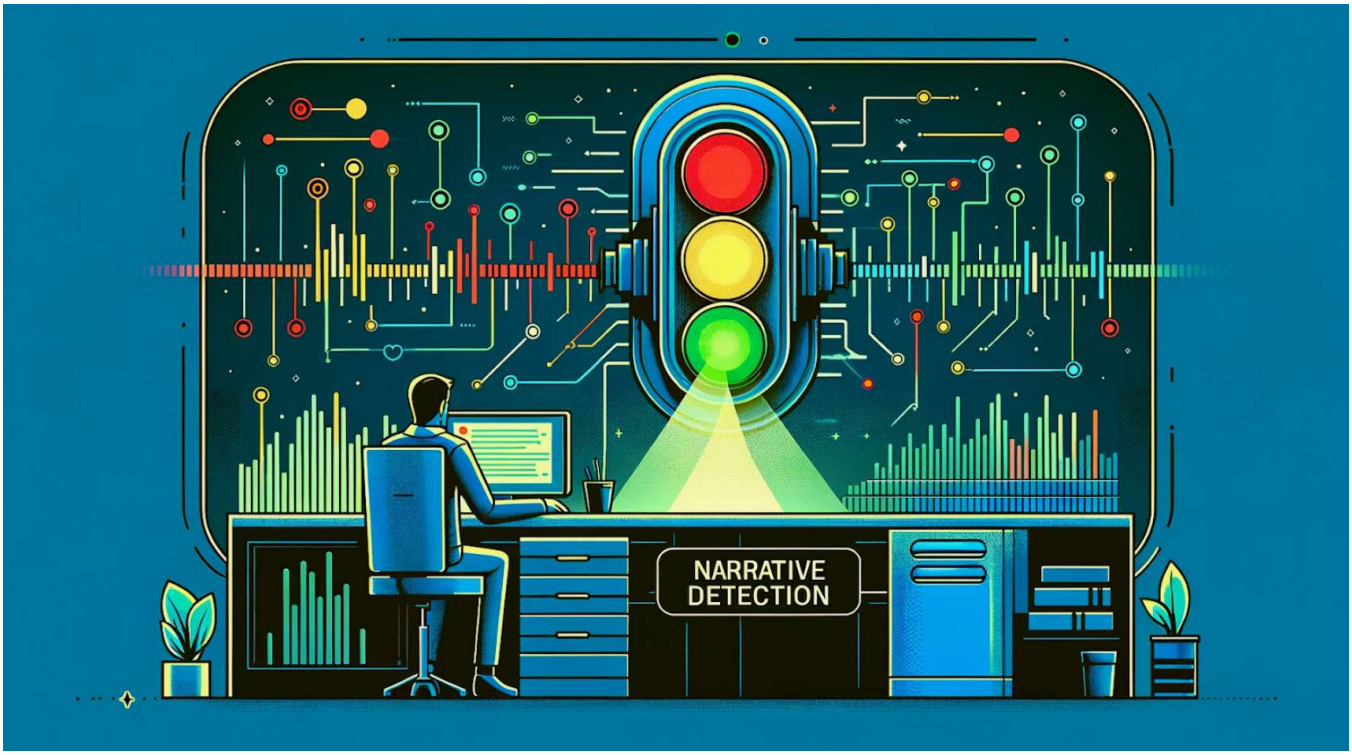
Reordering Unstructured Fact-Checked Claims into Narratives

Tomáš Nagy
RWTH Aachen
tomas.nagy@rwth-aachen.de

Sebastian Bock
RWTH Aachen
sebastian.bock@rwth-aachen.de



1 INTRODUCTION

Today's society is said to have - among others - an information problem and a knowledge problem. The information problem consists of the difficulty to choose which of the sheer endless sources of information to consume. The knowledge problem means the difficulty of knowing which concrete information is true or which information source can be trusted. With rising amounts of intentional and unintentional fake news in recent years, the practice of fact-checking has established itself as one strategy to improve the information environment.

Fact-checking organizations check whether pieces of information, primarily single claims, are true or false and usually provide an explanation for their conclusion. The research institution GESIS – Leibniz Institute for the Social Sciences created ClaimsKG, a database of these verified claims [3]. This resource is compiled and continually updated through a semi-automatic process that aggregates fact-checked claims and their associated metadata from 13 different fact-checking organizations. They provide them in a standardized RDF format. Additionally, they normalize the truth ratings of the different fact-checking organizations and apply entity detection and linking to the claim texts. The database contains about 72.000 claims as of March 2024 [11].

Where the splitting of longer texts like news articles into smaller pieces, i.e. single claims, provides benefits for executing fact checks, it simultaneously reduces the comprehensibility of the results. Even though GESIS provides a web interface for a structured search of their database [4], even an experienced user might quickly feel overwhelmed navigating a large number of results. Especially the general public is used to consuming information in narrative forms, mostly as news articles [9].

For this purpose, we have designed a framework to fix this shortcoming by presenting interrelated fact-checked claims belonging to

one context as labelled topics, with each topic consisting of multiple narratives. Thereby, we want to make the valuable and hard-earned knowledge within the ClaimsKG database better accessible to a broader public.

Our research project consists of 8 sections. We have begun with an Introduction (1), followed by a short overview of the dataset (2). Next, we provide a high-level description of our framework, explain general constraints and choices, and relate them to other work (3). Subsequently, each of the three levels of the framework will be discussed in detail, starting with defining a context (4), clustering the claims of one context into topics using BERTopic (5), and clustering the claims of one topic into narratives (6). This is followed by a brief description of our evaluation technique and its results (7). We conclude with the most significant findings, limitations, and avenues for further research (8).

2 DATASET

The fundamental unit of the dataset is a claim. We could retrieve 73310 claims with a unique claim text. We discarded 7.135 claims because they were either not in English or they were shorter than three words. Thus, we started our analysis with 66.175 claims. Please refer to the appendix for some more detailed statistics about the dataset.

A typical claim, which will serve as our running example, would be: “President Joe Biden’s climate plan includes cutting 90% of red meat from Americans’ diets by 2030.”

The topics contained in the dataset are quite diverse, with many claims being in areas of specific fake news domains, e.g. COVID or the Russo-Ukrainian war. Also, the claims were collected all over the world, with, for example, one fact-checking organization specializing on Africa (Africacheck) and one located in India (Vishvanews).

We made two observations about ClaimsKG. On the one hand, the claims are usually sensible complete sentences (mostly one sentence), so they are considerably more structured than a dataset consisting of Tweets. On the other hand, there are quite some claims which are much less precise and contentful than e.g. news headlines, additionally, many claims don’t have associated time data. This means that the claims we are working with are not that well accessible through a framework like the event extraction system 5W1H from the newsfeed tracking and organization community [6, 7].

3 FRAMEWORK OVERVIEW

The desired outcome of our clustering project is distinctively inspired by the paper "Story Forest". [9]

We believe that organizing individual pieces of information, such as claims or condensed news, in the form of narrative graphs is beneficial. Such a method allows the user understanding of complex information by enabling interactive exploration across different narratives. With the stated goal in mind, and given the challenge of working with a dataset characterized by unstructured, unlabeled, short claims, rather than well-ordered, topic or narrative-labeled or longer news articles (as in Story Forest), we came up with what we call an "entity-centric approach".

Following this strategy we build a structured pipeline to transform a collection of unstructured claims into cohesive single narratives, consisting of three major steps:
In the initial step, a context is defined by a collection of entities to identify a relevant subset of all claims.
The second step involves applying the unsupervised neural-based topic clustering algorithm, BERTopic [5], following the recent developments in the topic modelling community [1]. Within the BERTopic algorithm, each claim associated with a specific context is assigned to one topic. To enhance interpretability, each topic receives a label created through the use of a Large Language Model (LLM).
In the final step, claims within a single topic are further organized into narratives. This organization allows for each claim to be part of multiple narratives, achieved through the clustering of sentence embeddings. Within our algorithmic design, each narrative consists of a series of claims. Claims that fail to exhibit sufficient similarity to others are categorized as an individual narrative. Each narrative gets a label in the form of a narrative name using an LLM. We visualize each narrative as a sequence of interconnected nodes and each topic as a graph which encapsulates all narratives within it. The steps in our framework thereby are inspired by and built upon established practices from the narrative extraction community [8, 10].

4 DEFINING A CONTEXT

We refer to our methodology as ’entity-centric’ due to the initial step of defining a context through the selection of specific entities of interest. This process aims to simplify the complexity of the dataset by capturing a relevant subset. This is done by defining a set of words, which will be searched for in the list of detected entities of each claim and the text of each claim. The outcome of this first step is the set of claims for which one or more of these words were found.
Our running example, “President Joe Biden’s climate plan includes cutting 90 % of red meat from Americans’ diets by 2030.” would thus have been selected if the context, for example, included one of the words “Biden”, “America”, “meat” or “climate”. The detected entities for that claim are “2030”, “Joe Biden”, “climate plan”, and “red meat”, which in this case could be neglected because all of these strings are immediately present in the claims text itself. We chose this approach for the following reasons:

- ClaimsKG is already enriched with detected entities, which seemed reasonable to reuse.
- Replicating the keywords-based approach from [9] would have been infeasible as we don’t have labelled data. While searching for alternatives working on unlabeled data, we found that neural-based topic modelling, such as BERTopic has emerged in recent years and is a lean and efficient approach for topic clustering [5], especially also suited for shorter texts. We found that applying this technique immediately to the whole dataset does not make sense, because the dataset is too huge and too diverse and the number of outliers would thus be too large. Additionally, it can only assign each claim to exactly one topic. Our running example could then only be part of either the topic ‘Biden’ or the

topic ‘climate’, but not of both. Moreover, applying topic modelling in the first step would be computationally much more expensive than our string-matching approach.

- Defining a set of keywords, often entities, is a common approach in communication and political science research. Additionally, it seems plausible that a non-scientific consumer interested in exploring fake news, too, would start her exploration with a certain topic or person in mind. The key characteristic of this “divide and conquer” approach is that each claim can belong to multiple contexts. In this first step, we don’t aim to partition the dataset but select all claims that might be important for a user interested in a particular context. We have in mind the interested end user who wants to explore the narratives belonging to a certain topic.

Our context-capturing procedure significantly diverges from classical topic modelling approaches, which typically aim to find the ’best’ way to divide the dataset into distinct topics, usually assigning each document to single or multiple topics. While this can be an effective method for basic high-level topic discovery, especially when little is known about the dataset, it may not suffice for more specialized inquiries. In the realm of humanities research, for instance, there is often a pre-existing interest in a specific domain, necessitating a more nuanced analysis. Our approach, by focusing on a subset of the information relevant to the interest, is designed to address this need. It seeks to preserve the richness and complexity of the data, enabling more accurate and insightful analyses.

A significant limitation of our plain string matching technique is its inability to identify claims that share semantic similarity without using identical terms from the context. For instance, if one had searched for “policy” or “US Government,” our running example would not have been in the result set. Integrating a semantic analysis model could possibly enhance the results by recognizing semantically similar expressions that vary lexically. However, for the scope of this project, we opted not to explore it further. The benefits of string matching—its speed, reduced computational demand, and the straightforward nature of both the process and its outcomes—were deemed sufficient for our current objectives, prioritizing ease of understanding and practicality.

5 TOPIC MODELLING WITH BERTOPIC

To optimize the use of BERTopic for topic modelling, we first run the model multiple times on a dataset using the default parameters. This iterative process is not intended to apply BERTopic directly but rather to capture and analyze the variability in the number of unique topics identified in each iteration. By documenting these outcomes, we better understand the stability and robustness of BERTopic under its default settings, which helps us infer an optimal number of topics the algorithm will likely identify in the dataset.

After identifying the most frequent number of topics, we proceeded to run BERTopic. For this study, we conduct one iteration of BERTopic. The embedding process utilizes the ’sentence-transformers/all-MiniLM-L6-v2’ model for generating document embedding in 768-dimensional space, followed by the dimension reduction technique UMAP and the identification of dense clusters of documents with HDBSCAN. For each cluster, a custom representation model inspired by KeyBERT is employed to refine topic identification and representation.

Finally, the BERTopic model, including the custom and embedding models, is saved using the SafeTensors serialization method. This step ensures all aspects of the model are preserved for future use, further analysis, or study replicability. It’s important to note that in the default BERTopic implementation, each document is assigned to only one topic. This assignment results from the clustering step, during which documents are grouped based on their similarity in the embedding space. If we consider the Biden context that encompasses 1598 claims, the nature of the clustering technique results in approximately 30% of these claims not being assigned to any cluster. After we receive the list of topics, we automatically label the topics with human-readable names using ChatGPT-4. Subsequently, we use DataMapPlot to visualize the results.

6 NARRATIVE GENERATION

The foundational step in the narrative generation process involves systematically organizing and saving semantically related sentences. Utilizing an algorithm that computes the cosine similarity between sentence embeddings, our methodology clusters semantically cohesive groups within a specified threshold. This precision in leveraging embeddings and thresholding facilitates the assembly of highly coherent claims. We tested different configurations and observed that the ’all-MiniLM-L6-v2’ sentence transformer model, coupled with a 0.60 similarity threshold, yielded good results. This combination balances sensitivity and specificity in the detection of semantically close sentences. The resulting cluster of sentences that share semantic similarities is a narrative. It is important to note that each claim may be part of multiple narratives. This enhances the richness and diversity of the narratives generated and thus allows for a more expansive exploration of the thematic content. Each narrative is visualized as a sequence of interconnected nodes, where each node has a colour corresponding to the truth rating of the claim it represents (i.e. green for true, yellow for mixed/other and red for false). Each topic is thus visualized as a graph which encapsulates all narratives within the topic.

7 EVALUATION

The quality of our framework is not straightforward to determine. We don’t have any ground truth information available, and ground truth, in our case, would still be relative to every context. The ground truth describing one context would be all claims that human readers find relevant to that context. Afterwards, humans would have to decide what the best split of that context into topics and subsequently into narratives would be. This would require the evaluators to know all claims and be able to process them simultaneously, which is completely infeasible for many hundreds or even thousands of claims. Intrusion testing established itself as the standard procedure to evaluate topic clustering without ground truth available [2].

Following them in their approach, we conducted intrusion tests for our two case studies. We adopted a strategy incorporating a two-layer evaluation: one intrusion test focusing on topics and the other

on narratives. For these, we presented two human annotators with a series of test sets, with each test set containing up to five claims belonging to the same topic or narrative and one claim randomly sampled from a different topic/narrative. This yielded the following results:

Table 1: Evaluation of the coherence of topics and narratives after applying BERTopic and narrative clustering on the Biden dataset and the COVID dataset.

Context	Evaluated Level	Number of test sets	Both annotators correct	One annotator correct	Both annotators incorrect	Precision
Biden	Topics	28	17 / 60.7%	7 / 25%	4 / 14.3%	73.2%
Biden	Narratives	20	16 / 80%	3 / 15%	1 / 5%	87.5%
COVID	Topics	25	18 / 72%	4 / 16%	3 / 12%	80%
COVID	Narratives	25	13 / 52%	4 / 16%	8 / 32%	60%
Biden & COVID	Topics & Narratives	98	64 / 65.3%	18 / 18.4%	16 / 16.3%	74.5%

The precision is the quotient of correct answers overall answers. The expected value for random answers would be .17 for the topics and around .25 for the narratives. (As some narrative test sets contained less than five correct claims because the tested narrative only contained that few claims.) The overall precision of 74.5 is seen as a good but still improvable result. Translated to everyday terms, it means that 3 out of 4 topics/narratives obtained by our pipeline exhibit a relatively high coherence. Unfortunately, it is impossible to test the recall without having some curated expected topics/narratives, i.e. ground truth created by human annotators.

8 RESULTS, LIMITATIONS AND FUTURE RESEARCH

Our study adds value by enhancing human exploration of the knowledge contained in the information space established by fact-checking organizations. Our framework mirrors especially the analytical approach of researchers in fields like communication and political science, who initially concentrate on a specific domain or context before delving deeper into uncovering more profound insights from complex information. In our research, we present an innovative narrative discovery framework that employs an 'entity-centric' top-down approach. This approach makes the analysis accessible to undertake, easily understandable, efficient and scalable. Especially in the absence of labelled data or other cues to approach a dataset consisting of claims, our unsupervised approach helps identify prevailing narratives. Our method can be applied in fields such as computational social science, media analysis or political science research and anywhere else, where an understanding of narratives is beneficial.

However, our framework is not without limitations. The initial evaluation was confined to just two case studies: Biden and Covid, raising questions about its broader applicability across various contexts. Another weak point is the necessity to define a context of keywords as the first step because non-experts, in particular, may struggle to identify the most relevant keywords for their context or domain. We advise those new to the dataset to start their exploration process with embeddings visualization (Wizmap) or topic modelling to get an overview of the different topics contained within it. The lack of complete reproducibility of the labels for topics and narratives, influenced by the inherent variability in large language models like GPT-4, presents another challenge, potentially leading to inconsistent outcomes across various experiments. The reliance on sentence-transformer models during the BERTopic step and during the narrative extraction step also limits the models explainability and generalizability and are areas for future enhancement.

These limitations provide potential for future improvements. The task of selecting the right keywords to capture the context of interest can be assisted through the incorporation of Large Language Models (LLMs) or TF-IDF techniques. These solutions aim to streamline the initial task, allowing users to more intuitively identify relevant entities that encapsulate the essence of their specific context or domain. Moreover, generating concise human-readable labels for topics and narratives through open-source large language models instead of closed-source models is a logical step to make the pipeline accessible for everyone. Future efforts should focus on improving both the topic and narrative layers of our framework. In the topic layer, one big open challenge is outlier reduction. Another one is the challenge to allow for the assignment of one claim to multiple topics. Meanwhile, enhancing the narrative layer involves refining the representation layer by removing duplicates (i.e. highly similar claims referring to the same situation or utterance) and customizing it for specific end users.

In summary, our framework can contribute to the initially mentioned challenges our society faces. By enabling everyone to access the knowledge compiled in datasets like ClaimsKG in an easy fashion, we hope to make a small contribution to mitigating the information and knowledge crises.

REFERENCES

[1] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems* 112 (2023), 102131. https://doi.org/10.1016/j.is.2022.102131

[2] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), Vol. 22. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

[3] Gesis. 2024. *ClaimsKG*. Retrieved March 7, 2024 from https://data.gesis.org/claimskg/

[4] Gesis. 2024. *ClaimsKG Explorer*. Retrieved March 7, 2024 from https://data.gesis.org/claimskg/explorer/home

[5] Maarten R. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv* abs/2203.05794 (2022). https://api.semanticscholar.org/CorpusID:247411231

[6] Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5W1H : A Universal System for Extracting Main Events from News Articles. In *Proceedings of the 7th International Workshop on News Recommendation and Analytics (CEUR Workshop Proceedings, 2554)*, Özlem Özgöbek, Benjamin Kille, and Jon Atle Gulla (Eds.). CEUR, Aachen, 35–43. http://ceur-ws.org/Vol-2554/paper_06.pdf

[7] Felix Hamborg, Corinna Breitinger, Moritz Schubotz, Soeren Lachnit, and Bela Gipp. 2018. Extraction of Main Event Descriptors from News Articles by Answering the Journalistic Five W and One H Questions. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (Fort Worth, Texas, USA) *(JCDL '18)*. Association for Computing Machinery, New York, NY, USA, 339–340. https://doi.org/10.1145/3197026.3203899

[8] Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. 2023. A Survey on Event-Based News Narrative Extraction. *ACM Comput. Surv.* 55, 14s, Article 300 (jul 2023), 39 pages. https://doi.org/10.1145/3584741

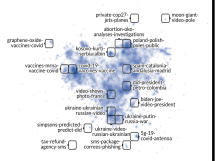
[9] Bang Liu, Fred X. Han, Di Niu, Linglong Kong, Kunfeng Lai, and Yu Xu. 2020. Story Forest: Extracting Events and Telling Stories from Breaking News. *ACM Trans. Knowl. Discov. Data* 14, 3, Article 31 (may 2020), 28 pages. https://doi.org/10.1145/3377939

[10] Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artif. Intell. Rev.* 56, 8 (jan 2023), 8393–8435. https://doi.org/10.1007/s10462-022-10338-7

[11] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In *The Semantic Web – ISWC 2019*, Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz,

Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon (Eds.). Springer International Publishing, Cham, 309–324.

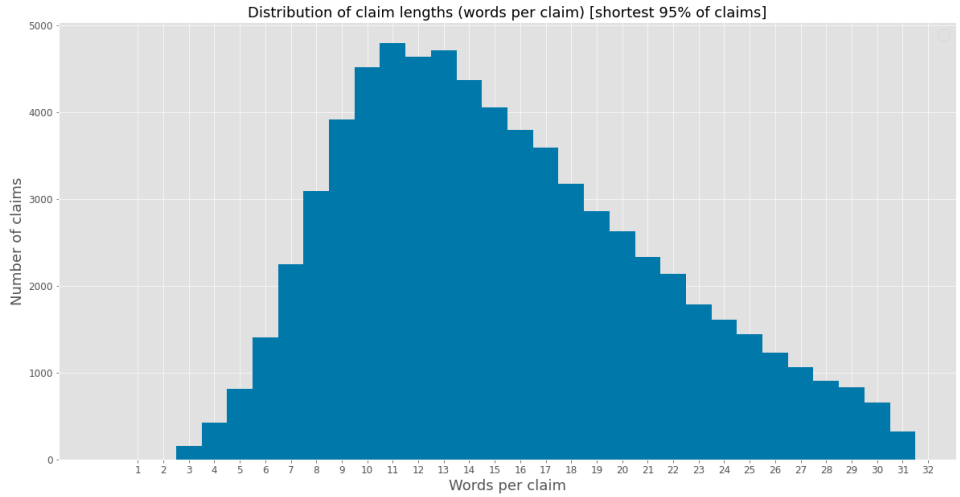
Appendix

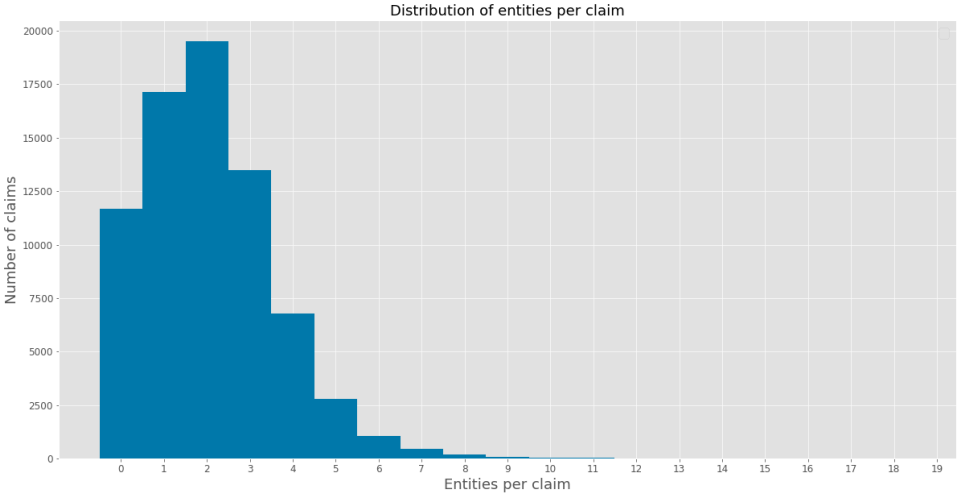
Component	Definition	Methodology	Instance	Illustrative Example
Dataset	The ClaimsKG dataset compiles 66,175 fact-checked claims from 13 diverse fact-checking organizations.	Retrieval and preprocessing of claims.	ClaimsKG	
Context	Refers to a subset of claims characterized by common entities or a connected collection of topics.	Entity-based search	Biden = [“Biden”, "Sleepy Joe", "Uncle Joe"]	
Topic	A coherent group of narratives that share consistency or similarity in their claims.	BERTopic algorithm + LLM-based labeling	Topic1: Biden's Keystone Pipeline Decision and Impact on Gas prices and other Climate Related Claims	
Narrative	A sequence of claims that together construct a coherent narrative, with (logical) connections between them.	Sentence embeddings clustering based on cosine similarity + LLM-based labeling	Narrative 1: Biden's Climate Plan and Red Meat Consumption	
Claim	A claim is a real-world statement. It can be categorized with varying levels of veracity: true, false, mixed, or other.	N/A	Claim1: President Joe Biden's climate plan includes cutting 90% of red meat from Americans' diets by 2030.	
Entity	Represents elements of the claim that exist as distinct and individual units, such as persons, organizations, objects, concepts, or any identifiable factor.	N/A	[“Joe Biden”, “climate”, “2020”, “red meat”]	N/A

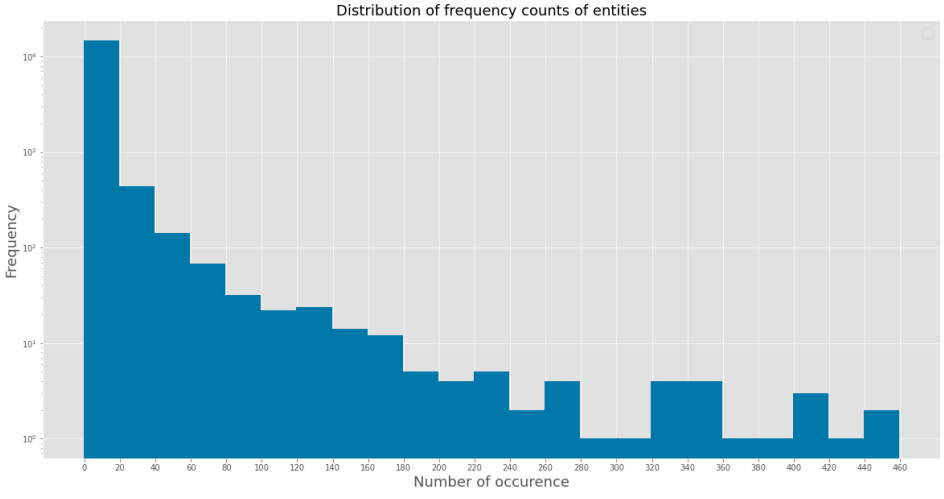
Overview of the Narrative Discovery Framework for Unstructured Fact-Checked Claims

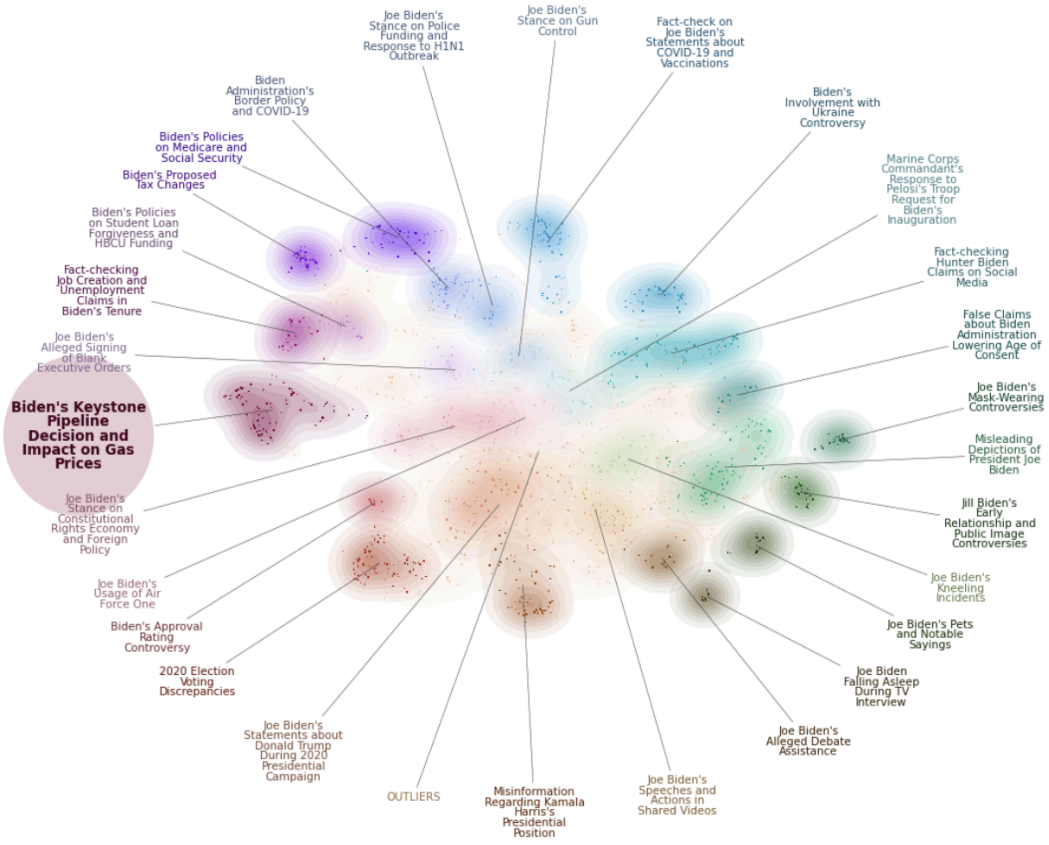
Term	Occurrences
Facebook	4530
COVID-19	1390
Donald Trump	964
U.S.	936
Texas	723
President Donald Trump	722
Joe Biden	671
United States	623
Hillary Clinton	598
Kenya	594
China	570
President Obama	558
Florida	551
Nigeria	514
Wisconsin	487

Overview of the most frequent Entities in the ClaimsKG dataset

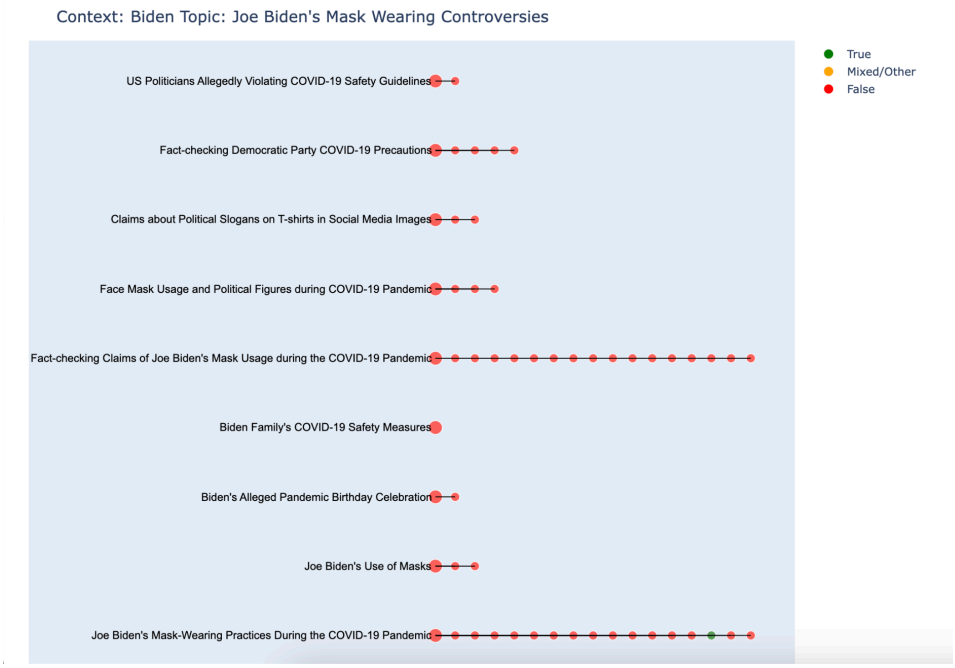

Distribution of claim lengths (words per claim) [shortest 95% of claims]
Number of claims
Words per claim


Distribution of entities per claim
Number of claims
Entities per claim

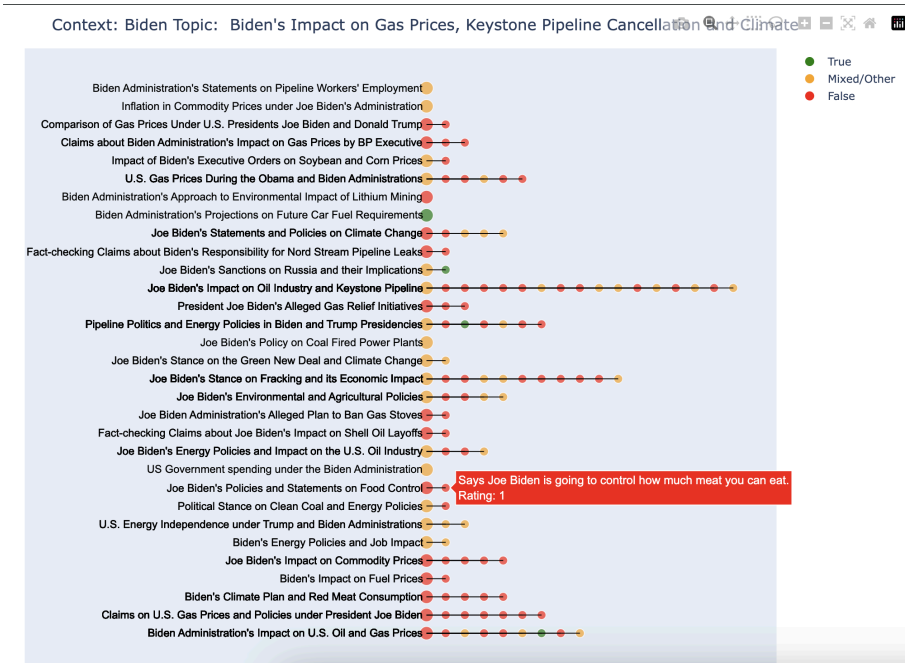

Distribution of frequency counts of entities
Frequency
Number of occurence



Visualization of Topic Modelling: Biden-Related Context



Visualization of Narratives: Joe Biden’s Mask Wearing Controversies



Visualization of Narratives: Biden's Impact on Gas Prices, Keystone Pipeline Cancellation and Climate