POLAR: A Holistic Framework for the Modelling of Polarization and Identification of Polarizing Topics in News Media

Demetris Paschalides
Computer Science Department
University of Cyprus
Nicosia, Cyprus
dpasch01@cs.ucy.ac.cy

George Pallis
Computer Science Department
University of Cyprus
Nicosia, Cyprus
gpallis@cs.ucy.ac.cy

Marios D. Dikaiakos
Computer Science Department
University of Cyprus
Nicosia, Cyprus
mdd@cs.ucy.ac.cy

***Abstract*—Polarization is an alarming trend in modern societies with serious implications on social cohesion and the democratic process. Typically, polarization manifests itself in the public discourse in politics, governance and ideology. In recent years, however, polarization arises increasingly in a wider range of issues, from identity and culture to healthcare and the environment. As the public and private discourse moves online, polarization feeds in and is fed by phenomena like fake news and hate speech. The identification and analysis of online polarization is challenging because of the massive scale, diversity, and unstructured nature of online content, and the rapid and unpredictable evolution of polarizing issues. Therefore, we need effective ways to identify, quantify, and represent polarization and polarizing topics algorithmically and at scale. In this work, we introduce POLAR - an unsupervised, large-scale framework for modeling and identifying polarizing topics in any domain, without prior domain-specific knowledge. POLAR comprises a processing pipeline that analyzes a corpus of an arbitrary number of news articles to construct a hierarchical knowledge graph that models polarization and identify polarizing topics discussed in the corpus. Our evaluation shows that POLAR is able to identify and rank polarizing topics accurately and efficiently.**

***Index Terms*—Natural Language Processing, Polarization, Polarizing Topic Extraction, Inter-group Conflict, Signed Networks**

I. Introduction

Polarization is becoming a major concern around the world with dire consequences for social cohesion and stability. Polarization seems to be playing a key role in shaping how events and public debates evolve and take shape, such as for instance in the US presidential elections [1], the refugee crisis in Germany [2], the Brexit referendum [3], and the recent storming of the United States Capitol [3]. Polarization is increasingly reflected in digital content and the interactions that take place online, on the Web and in social media [4]–[7]. It seems to be fed by and to feed into alarming online phenomena that undermine the democratic process, such as misinformation, disinformation, media manipulation, and hate speech [8]–[11]. Therefore, effective approaches are needed to model polarization as it is manifested in digital content and online interactions, so that it can be monitored, analyzed and understood with algorithmic means [8], [12]. However, the scale, noise and unstructured nature of online content make the definition and identification of polarization and polarizing topics extremely difficult [13]. In an attempt to alleviate this difficulty, existing efforts focus on exploring polarization in specific contexts, determined by selected issues, people or events. Also, most studies confine polarization in the context of the struggle between *Left* and *Right* in the political spectrum [4], [6], [14], and apply topic-specific approaches tailored to social networking platforms like Twitter [13]–[15] (use of replies, hashtags etc.). These approaches, however, do not provide sufficient tools for a wider mapping and understanding of polarization as it manifests in different topics and involves various actors, coalitions and diverse conflicts.

To address this limitation, we introduce POLAR, a framework for modeling and identifying polarizing topics in any domain without prior domain-specific knowledge. POLAR processes a corpus of news articles and constructs a representation of domain knowledge as a *Sentiment Attitude Graph* (SAG). SAG vertices correspond to entities extracted from the corpus (e.g. political figures, organizations, countries), and SAG edges represent associations between vertices and the stance thereof as captured from the corpus texts (e.g. supportiveness or opposition). POLAR groups SAG nodes into "*fellowships*," namely factions of entities that demonstrate supportiveness between each other [16]. Polarization is identified as fellowship pairs ("*dipoles*") with antagonizing stances between them. From the knowledge associated to fellowship dipoles, POLAR extracts discussion topics and quantifies their polarization potential by estimating the extent of conflict between supportive and antagonistic attitudes that originate from entities in the fellowship dipole [15]. Topics with high polarization are labeled as *polarizing*.

ASONAM '21, November 8–11, 2021, Virtual Event, Netherlands

ACM ISBN 978-1-4503-9128-3/21/11?/$15.00
http://dx.doi.org/10.1145/3487351.3489443

POLAR identifies polarizing topics from news articles in an unsupervised and domain-agnostic way with minimal parameterization. To the best of our knowledge, this is the first general approach to model and identify polarizing topics inside a corpus of news articles. POLAR bridges the theoretical background in group polarization [16] with algorithmic techniques and tools from NLP and Machine Learning, by proposing a novel hierarchical modeling of polarization. Also, an unsupervised approach for the construction of a hierarchical domain knowledge in terms of the SAG, entity fellowships, fellowship dipoles, and polarizing topics is introduced. The key contributions of this work are described below:

- The development of POLAR, a robust, domain-agnostic, and unsupervised framework for the modeling and identification of polarized communities as dipoles, and accurate identification of polarizing topics across news media.
- POLAR bridges the theoretical background in group polarization [16] with algorithmic techniques and tools from NLP and Machine Learning, by proposing a novel hierarchical modeling of polarization.
- An unsupervised approach for the construction of a hierarchical domain knowledge in terms of the SAG, entity fellowships, fellowship dipoles, and polarizing topics.

The rest of the paper is organized as follows. Section II presents the modeling of polarization and polarizing topics. Section III presents the flow and architecture of POLAR, and Section IV the evaluation method, along with the dataset and our findings. Finally, Section V concludes this work.

II. Modeling Polarization

A. Defining Polarization

In social sciences, polarization is defined as "*the social process whereby a social or political group is segregated into two or more opposing sub-groups with conflicting beliefs*" [17]. The term *group* refers to a set of two or more *entities* that relate to one another, share common characteristics and have a collective sense of unity [18]. A relationship is defined as a number of recurring interactions between two or more entities, and is considered as the basis of various social structures (e.g groups). Despite the commonness of a group, inter-group conflict [16] is a phenomenon that occurs frequently, as group entities tend to have conflicting attitudes on specific topics. Polarization occurs when group attitudes on one or more topics move toward more extreme positions, thus causing the division of the group into two or more conflicting *sub-groups* [16]. A *sub-group* of entities is characterized by their common beliefs, ideologies, and attitudes toward a number of topics. Because of their overall supportiveness, we name sub-groups as *fellowships*. Similarly, the conflict between a pair of entity fellowships can be viewed as a *fellowship dipole*.

Figure 1a depicts an example of a large group A, which comprises three fellowships and two dipoles.

B. Polarization Data Model

As described above, a group comprises of entities and their interactions. Thus, we model a group as an undirected, weighted, heterogeneous graph $G(V, E)$, where key entities correspond to vertices V and entity interactions correspond to edges E. An entity $v_i \in V$ represents a real world object with an abstract or physical existence and is defined as a tuple $v_i = (id_{v_i}, t_{v_i})$, where: id_{v_i} is a unique entity identifier and t_{v_i} is an *entity type*. An entity type can take values: *person* (real or fictional), *nationality*, *religion*, *political group*, *organization* (including companies, agencies and institutions), *location* (e.g. country, city etc.), *product*, *event*, *law*, and *legislation*.

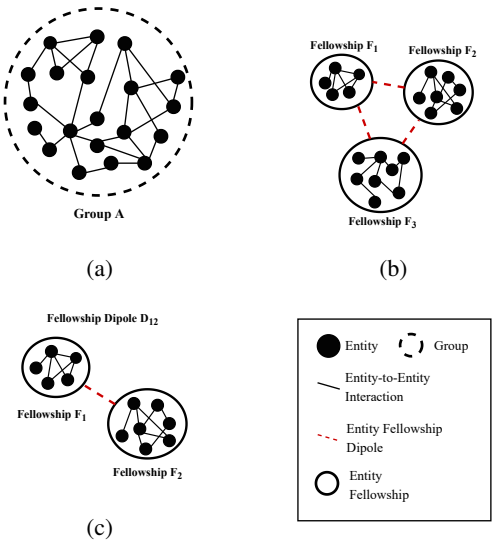


Fig. 1: The inter-group conflict steps: *1a)* existence of a general group A; *1b)* segregation to fellowships (i.e. sub-groups) F_1, F_2 and F_3 ; *1c)* formation of fellowship dipoles.

An edge between two entities v_i and v_j is defined as a triplet $(v_i, v_j, w_{ij}) \in E$. The edge weight w_{ij} indicates the nature of the relationship between v_i and v_j (i.e. *positive*, *neutral*, or *negative*), with values ranging from -1 (extremely negative) to 1 (extremely positive). The value of w_{ij} is determined by the overall sentiment attitude between an extracted entity-pair. We refer to G as the *Sentiment Attitude Graph* (SAG).

An *entity fellowship* is a dense sub-graph of the SAG, which contains predominantly positive relationships among its vertices. We denote a fellowship as $F_k = G(V_{F_k}, E_{F_k})$, where $V_{F_k} \subseteq V$ and $E_{F_k} \subseteq E$. A *dipole* D_{kl} is a sub-graph of SAG, which isolates two fellowships F_k and F_l and their (mostly) negative inter-connections: $D_{kl} = SAG[V_{F_k} \cup V_{F_l}]$.

We introduce dipole as the basic model that represents polarization and inter-group conflict between fellowships F_k and F_l in our framework.

Polarizing Topic: A fellowship dipole can be described by a set of "*discussion topics*" across its poles. A characteristic of the dipole's topics is that they bear attitudes (i.e. positive, neutral or negative) of fellowship entities. These opinions are observable manifestations of supportive or oppositional stances towards the discussion topic. When the attitudes on a topic reach a significant level of disagreement, then it is labeled as *polarizing*. This is measured by the extent to which attitudes on a topic are opposed with respect to a theoretical

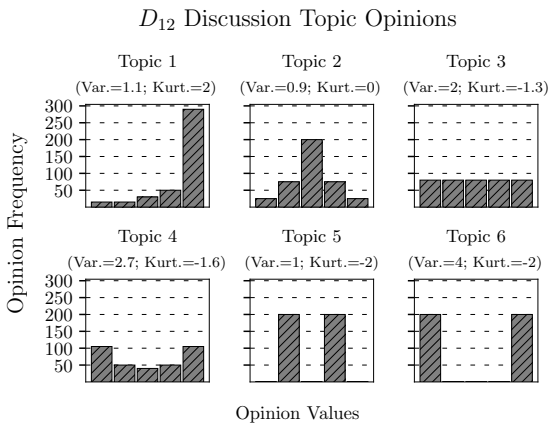


Fig. 2: Frequency of topic opinions for dipole's D_{12} using simulated data of 400 opinion observations on a scale of 1-5.

maximum [19]. The maximum is determined by the unit of measurement of attitude values (e.g., using sentiment score $\in [-1, 1]$ or a Likert scale with values $\in \{1, 2, 3, 4, 5\}$).

From a technical perspective, this translates to a bimodal distribution of topic opinions which aggregate around two modes representing positive and negative values [19], [20], and their distance representing how polarizing the topic is. In the literature, there are different approaches in measuring this distance statistically, like spread, variance, mode difference, entropy, or kurtosis [15], [19].

Figure 2 depicts an example of topics for a dipole D_{12} between fellowships F_1 and F_2. In the example, we can see the quantity of attitudes on six topics, with attitudes represented in a Likert scale of 1 to 5. Topic 1 and 2 are considered unimodal distributions, as they form a consensus around a specific attitude. Topic 3 has a flat-shaped distribution, indicating the existence of attitude balance across values. In contrast to these, topics 4, 5 and 6 show indications of polarization. Topic 4 starts to form a U-shaped distribution, a known characteristic of polarization [19], [20], but not well separated to be labeled as polarizing. In contrast with topic 4, topics 5 and 6 present a clear separation of attitudes. From the last two topics, topic 6 is considered as more polarizing, as the two modes reside on the extremes. To identify polarizing topics, POLAR employs a metric called polarization index [15], which considers the set of sentiment attitude on a topic. We provide a detailed description of polarization index in Section III-G.

III. POLAR Architecture and Algorithms

A. Overview of POLAR Algorithms

Figure 3 presents the overall POLAR framework pipeline. Upon initialization, POLAR collects news articles related to the theme of study. At first, the collection undergoes Named Entity Recognition and Linking (NERL) to identify and disambiguate entity mentions. POLAR outputs the entities V and their occurring sentences S, and proceeds in generating SAG, using the *Entity-to-Entity Relationship Extraction* (see Section III-D), and *Entity-to-Entity Attitude Extraction* (see Section III-D). Then, POLAR identifies the polarizing fellowship dipoles, by first extracting the fellowships F (see Section III-E) and generating the dipoles D (see Section III-F). For each of the dipoles D, POLAR extracts its discussion topics, and quantifies each topic's polarization (see Section III-G). As an output, POLAR[1] generates a representation of polarization knowledge, which includes: *i)* the SAG, *ii)* the fellowships, *iii)* the polarized dipoles, and *iv)* polarizing topics.

B. Collection of News Articles

News articles are the primary source of data for POLAR. Most existing works on polarization rely on messages circulating in Online Social Networks (OSNs) [4], [8], [13], [21], [22]. However, such messages are typically short, noisy, and informal [13]. News articles, instead, are typically longer, more formal, and more descriptive than the messages in OSNs. Therefore, they represent a richer and more reliable source of information for a) extracting knowledge about polarizing topics, their semantics and structure; b) identifying factors that may contribute or have the potential to instigate or mitigate polarization (individuals, concepts, events etc.). Also, evidence suggests that attitudes expressed in news articles (e.g. bias or hyper-partisanship) often play an important role in instigating polarization [4], [8].

POLAR requires two input parameters to run: a theme, which sets the general topic of the study, and a time-span, which limits the focus of the inquiry to a particular time frame. These parameters are used to specify the granularity of the analysis and focus the scope of the study. For example, if we want to explore the polarization for the COVID-19 pandemic in the US, we define the theme as "*US COVID-19 Pandemic*" and the period between years 2020 and 2021. For a more confined study, we could use as input parameters "*COVID-19 Vaccine Candidates*" and the period between March and July of 2021. POLAR deploys parallel collectors that fetch news articles from the GDELT Project[2] - a large, comprehensive, and open database of global news articles. Specifically, it selects articles whose content and publication date match the given theme and time-span parameters, and processes them in batches using parallelism, decoupled processing components, and distributed execution.

C. Identification of Entities

Named Entity Recognition: Entities are fundamental components in our approach as they populate the social and political groups we analyze. POLAR needs to locate and classify entities, mentioned in news articles into pre-defined types, namely persons, organizations, and locations. To this end, it employs a Named Entity Recognition (NER) transformer model, trained over an entity annotated dataset [23], able to identify entities within texts as sequences of tokens along with their types. Given a news article, POLAR identifies the entity mentions, the sentences they occur in, and their entity types. As a result, it returns a set of annotated sentences $S = \{(s_i, V_{s_i}), ...\}$, where each sentence s_i is linked to a set V_{s_i} of entity mentions occurring within s_i.

[1] Available at https://git.io/JzR1y [2] https://www.gdeltproject.org/

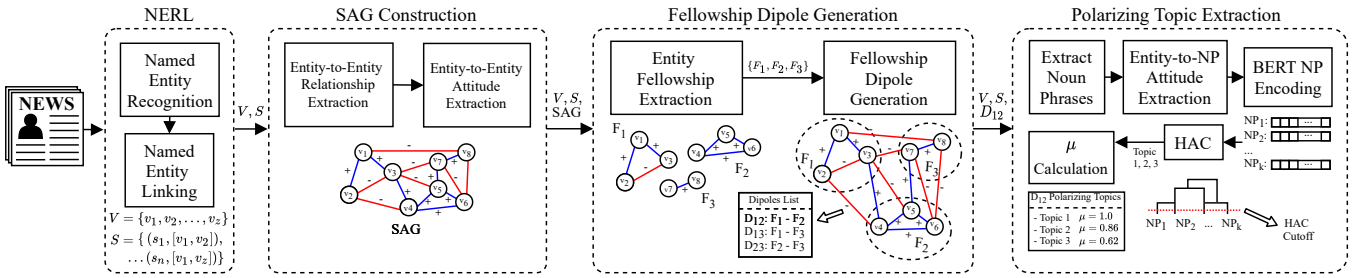


Fig. 3: An overview of the proposed framework for polarizing topic extraction from news articles.

Named Entity Linking: Applying NER to a large set of news articles results in large dimensionality, since mentions to the same entity are not recognized as such; for instance, references to "*Trump*," "*President Trump*," and "*Donald Trump*," are treated as if they referred to separate entities. Named Entity Linking (NEL) provides a solution to this issue by assigning entity mentions to unique identifiers with the help of existing Knowledge Graphs (KG), such as Wikipedia. The NEL task is split in 2 steps: *i)* finding entity candidates; and *ii)* applying collective disambiguation.

To find entity candidates, we employ a snapshot of the Wikidata[3] KG with ≈ 1 billion item entries and 1.1 TB of data. We index the Wikidata entries in Elasticsearch[4] for efficient access. To identify the candidates of an entity mention, for each entity mention in S we obtain a small set of possible entities within KG. The candidates are retrieved by executing string similarity query over Elasticsearch, using the Token Sort Ratio (TSR) measure. TSR splits the mentions into tokens, sorts them, and then compares them using the simple ratio mechanism. We consider as candidates the KG vertices with a TSR score ≥ 0.5, indicating that at least 50% of the strings are similar. Then, we proceed with collective disambiguation over the selected candidate entities. Traditional NEL methods utilize ML models that require large training corpora and extensive feature engineering. In POLAR, we employ a more optimal and domain-agnostic solution, by encoding the KG vertices in vectors of low dimensional space, called vertex embeddings [24]. Each vertex is encoded to a d-dimensional vector which captures unique characteristics of the vertex structural positioning in the KG. The unsupervised setting of this method deems it content-agnostic and applicable to any KG. We train the vertex embeddings using the DeepWalk algorithm [25] with the suggested configuration [24]. DeepWalk was selected, as it efficiently learns a latent representation of adjacency matrices in KG. Additionally, DeepWalk is scalable, and can be executed in an online manner with incremental results, and is trivially parallelizable. POLAR selects the best candidate vertex for each entity mention and provides the final entity set for S. This is achieved by measuring the semantic similarity of entities using the cosine similarity between their embeddings. To ease the complexity of this approach, we employ a greedy optimization algorithm [24].

[3] https://wikidata.org [4] https://www.elastic.co/elastic-stack

D. Sentiment Attitude Graph Generation

The Sentiment Attitude Graph (SAG) is the basic data structure of POLAR, comprising entities and their interactions. To construct the SAG, we need to extract *Entity-to-Entity* relationships, and calculate their *Sentiment Attitude*.

Entity-to-Entity Relationship: The interactions and relationships between entities are fundamental for the existence of a social or political group. POLAR identifies the existence of pair-wise entity links by quantifying their co-occurrences in the news articles. The intuition is that the higher the co-occurrence frequency of an entity pair, the more probable the existence of a real-life connection between them. Within POLAR, we calculate a *binary occurrence matrix* X based on whether or not an entity v_j is referred within the sentence $s_i \in S$. After that, we calculate the co-occurrence matrix $C = X^T \cdot X$. We keep only entity pairs whose co-occurrence frequency is over the 95_{th} percentile of the overall pair frequencies. This maximizes the accuracy of determining the pair relationships, whilst keeping the SAG at a reasonable scale for further processing.

Entity-to-Entity Sentiment Attitude: The relationship between two entities can be described as *positive*, *neutral* or *negative*, depending on the level of supportiveness or opposition between the entities discovered in the text. A simple method to determine the nature of an entity relationship is to capture the contextual sentiment of their co-occurring sentences. However, this is a naive approach as it does not consider the syntactical dependencies between words. Another approach is the *sentiment attitude identification task* [26], which seeks to identify the sentiment directed from one element in the text to another. This can be achieved by finding the explicit syntactical dependency path between the entity pair and calculating its sentiment score. By doing so, the sentiment is restricted to the syntactical relation of the two entities within the sentence, which addresses the limitations of the initial contextual sentiment approach.

POLAR calculates the sentiment attitude between a pair of entities using a lexicon-based classifier of sentence-level syntactical dependencies [26]. Given a sentence s_i, we identify a SAG entity pair (v_x, v_y), where $v_x \in V$ is the *attitude holder*, and $v_y \in V$ is the *attitude target*, by extracting all possible entity pairs. Afterwards, we calculate the sentiment attitude from the holder v_x towards the target v_y that we denote as $att(s_i, v_x, v_y) \in \{positive, neutral, negative\}$. As features

of the classifier we consider all the syntactical dependency paths between head word of v_x and v_y in sentence s_i. These features include: *i)* the sentiment label of the path that contains the dependencies between the subject $nsubj$ and direct object $dobj$ of the sentence s_i; *ii)* the sentiment label of the path containing the dependency pattern of $(nsubj, ccomp, nsubj)$ of s_i; and *iii)* an indicator of $nmod : against$, a negative relation (nominal modifier) between the two entities within s_i. Taking into account that SAG is an undirected graph, we want to consider bi-directional relationships. Thus, for each entity pair we calculate both $att(s_i, v_x, v_y)$ and $att(s_i, v_y, v_x)$. To calculate the sentiment label, we use the IBM Debater Sentiment Composition Lexicon [27], which captures the semantics of conflict and debate. After calculating the sentiment attitudes for each entity pair, we calculate the average sentiment attitude w_{xy}, and populate the edges (v_x, v_y, w_{xy}) of SAG.

E. Extraction of Entity Fellowships

POLAR identifies the formation of entity fellowships within SAG. As described in Section II-B, an entity fellowship F_i is characterized by the general supportiveness of its members. In a SAG, this characterization is analogous to densely connected graph partitions with *positive* attitudes, similar to clusters in the signed network clustering [28]. A signed network is a graph where each edge has a positive (+1) or negative (-1) sign [28]. The task of clustering signed networks amounts to finding clusters such that most edges within are positive, and most edges across are negative. Several algorithms have been proposed in recent literature for signed network clustering, based on correlation clustering, k-balanced social theory, and signed modularity [28]. However, these algorithms are limited to their dependency on modularity, which was shown to suffer a resolution limit, making the detection small communities difficult [29]. Such small communities cannot be ignored as they may represent important minorities. Also, these algorithms require a knowledge of the number of clusters k, which is undesirable as the size of SAG and the number of its fellowships is not known in advance.

To overcome the above limitations, we employ the SiMap method for the identification of fellowships within SAG [30]. Instead of the number of clusters k, SiMap accepts a resolution parameter λ, and is able to produce smaller and denser partitions as λ slides from $0 \rightarrow 1$, thus overcoming the resolution limit. As a result, SiMap is able to partition SAG into an arbitrary number of positive clusters, which we name fellowships F. We set $\lambda = 0.05$, as suggested by [30].

F. Generation of Polarized Dipoles

To identify polarizing dipoles out of the set of all possible fellowship pairs, we apply two heuristic rules: $negative\_across$ and $frustration$.

Heuristic 1 $negative\_across$**:** This heuristic measures the ratio r^- of negative edges connecting two fellowships F_i and F_j of a possible dipole D_{ij}. The intuition is that dipoles with a higher r^-, are more likely to be polarized. After a manual inspection, we find that $r^- \geq 0.5$ maximizes the probability of a dipole being polarized. The $frustration$ heuristic is applied to the remaining dipoles.

Heuristic 2 $frustration$**:** The $frustration$ heuristic takes into account the structural balance [31] of a dipole. According to structural balance theory [31], a signed graph is said to be balanced iff *i)* all the edges are positive, or *ii)* the nodes can be partitioned into two disjoint sets such that positive edges exist only within clusters, and negative edges are only present across clusters. Research has shown that balanced structural configurations of entities with signed relations (positive or negative) lead to social polarization [32]. As a result, a balanced signed graph, can be segregated into two completely opposing and conflicting fellowships. Thus, a fellowship dipole with high structural balance indicates a higher opposition between the fellowships, and a highly polarized state [12]. Among various measures is the frustration index [33] that indicates the minimum number of edges whose removal results in balance. The $frustration$ heuristic utilizes the normalized frustration index for each dipole D_{ij} as $L(D_{ij})$. L produces values from 0 to 1, with 0 being totally imbalanced, and 1 perfectly balanced. Dipoles with higher values of L indicate a higher probability of a polarized state. POLAR maximizes the number of polarized dipoles by removing the dipoles with $L < 0.7$.

G. Extraction of Polarizing Topics

Given a polarized fellowship dipole D_{ij}, POLAR identifies the discussion topics between the opposing fellowships F_i and F_j and measures the polarization around them. POLAR retrieves the sentences $S_{D_{ij}} \subseteq S$ where fellowship dipole entities co-occur. Within POLAR, we define topics as clusters of Noun Phrases (NPs) of $S_{D_{ij}}$. Following, we describe the process of NP and topic extraction, and polarization measurement.

Noun Phrases: Discussion topics can be a collection of any arbitrarily long texts, but we confine them to be a set of Noun Phrases (NPs). Grammatically, a NP functions as a noun in a sentence. One way to identify the NPs of a sentence is using constituency parsing, the task of breaking a text into sub-phrases or constituents. Consider the following sentence: "*Anti-abortionist David Daleiden caused substantial harm to Planned Parenthood*." By applying constituency parsing, we extract the following NPs: "*David Daleiden*," "*Planned Parenthood*", "*anti-abortionist*," and "*substantial harm*." POLAR parses each sentence $s_i \in S_{D_{ij}}$ for a given dipole D_{ij} using a minimal neural model for constituency parsing [34]. The model simply encodes the sentence with stacked sequence-to-sequence encoders, extracts the tokens' representations and returns the constituents. As a result, POLAR identifies the NPs within each sentence of a given dipole.

Clustering NPs to Topics: Topics are formed by clustering the NPs into groups with similar semantic meanings. To semantically cluster the NPs we encode them into word vectors. The encoding is done using the context-based BERT [35], a transformer-based pre-trained language model. The advantage of BERT against other word vector techniques [36] is that it makes use of Transformer, an attention mechanism that learns

Abortion			Immigration			Gun Control		
Topic	$\mu_A \downarrow$	μ_D	**Topic**	$\mu_A \downarrow$	μ_D	**Topic**	$\mu_A \downarrow$	μ_D
Pro-life	0.843	0.732	Racial Identity	0.845	0.732	Gun Buyback Program	0.866	0.561
Birth Control	0.842	0.642	DACA	0.809	0.681	Gun Control to Restrain Violence	0.827	0.698
Anti-abortion	0.749	0.617	Border Protection	0.787	0.718	Second Amendment	0.787	0.688
Life Protection	0.723	0.404	Refugee	0.777	0.715	Right to Self-defense	0.754	0.691
Pro-choice	0.720	0.585	Born Identity	0.773	0.655	Gun Business Industry	0.740	0.701

TABLE I: Top-5 ranked topics for *Abortion*, *Immigration*, and *Gun Control*.

contextual relations between words in the text. As a result, each NP is encoded into a $1024d$ vector.

To identify the discussion topics of the dipole D_{ij}, we semantically cluster its encoded NPs. To do so, we employ the Hierarchical Agglomerative Clustering (HAC) algorithm [37], as it does not require a predefined number of clusters k. Instead, HAC requires a cutoff threshold that provides the final set of clusters. Specifically, given a set of elements to cluster, HAC produces a distance matrix between them and iteratively merges them, thus generating a hierarchical dendrogram of clusters. The cut-off threshold, which corresponds to the maximum distance between the clusters, is then used to 'cut' the hierarchical dendrogram and provide the final set of the clusters. In POLAR, we use the cosine distance metric and set the cutoff threshold to 0.2. The resulting NP clusters represent the dipole discussion topics, denoted as $T_{D_{ij}}$.

Measuring Topic Polarization: Given a topic $t_z \in T_{D_{ij}}$, we identify the extend of its polarization, by calculating its polarization index [15]. The intuition for the polarization index is that "*a population is perfectly polarized when divided into two groups of the same size and with opposite attitudes*." For a dipole topic t_z, the population refers to the set sentiment attitudes expressed from dipole fellowship entities $v_x \in F_i \cup F_j$ towards t_z. These attitudes are determined using an adaptation of the sentiment attitude approach described in the Section III-D. Instead of a target entity v_y, we define a target NP as np_y, within the retrieved dipole sentence $s_i \in S_{D_{ij}}$. This is done by taking every pair of v_y and available NPs in the sentence. If there exist a dependency path between v_y and np_y, then we calculate the attitude as $att(s_i, v_x, np_y)$. The final set of sentiment attitudes for topic t_z is denoted as A_{t_z}.

After the extraction of the topic's sentiment attitudes A_{t_z}, the polarization index $\mu_{t_z} = (1 - \Delta_{A_{t_z}})\delta_{A_{t_z}}$ is calculated to determine the polarization of the topic t_z in the context of the specific dipole. $\Delta_{A_{t_z}}$ is the normalized difference in set sizes between the positive and negative sentiment attitudes, $A^+_{t_z}$ and $A^-_{t_z}$ respectively. $\delta_{A_{t_z}}$ is the attitude difference and is calculated as $\delta = |gc^+ - gc^-|/2$ with gc^+ and gc^- equal to the average attitude values of $A^+_{t_z}$ and $A^-_{t_z}$. The values of μ_{t_z} range from 0 to 1, with $\mu_{t_z} = 1$ if the attitudes are perfectly polarized, and $\mu_{t_z} = 0$ if there is no polarization.

The same process is repeated for all the dipoles' topics, resulting in a polarization value for each topic in each dipole. As a result, POLAR is able to produce a ranked polarizing topic list, by taking into account the average polarization index for each topic.

IV. Experiments & Preliminary Evaluation

For this work, we conduct a preliminary evaluation with main objectives the accurate identification of discussion topics, and the correctness in quantifying their polarization. We use topic-annotated news articles on specific themes, namely *i)* abortion; *ii)* gun control, and *iii)* immigration. Specifically, we measure the accuracy of POLAR in identifying and ranking the annotated topics using fellowship dipoles. Furthermore, we evaluate the robustness of POLAR by injecting different ratios of irrelevant articles to each theme.

A. Dataset

To evaluate POLAR, we use the dataset derived from Roy et al. 2020 [38]. The dataset consists of $16,475$ news articles, categorized into three divisive themes, each manually annotated with specific discussion topics: *abortion* with 4220 articles and 20 topics, *immigration* with 7794 articles and 22 topics, and *gun control* with 4461 articles and 19 topics.

In order to evaluate the correctness of POLAR, we produce a topic polarization rank list for each theme. We process each theme's articles and identify sentences that include topic-specific terms based on the topic annotations in [38]. Following, for each sentence, we identify the sentiment attitudes toward the topic-specific terms and aggregate them. As a result, we have a set of sentiment attitudes for each topic per theme. Finally, we calculate the polarization indices for each topic and produce the article-level polarizing topic rank lists, denoted as μ_A (see Table I).

B. Polarizing Topic Accuracy and Ranking

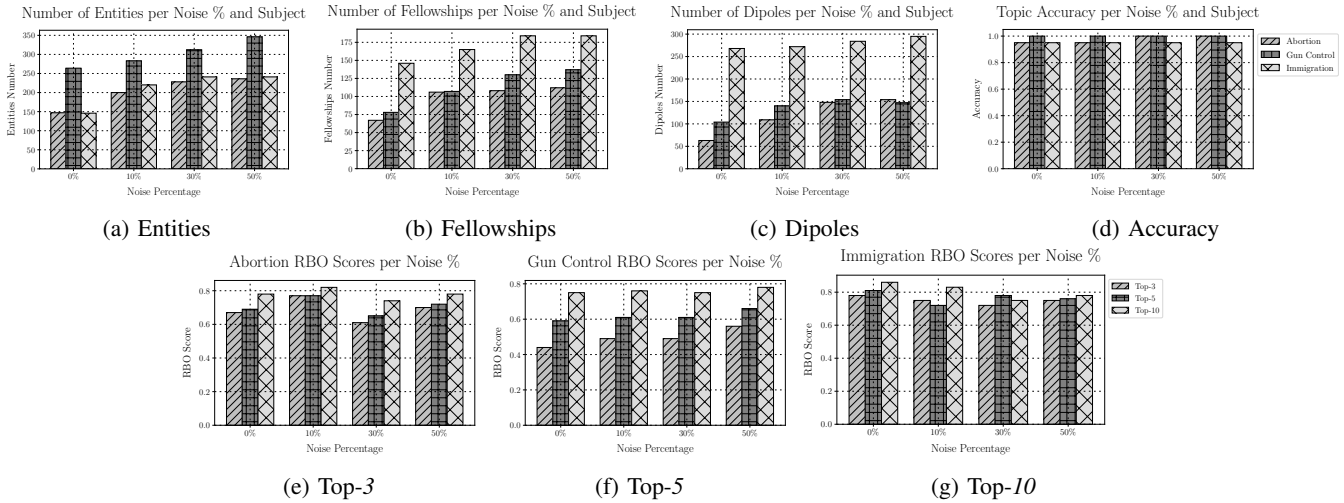
We apply POLAR to each theme's news collections, generating a SAG for each theme (SAG_{Ab}, SAG_{Im}, and SAG_{GC}) (see Table II). Next, the entity fellowships for each SAG are extracted, with SAG_{Ab}, SAG_{Im}, and SAG_{GC} partitioned to 67, 146, and 78 fellowships respectively.

Subject	**SAG**	$\|V\|$	$\|E\|$	$\|F\|$	$\|D\|$
Abortion	SAG_{Ab}	147	313	67	66
Immigration	SAG_{Im}	164	656	146	268
Gun Control	SAG_{GC}	146	291	78	104

TABLE II: Summary of SAG per theme.

Subsequently, we generate the fellowship dipoles for each SAG. Initially, we get 83 dipoles for SAG_{Ab}, 305 for SAG_{Im}, and 118 for SAG_{GC}. We filter out the dipoles less likely to be polarized by applying the $negative\_across$ and $frustration$ heuristics (see Section III-F), and get a $\approx 15\%$ reduction to the dipoles size (see Table II).

Next, we extract the dipoles' discussion topics and their sentiment attitudes. For the evaluation, we automatically label



(a) Entities (b) Fellowships (c) Dipoles (d) Accuracy

(e) Top-*3* (f) Top-*5* (g) Top-*10*

Fig. 4: Figures (a), (b), (c), and (d) describe the noise impact on each element, and (e), (f) and (g) the RBO scores per theme.

the identified dipole topics using the manually annotated topics from the dataset. We refer to the set of dipole topics as T_D, and the set of pre-defined dataset topics as T_A. The labeling is done by measuring the semantic similarity between the centroid of the NPs vectors for each T_D, with the ones in T_A. If the similarity is ≥ 0.85, we label the T_D as related to T_A. A T_D can be related to multiple T_As. For each dipole-topic pair, we calculate the polarization indices μ_D. Finally, we rank the topics of each theme based on the average aggregate of the dipoles' ranking lists μ_D (see Table I).

Topic Identification Accuracy: We proceed in evaluating POLAR's topic identification accuracy. For each theme, we compare the set of topics T_D identified by POLAR, against the manually annotated topics T_A from the dataset. We define accuracy as $|T_A \cap T_D|/|T_A|$ which indicates the coverage ratio of the annotated topic set from the POLAR topic set. Results show that POLAR manages to have high accuracy scores in all themes, with 1.00 in gun control, and 0.95 in immigration and abortion. In immigration and abortion, POLAR fails to identify the topic of "*hobby lobby*" in the former and "*wealth gap*" in the latter, both of which have significantly low observations.

Subject	**Top-*3***	**Top-*5***	**Top-*10***
Abortion	0.67	0.69	0.78
Immigration	0.78	0.81	0.86
Gun Control	0.44	0.59	0.75

TABLE III: RBO scores for top-3, top-5 and top-10 topics.

Polarizing Topics Ranking Agreement: To measure the correctness of POLAR in ranking polarizing topics, we calculate the ranking agreement between the article-level list μ_A and dipole-induced list μ_D. As a ranking agreement metric, we use the Ranked Biased Overlap (RBO) [39], an intersection-based ranking agreement measure, compared to the traditional correlation-based measures (e.g. Spearman ρ and Kendal τ). We employ the RBO as the consecutive differences between the ranked topics in both μ_A and μ_D suggest that the rankings are prone to small changes. RBO takes values from 0 to 1, with 1 indicating a full overlap (practically the same), and 0 indicating that the ranked lists are disjoint. RBO accepts a parameter d, which corresponds to the top-d elements in the list that contribute the most to the scoring. For our evaluation we check for $d = \{3, 5, 10\}$ giving the top-d a 75% contribution to the score (see Table III). Our findings suggest that, in all themes, the dipole topic rank list μ_D is highly in line with the baseline article rank list μ_A. This indicates the overall correctness of the POLAR pipeline and the use of fellowships dipoles in identifying and ranking polarizing topics.

C. Robustness to Noise

News articles do not always relate to controversial topics, and if they do, they often do not explicitly focus on them. As a result, we can have a dataset with significant noise and limited references to polarizing topics. To this end, we evaluate the impact of noise and the robustness of POLAR. Specifically, we inject percentages of irrelevant theme articles into the data being processed. The selected percentages are 10%, 30%, and 50% of the monthly news articles per theme.

Figures 4a, 4b, 4c indicate that the order of SAG, and the number of fellowships and dipoles, increase linearly with noise injection. This is expected, as more noise implies more articles to process, and additional entities in SAG.

Additionally, we evaluate the ability of POLAR to identify the fellowship dipoles. To do so, we look into the similarity of dipoles between the initial execution and each of the execution with noise percentage. For each of initial dipoles D_{ij} and noisy dipoles D^N_{kl} (N represents the noise %), we use the Jaccard index $J(D_{ij}, D^N_{kl})$ over their entities, and identify the most similar pairs. We consider the average Jaccard indices of the most similar dipoles, presented in Table IV. Despite the increment in noise, POLAR is still able to identify a large percentage of the dipoles from the original execution.

Finally, we evaluate the topic identification accuracy (Figure 4d) and ranking agreement (Figures 4e to 4g) of the noisy executions, compared to the initial ones. Based on the results, it is clear that, despite the noise injections, POLAR is able to

Noise Ratio	Abortion	Gun Control	Immigration
10%	0.83	0.68	0.89
30%	0.78	0.62	0.84
50%	0.75	0.63	0.83

TABLE IV: Dipole coverage of each subject's initial POLAR execution compared to its noise percentage.

maintain high accuracy in identifying the discussion topics and polarizing ranking agreements. The above findings indicate that POLAR is robust and tolerant to noise.

D. Processing Time Performance

We also evaluate the overall POLAR processing time for different collection sizes per theme, along with the noise injections from Section IV-C. The experiments were executed on an Ubuntu VM (64 VCPUs and 64GB RAM) along with a Tesla K60 16GB GPU. POLAR processing time scales linearly with the news article size.

V. Conclusion & Future Work

In this paper, we presented POLAR, a large-scale, unsupervised framework for the identification and ranking of polarizing topics from news articles. POLAR is able to address the literature limitations, by employing a novel hierarchical modeling of polarization, and extracting this information using state-of-the-art NLP and clustering methods. Our evaluation showed that POLAR achieves high accuracy in topic identification and polarizing topic ranking, is robust to noise, and achieves linear time performance.

Acknowledgment

This work is partially supported by the Cyprus Research and Innovation Foundation through COMPLEMENTARY/0916/0036 project.

References

[1] O. Solon, "Facebook's failure: did fake news and polarized politics get trump elected?" *The Guardian*, 2016.

[2] A. Taub and M. Fisher, "Facebook fueled anti-refugee attacks in germany, new research suggests," *NYTimes*, 2018.

[3] M. Spring and L. Webster, "European elections: How disinformation spread in facebook groups," *BBC*, 2019.

[4] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," in *Proc. of LIkKDD*. NY, USA: ACM, 2005, p. 36–43.

[5] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini, "Political polarization on twitter," *5th ICWSM*, 01 2011.

[6] K. Garimella and I. Weber, "A long-term analysis of polarization on twitter," *CoRR*, vol. abs/1703.02769, 2017.

[7] S. Aral, *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health–and How We Must Adapt*. Currency, 2020.

[8] S. Aral and D. Eckles, "Protecting elections from social media manipulation," *Science*, 2019.

[9] P. N. Howard, *Lie Machines: How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives*. Yale University Press, 2020.

[10] D. Paschalides, D. Stephanidis, A. Andreou, K. Orphanou, G. Pallis, M. D. Dikaiakos, and E. Markatos, "Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech," *ACM Trans. Internet Technol.*, vol. 20, no. 2, Mar. 2020.

[11] D. Paschalides, C. Christodoulou, K. Orphanou, R. Andreou, A. Kornilakis, G. Pallis, M. D. Dikaiakos, and E. Markatos, "Check-It: A plugin for detecting fake news on the web," *Online Social Networks and Media*, vol. 25, p. 100156, sep 2021.

[12] Z. Neal, "A sign of the times? weak and strong polarization in the u.s. congress," *Social Networks*, vol. 60, 2020.

[13] K. Garimella, G. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy in social media," *CoRR*, vol. abs/1507.05224, 2015.

[14] D. Demszky, N. Garg, R. Voigt, J. Zou, M. Gentzkow, J. Shapiro, and D. Jurafsky, "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings," *CoRR*, 2019.

[15] A. Morales, J. Borondo, J. Losada, and R. Benito, "Measuring Political Polarization: Twitter shows the two sides of Venezuela," *Chaos (Woodbury, N.Y.)*, vol. 25, 2015.

[16] H. Tajfel and J. Turner, "An integrative theory of intergroup conflict," *The Social Psych. of Intergroup Rel.*, vol. 33, 1979.

[17] C. R. Sunstein, "The law of group polarization," *University of Chicago Law School*, no. 91, 1999.

[18] J. C. Turner, "Towards a cognitive redefinition of the social group." *Current Psychology of Cognition*, 1981.

[19] P. DiMaggio, J. Evans, and B. Bryson, "Have american's social attitudes become more polarized?" *American journal of Sociology*, vol. 102, no. 3, pp. 690–755, 1996.

[20] M. Badami, O. Nasraoui, W. Sun, and P. Shafto, "Detecting polarization in ratings: An automated pipeline and a preliminary quantification on several benchmark data sets," in *IEEE Big Data*, 2017.

[21] M. Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *PNAS*, vol. 113, 2016.

[22] F. Zollo, A. Bessi, M. Del Vicario, A. Scala, G. Caldarelli, L. Shekhtman, S. Havlin, and W. Quattrociocchi, "Debunking in a world of tribes," *PLOS ONE*, vol. 12, no. 7, pp. 1–27, 07 2017.

[23] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147.

[24] A. Parravicini, R. Patra, D. Bartolini, and M. Santambrogio, "Fast and accurate entity linking via graph embedding," in *Proc. of GRADES-NDA*. ACM, 2019.

[25] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proc. of the 20th ACM SIGKDD*, 2014.

[26] E. Choi, H. Rashkin, L. Zettlemoyer, and Y. Choi, "Document-level sentiment inference with social, faction, and discourse context," in *In Proc. of 54th ACL*, 2016.

[27] O. Toledo, R. Bar, A. Halfon, A. Menczel, C. Jochim, N. Slonim, and R. Aharonov, "Learning sentiment composition from sentiment lexicons," *COLING*, 2018.

[28] J. Tang, Y. Chang, C. Aggarwal, and H. Liu, "A survey of signed network mining in social media," *ACM CSUR*, vol. 49, no. 3, pp. 1–37, 2016.

[29] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 36–41, 2007.

[30] P. Esmailian and M. Jalili, "Community detection in signed networks: the role of negative ties in different scales," *Scientific reports*, 2015.

[31] D. Cartwright and F. Harary, "A generalization of heider's theory," *Psychological Review*, vol. 63, pp. 277–292, 1956.

[32] S. Aref and Z. Neal, "Detecting coalitions by optimally partitioning signed networks of political collaboration," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[33] S. Aref and M. Wilson, "Balance and frustration in signed networks," *Journal of Complex Networks*, vol. 7, 2019.

[34] O. Vinyals, L. u. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," vol. 28, 2015.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *North American Chapter of the Assoc. for Computational Linguistics*, 2019.

[36] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[37] U. V. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing 17(4)*, 2007.

[38] S. Roy and D. Goldwasser, "Weakly supervised learning of nuanced frames for analyzing polarization in news media," in *Proc. of EMNLP*. ACL, 2020.

[39] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM TMIS*, vol. 28, 2010.