R., Liu, S., Wang, X., Xu, G.: Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining **10**(1), 82 (2020)
17. Ma, J., Chen, C., Hou, C., Yuan, X.: Kapalm: Knowledge graph enhanced language models for fake news detection. In: EMNLP (2023)
18. Mendes, P.N., Jakob, M., Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: I-SEMANTICS (2011)
19. Morales, A., Borondo, J., Losada, J., Benito, R.: Measuring Political Polarization: Twitter shows the two sides of Venezuela. Chaos (2015). https://doi.org/10.1063/1.4913758
20. Morkūnas, M.: Russian disinformation in the baltics: Does it really work? Public Integrity (2023)
21. Osmundsen, M., Bor, A., Vahlstrup, P.B., Bechmann, A., Petersen, M.B.: Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. American Political Science Review (2021)
22. Paschalides, D., Pallis, G., Dikaiakos, M.: Polar: A holistic framework for the modelling of polarization and identification of polarizing topics in news media. ASONAM (2021)
23. Paschalides, D., Christodoulou, C., Orphanou, K., Andreou, R., Kornilakis, A., Pallis, G., Dikaiakos, M.D., Markatos, E.: Check-It: A plugin for detecting fake news on the web. OSNEM (2021)
24. Pavlov, T., Mirceva, G.: Covid-19 fake news detection by using bert and roberta models. In: MIPRO (2022)
25. Przybyła, P.: Capturing the style of fake news. In: AAAI (2020)
26. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: EMNLP-IJCNLP (2019)
27. Shi, P., Lin, J.: Simple bert models for relation extraction and semantic role labeling. arXiv preprint (2019)
28. Shu, K., Wang, S., Liu, H.: Beyond news contents: The role of social context for fake news detection (2019)
29. Tajfel, H., Turner, J.: An integrative theory of intergroup conflict. The Social Psych. of Inter. Rel. (1979)
30. Vicario, M.D., Quattrociocchi, W., Scala, A., Zollo, F.: Polarization and fake news: Early warning of potential misinfo. targets. TWEB (2019)
31. Wolpert, D.H.: Stacked generalization. Neural Networks (1992)
32. Wu, K., Yuan, X., Ning, Y.: Incorporating relational knowledge in explainable fake news detection. In: PAKDD. Springer (2021)
33. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. (2021)
34. Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J.: Gnnexplainer: Generating explanations for graph neural networks. NIPS (2019)
35. Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Comput. Surv. (2020)
36. Zollo, F.: Dealing with digital misinformation: a polarised context of narratives and tribes. EFSA (2019)