Semi-Automatic Topic Discovery and Classification for Epidemic Intelligence via Large Language Models

Federico Borazio[†], Danilo Croce[†], Giorgio Gambosi[†], Roberto Basili[†], Daniele Margiotta[‡], Antonio Scaiella[‡], Martina Del Manso[∗], Daniele Petrone[∗], Andrea Cannone[∗], Alberto Mateo Urdiales[∗], Chiara Sacco[∗], Patrizio Pezzotti[∗], Flavia Riccardo[∗], Daniele Mipatrini[+], Federica Ferraro[+], Sobha Pilati[+]

[†] Department of Enterprise Engineering, University of Rome Tor Vergata, Italy
[‡] Reveal s.r.l.
[∗] Infectious Diseases Department - Istituto Superiore della Sanità
[+] General Directorate for Health Prevention - Italian Ministry of Health
borazio@ing.uniroma2.it, {croce,basili}@info.uniroma2.it

Abstract

This paper introduces a novel framework to harness Large Language Models (LLMs) for Epidemic Intelligence, focusing on identifying and categorizing emergent socio-political phenomena within health crises, with a spotlight on the COVID-19 pandemic. Our approach diverges from traditional methods, such as Topic Models, by providing explicit support to analysts through the identification of distinct thematic areas and the generation of clear, actionable statements for each topic. This supports a Zero-shot Classification mechanism, enabling effective matching of news articles to fine-grain topics without the need for model fine-tuning. The framework is designed to be as transparent as possible, producing linguistically informed insights to make the analysis more accessible to analysts who may not be familiar with every subject matter of inherently emerging phenomena. This process not only enhances the precision and relevance of the extracted Epidemic Intelligence but also fosters a collaborative environment where system linguistic abilities and the analyst's domain expertise are integrated.

Keywords: Epidemic Intelligence, Topic Discovery, Large Language Models, Zero-shot Classification

1. Epidemic Intelligence: Objectives and Challenges

Following the paradigmatic change from disease specific to an all-hazard approach to the assessment of public health introduced in the 2005 revision of the International Health Regulations[1], the concept of Epidemic Intelligence was defined as a complex of activities related to the early identification of potential health hazards, their verification, assessment, and investigation that aim to generate information to guide appropriate actions in public health (Paquet et al., 2006), (World Health Organization, 2014). Within this global framework, Member States have developed ways to implement this concept to support situation awareness and evidence-based decision-making in public health. Italy started to develop its own national approach to Epidemic Intelligence in 2007 as part of a project funded by the Italian Ministry of Health coordinated by the Istituto Superiore di Sanità (ISS) (Del Manso et al., 2022). At this time a situation and need assessment was performed in order to identify existing capacities and areas with additional implementation requirements.

The results led to the conclusion that while the epidemiological monitoring conducted on data generated by existing national surveillance systems for infectious diseases (clinical, laboratory-based, and syndromic) could support an indicator-based component for the early detection of transmission events in the country, an Epidemic Intelligence system in Italy would need to develop ex novo an event based surveillance component. This component would be an extremely sensitive and flexible surveillance system based on open-source unstructured information published online concerning cases and clusters of infectious disease occurring in Italy in order to inform as soon as possible decision-making and public health experts or to provide information to clinicians and improve the timeliness of diagnoses. Some of this information would be validated (i.e. verified with public health officials within the country). The selection and assessment of news items would be performed by trained analysts to detect events of public health importance according to the methodology developed by the European Centre for Disease Prevention and Control (ECDC[2]). Following several pilots to design and test this national event-based surveillance component of Epidemic Intelligence, Italy chose to follow the implementation model developed and sustain-

[1] International Health Regulations (2005): https://iris.who.int/bitstream/handle/10665/43883/9789241580410_eng.pdf

[2] ECDC: https://www.ecdc.europa.eu/en/news-events/e-learning-course-epidemic-intelligence-ei

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 68–84
21 May, 2024.

ably implemented by the Global Health Security Action Group Early Alerting and Reporting project (EAR) (Riccardo et al., 2014). This consisted of a decentralized approach in which participating countries contributed analysts that were operational on a rotation basis.

In order to apply this to the Italian regionalized health care system, since 2017, Italy has adopted a decentralized method of setting up a network of analysts (Network Italiano di Epidemic Intelligence - Italian Network of Epidemic Intelligence) nominated by regional authorities among subject-matter experts employed within the national health system at the national, regional and local level. Analysts of the Italian Network of Epidemic Intelligence work in rotating teams. Each day they screen news items, identifying those that are relevant to the surveillance focus (e.g., cases or clusters of infectious diseases in Italy or/and signs and symptoms in unexpected frequency) that are called signals. Signals are then individually risk assessed by the analysts using a common methodology (Intelligence and Miglietta, 2022) to identify those of public health relevance that are called events and that are then reported. At any given time, analysis is required to manually screen thousands of news items, reject irrelevant ones categorize signals, and assess them as events. Especially the screening phase of this work is extremely time-consuming and resource intensive and this undermines the long-term sustainability of this surveillance system. The integration of NLP techniques brings a significant contribution to the quality enhancement of the Epidemic Intelligence efforts: (*i*) *Enhanced Text search capabilities*, enabling the processing of larger text data volumes to uncover emerging threats, thus aiding to identify otherwise overlooked information; (*ii*) *Reduced monitoring time*, through the automation of routine monitoring tasks, allowing to allocate more time to complex and strategic analyses; and (*iii*) *Improved accuracy*, fostering for well-informed and documented decisions.

This paper introduces an advanced framework designed to harness the capabilities of Large Language Models (LLMs) for Epidemic Intelligence, addressing the specific challenges of identifying and categorizing emergent socio-political phenomena within the context of health crises, notably the COVID-19 pandemic. The objective is not to replace the analyst with an opaque, black-box approach for topic discovery but to ensure each analytical step is as self-explanatory as possible. By producing linguistically informed insights, we aim to elucidate the rationale behind the interpretation, for the analysts not familiar with subject matters about inherently emerging phenomena. The methodology begins with the user need to be defined as a set of seed terms to delineate a high-level conceptual document perimeter. This produces a corpus of retrieved news, specifically focused on the general need. Then, linguistic triples are generated to capture fine-grain concepts (e.g. specific activities or clinical concepts) implicit in the corpus. These triples can then be validated manually against the collected news, culminating in their automatic translation into prompts. Notice that the user can continuously fine-tune the system's proposed prompts, further customizing the analysis to his own specific needs. The overall process fosters a collaborative environment where the agent's intelligence and the analyst's domain expertise cooperate according to the investigative goals. This enhances the accuracy and coverage of the resulting Epidemic Intelligence activities. Preliminary results from our empirical investigation confirm the significant benefits of our workflow's capability in accurately mapping news articles to pertinent fine-grain topics. In the remaining, Section 2 reports related work, Section 3 presents the proposed workflow, Section 4 discusses the empirical evaluation, while Section 5 presents the conclusions.

2. Related Work

The use of NLP and text mining techniques in order to extract relevant information from vast amounts of text data available on the internet, thus allowing the identification of relevant epidemiological events, has been extensively studied in the previous years (see (O'Shea, 2017) for a systematic review of proposals dating a few years ago). For instance, the Medical Information System (MedISys) (Rortais et al., 2010) supports the timely detection of emerging diseases by crawling online news articles and applying hierarchical clustering algorithms to classify them into predefined categories. The Pattern-based Understanding and Learning System (PULS) (Yangarber and Steinberger, 2009) extends the MedISys by applying Natural Language Processing. The Global Health Monitor system (Collier et al., 2008) uses instead an ontology-based approach to text mining text data from the web to detect and track infectious disease outbreaks.

Text classification is a fundamental approach to the identification of relevant events. After early works applying classical machine learning approaches (Kowsari et al., 2019) (Khan et al., 2010), deep learning architectures introduced a new set of general methods (Minaee et al., 2021), (Luan and Lin, 2019) for text and news classification. Interest in the topic received a boost with the outbreak of the COVID-19 pandemic (Al-Garadi et al., 2022), (Raza et al., 2022), (Raza and Schwartz, 2023). Moreover, the advent of the attention mechanism in neural networks (Vaswani et al., 2017) and the adoption of transformer-based encoders

(Devlin et al., 2019), (Gillioz et al., 2020) made it possible effective information extraction from texts (Gupta et al., 2022), (Choudhary et al., 2023) as well as document classification (Li et al., 2022), (Deping et al., 2021), (Kaliyar et al., 2021). The use of transformer architectures, and the related Large Language Models, to news classification is an active research area (see for example (Khosa et al., 2023), (Deping et al., 2021), (Santana et al., 2022), (Gunes and Florczak, 2023)). However, the evaluation of the use of such approaches to the medical, and in particular in the epidemiological, field has been performed only quite recently (Wang et al., 2023), (Adaszewski et al., 2021) and it is still in its infancy. Unlike traditional approaches that might rely on probabilistic distributions akin to Topic Models (Blei et al., 2003), (Blei and Lafferty, 2009), (Mcauliffe and Blei, 2007),(Churchill and Singh, 2022), our method aims to provide explicit support to analysts. It does so by identifying distinct thematic areas and generating clear, actionable statements for each topic, such as "*A news article pertains to this topic if it addresses* . . . ". These statements are straightforward triggers for a Zero-shot Classification mechanism (Yin et al., 2019), effectively matching news articles to meticulously defined topics.

3. Automatic Topic Discovery and Classification

This section describes the workflow from initial data gathering to the application of Zero-shot classifiers for identifying and categorizing emergent socio-political phenomena.

The **Data Gathering** phase initiates our workflow, where analysts input keywords, such as “*Coronavirus outbreak*” or “*Covid contagion*”, to guide the system in collecting news articles relevant to the defined scope. This stage aims to cast a wide net to ensure comprehensive coverage, setting the stage for subsequent refinement and analysis.

The cornerstone of our methodology lies in harnessing the initial intuition of analysts to seed the discovery of diverse topics within the vast landscape of collected data. This phase, **Topic Discovery**, begins with the identification of *seed-words*, terms that encapsulate the essence of the analyst’s goal. In the Topic Discovery phase, seed words serve as descriptors of a general domain of analysis and let the system suggest potential themes. The analysts can then refine or expand upon these suggestions. This interaction balances automation and human expertise, ensuring a form of both guided and nuanced analysis. Terms such as “*Covid*” and “*Hospital*” could serve as initial seeds, delimiting the text perimeter within the broader domain of public health and epidemic preparedness. The output of this phase is a set of specific concepts that should emerge directly from the news in the perimeter, different from abstract word distribution, often associated with traditional Topic Modeling approaches, (Blei et al., 2003; Abdelrazek et al., 2023). Our target is not generating probabilistic topic models but conceptual sub-topics that are immediately comprehensible to an analyst. In this work, we thus conceive a topic as a collection of assertions such as “*This text discusses a concerning increase in infections led to a rise in Covid patients*." or “*The text discusses the monitoring of the increase in deaths caused by Covid*” for a topic possibly named “PANDEMIC PEAKS”. Another such topic as “COVID PATIENT CARE” could be inspired by assertions like “*The text addresses the challenges hospitals face in accommodating new patients*” or “*The text discusses hospitals continuing to vaccinate patients for COVID prevention*”. This approach makes topic interpretation easy, given the assertions and the title. Whether a news article aligns with a specific topic depends upon the verification of its assertions. A news article is assigned to a topic proportionally to how much the system judges one of its assertions to be true. Multiple true assertions incrementally contribute to the overall confidence in the topic. Crucially, as assertions are interpreted first by the analysts, he may wish to refine a topic by adjusting, deleting, or introducing new assertions.

In the final phase, we apply **Zero-shot Classification** through Large Language Models (LLMs) to conclusively label news articles with the specific topics identified earlier, avoiding model fine-tuning. This approach, grounded in natural language inference (Yin et al., 2019), exploits logical alignment between text and topic-defining assertions in the form of prompts. Notice that this enhances article-topic associations, beyond mere categorization, as assertions also provide explanations of individual classification inferences.

Upon completion of the workflow, once topics and corresponding prompts are made available, all news can be classified accordingly. Users can then exploit specific topics, e.g.,“PANDEMIC PEAKS”, as news filters, based on the metadata associated with prompts. This enables further analyses, such as focused news retrieval, filtering, and aggregation.

3.1. Data Gathering

The foundation of our approach begins with the **Data Gathering** phase, a crucial step designed to amass a comprehensive corpus of news articles pertinent to specific events or phenomena. For instance, an analyst may conduct an inquiry into the societal impact of afflictions, such as the Coronavirus within the Italian territory over the past fortnight. Accordingly, by providing pivotal terms such as “*Coronavirus outbreak*” and/or “*Covid contagion*” (possibly accompanied by time constraints) the process autonomously assembles a specific document collection, through the systematic extraction of Web news articles. In the initial phase, broad or generic query terms are used to maximize coverage which means extending article retrieval also to possibly irrelevant data. This strategy aims to extend the corpus, leaving its refining and validating to subsequent, more informed, stages. To facilitate this process, we developed a dedicated crawling service, aimed at collecting unstructured data using Google News as a primary but not exclusive source.

3.2. Topic Discovery

The Topic Discovery phase is pivotal in our workflow, aimed at moving from a small set of seed words to a possibly comprehensive collection of specific epidemic topics. Consider the previous example of an analyst inputting seeds such as "*Covid*" and "*Hospitals*". The initial step of lexical expansion endeavors to broaden the analyst's query using Word Space models, such as those created by the Contextual Bag of Words (CBOW) model implemented in `Word2Vec` (Mikolov et al., 2013). This distributional representation embeds terms within high-dimensional spaces where metric distances mirror paradigmatic relations, like quasi-synonymy, facilitating the exploration of related lexical fields (Sahlgren, 2006). Expanding upon the initial seed terms involves selecting the terms closest to each seed, aiming to broaden the initial semantics for topic generation. For example, the most similar words to "*Hospital*" are "*clinic*" and "*infirmary*", while "*coronavirus*" and "*pandemic*" are the corresponding words for "*Covid*". These entries offer a useful semantic expansion for the analysts to explore related themes. However, complex prompts for classification (i.e. assertions) require more informative linguistic structures corresponding to concepts, such as biological or clinical entities or events. This requires not just the selection of individual relevant terms but also complex well-formed definitions.

In this work, to automatically discover meaningful statements, such as "*This text discusses a concerning increase in infections led to a rise in Covid patients*.", we employ a form of grammatically controlled lexical expansion. From seed terms, we aim to generate structured forms like Subject-Verb-Object (SVO) triples, which can be easily transformed into coherent sentences. This approach ensures that an expansion can be easily understood by the analysis, but also facilitates the automatic creation of meaningful textual prompts. We call this step the **Linguistic Triple Generation** process. Assuming the inserted seeds are nouns, our process begins by identifying the set of e_v verbs closest to them. We use cosine similarity in the employed Word Space, also assuming that a seed noun can function either as a subject or as an object of the selected verb. For each such verb v, we then in turn retrieve the set of e_n nouns closest to v to completely fill an SVO structure. This approach ensures that the expansion from seed words to SVO triples is both deliberate and meaningful, providing a semantically rich lexicon from which complex thematic prompts can be derived. For instance, from the seed "*Covid*", we may derive closely related verbs such as "*record*" and "*infect*", while "*hospital*" might lead us to "*admit*" or "*vaccinate*". The expansion from verbs to nouns allows for the generation of SVO triples by further associating these verbs with relevant nouns: "*infect*" leads to "*patients*", "*lung*", and "*infections*"; "*record*" to "*deaths*", "*recovered*", and "*amount*". These expansions facilitate the construction of SVO triples such as ("*Covid*", "*record*", "*deaths*"), ("*Covid*", "*record*", "*recovered*"), ("*Covid*", "*increase*", "*infections*"), ("*Covid*", "*infect*", "*patients*"), ("*Covid*", "*infect*", "*lung*"), ("*Covid*", "*record*", "*infections*"), ("*Hospitals*", "*admit*", "*patients*"), ("*Hospitals*", "*vaccinate*", "*patients*"), ("*Covid*", "*record*", "*amount*"), and ("*Covid*", "*increase*", "*quantity*"), ...

Obviously, the growth of the number of triples given e_s seeds alongside e_v verbs and e_n nouns impacts significantly on complexity. However, the news collection provided by the Data Gathering phase is crucial in filtering triples whose frequency in the corpus is too low, e.g. below a threshold of τ sentences. We can thus manage the proliferation of triples. This approach integrates the semantics of the wordspace with distributional information related to the topics implicitly expressed by the gathered collection.

After the generation of SVO triples, the workflow progresses to **Triple Clustering**, a crucial step designed to detect k distinct thematic areas relevant to the analyst's interests. Notice that each triple is defined in a metric space depending on its Compositional Distributional Semantics (Mitchell and Lapata, 2008). By representing triples as centroids of their constitutive vectors we may map triples in the same wordspace as the lexical entries. This representation supports the clustering of triples whereas the compositional nature of the mapping is useful to preserve semantic proximity, i.e. relatedness between triples. The clustering (via a k-mean-like algorithm) operates in the structured space and induces coherent groups. Each cluster emerges as a thematic entity, characterized by a given unique narrative thread, but described by the semantic proximity among its constituent triples. As an example, applying k-means with $k = 3$ to the SVO triples mentioned above, we derive the following groups: C_1 = {("*Covid*", "*increase*", "*infections*"), ("*Covid*", "*record*", "*deaths*"), ("*Covid*", "*record*", "*recovered*")}, which encapsulates the theme of the increase in Covid-related infections. C_2 = {("*Hospitals*", "*admit*", "*patients*"), ("*Hospitals*", "*vaccinate*", "*patients*"), ("*Covid*", "*infect*", "*patients*"), ("*Covid*", "*infect*", "*lung*"), ("*Hospital*", "*treat*", "*cure*")}, focusing on hospital responses to Covid, including admissions, vaccinations, and treatments. C_3 = {("*Covid*", "*increase*", "*quantity*"), ("*Covid*", "*record*", "*amount*")}, centering on quantitative aspects of the Covid pandemic.

In the provided toy example, the number of triples and clusters is modest, but one can easily envision scenarios with significantly larger outcomes. Moreover, given the limited textual context, triples may emerge redundantly within a given cluster. However triple similarity in the metric space (modeling for example too similar subjects or objects across triples) can be used to automatize a further stage, called **Triple Pruning**. It ensures the satisfaction of some constraints onto triples: (*i*) each triple must be locally informative (provide high levels of *inner novelty*), (*ii*) triples within the same cluster must exhibit large diversity (*outer novelty*), and (*iii*) triples must be *relevant* within the collection of retrieved documents. Ranking triples refines clustering results, enhancing topic specificity and relevance.

Formally, we define a generic triple of terms such that $t_i = (\mathcal{S}_i, \mathcal{V}_i, \mathcal{O}_i)$ where $\mathcal{S}$ denotes the subject, $\mathcal{V}$ denotes the verb and $\mathcal{O}$ denotes the object, each represented by a corresponding embedding vector $\mathbf{s}_i$, $\mathbf{v}_i$, $\mathbf{o}_i$ in a normalized space, i.e., $\|\mathbf{s}_i\| = \|\mathbf{v}_i\| = \|\mathbf{o}_i\| = 1$. After the system has generated the clusters, the selection procedure of best informative triples within a cluster C_j ($j \in 1, \ldots, m$) is set up by combining the semantic signal provided by the documents and the terms of the triples.

First of all, we introduce a cluster such that $C = \{(t_i, w_i)\}$, where w_i is a semantic weighting function for t_i that will be hereafter defined. In order to pick up the triples that provide additional semantic information, we introduce the `Inner Novelty` with the aim of selecting only triples exhibiting meaningful signals through higher internal information heterogeneity. Let

$$in(t_i) = in(\mathcal{S}_i, \mathcal{V}_i, \mathcal{O}_i) =$$
$$= 1 - \Big(\beta^{\mathcal{SV}}(\mathbf{s_i} \cdot \mathbf{v_i}) + \beta^{\mathcal{SO}}(\mathbf{s}_i \cdot \mathbf{o}_i) + \beta^{\mathcal{VO}}(\mathbf{v}_i \cdot \mathbf{o}_i)\Big)$$

with $\beta^{\mathcal{SV}}, \beta^{\mathcal{SO}}, \beta^{\mathcal{VO}} \in \Re^+$, such that $\beta^{\mathcal{SV}} + \beta^{\mathcal{SO}} + \beta^{\mathcal{VO}} = 1$. In this scenario, we postulate that a triple such as ("*Hospital*", "*treat*", "*cure*") might exhibit low *Inner novelty*, contributing minimally to the analysis due to the high similarity between "*treat*" and "*cure*". The object in this case adds little to the action's significance, leading us to consider its utility in the analysis as marginal. An additional measure is the `Outer Novelty` that captures the relevance of the semantic signal provided by a triple, evaluating the diversity between pairs of triples. Let

$$nov(t_i, t_h) = nov\Big((\mathcal{S}_i, \mathcal{V}_i, \mathcal{O}_i), (\mathcal{S}_h, \mathcal{V}_h, \mathcal{O}_h)\Big)$$
$$= 1 - \Big(\gamma^S(\mathbf{s}_i \cdot \mathbf{s}_h) + \gamma^V(\mathbf{v}_i \cdot \mathbf{v}_h) + \gamma^O(\mathbf{o}_i \cdot \mathbf{o}_h)\Big)$$

and with $\gamma^S, \gamma^V, \gamma^O \in \Re^+$ that regulate the *term-wise similarity*, of pairs of triples, with and $\gamma^S + \gamma^V + \gamma^O = 1$. Then, we compute the *Outer Novelty* of a triple relative to a set C of other triples already selected such that:

$$on_i(C) = \begin{cases} \min_{t_h \in C} \; nov(t_i, t_h) & \text{if } C \neq \varnothing \\ 1 & \text{otherwise} \end{cases}$$

where $i \neq h$ and C is the set that contains already chosen triples. For example, against the cluster $C = \{$("*Hospital*", "*treat*", "*case*")$\}$, we hypothesize that a triple such as ("*Hospital*", "*treat*", "*patient*") has a lower *Outer Novelty* when it is less frequent in the collected news corpus than the member of C.

Given a cluster C made of triples ranked according to the defined novelty weights, the overall weight $w_i(C)$ of a triple t_i is:

$$w_i(C) = \log df_i \cdot in_i \cdot on_i(C) \qquad (1)$$

where $\log df_i$ denotes the logarithm of triple's *document frequency*.

The following algorithm in 1 computes the target set C^* of the most informative tuples from a set C, i.e.

$$C^* = \textsc{bestTriples}(C) \subseteq C$$

Obviously, $C^* = \varnothing$ is the initial set of selected triples.

Every generic element $x_i \in C$ corresponds to a 4-tuple

$$x_i = \langle t_i, \log df_i, in_i, on_i(C)\rangle$$

where at the beginning $\forall x_i \in C \; on_i = 1$ as $C^* = \varnothing$, and $w_i(C^*) = \log df_i \cdot in_i$.

Algorithm 1 Selection of best triples

```
procedure bestTriples($C$)
    $C^* \leftarrow \varnothing$
    $R = \{x_i \in C \mid \log df_i > \tau\}$
    while $R \neq \varnothing$ do
        $x \leftarrow \mathrm{argmax}_{\forall x_i \in R}, w_i(C^*)$
        $C^* \leftarrow C^* \cup \{x\}$
                ▷ pruning the less informative triples
                        ▷ according to Outer Novelty
        $R = \{x_i \in R \mid t_i \neq t \wedge$
                $\min(on_i(C^*), nov(t_i, t)) > \epsilon\}$
    end while
    return $C^*$
end procedure
```

Notice that τ and ϵ as the two parameters of the algorithm: τ is the lower bound of frequencies needed to discard too rare triples that are not relevant the a news collection, while $\epsilon \in [0, 1]$ regulate the overall novelty of a new triple against the already selected ones. Moreover, selecting the $\min(on_i(C^*), nov(t_i, t))$ requires constant time thanks to caching. In fact, $\forall x_i \in R$, $on_i(C^*)$ changes at each step as $x_i = \langle t_i, \log df_i, in_i, \min(on_i(C^*), nov(t_i, t))\rangle$, according to the new C^*.

In essence, for each cluster, the algorithm initially selects the triple that simultaneously maximizes *document frequency* in the news corpus and *Inner Novelty* from the set of viable candidates. Subsequently, it iterates the selection process. At each iteration, a newly selected triple must itself exhibit high relevance and *Inner Novelty*, while also demonstrating substantial *Outer Novelty* concerning the previously selected triples.

Keeping in mind the $C1, C2, C3$ clusters from the example described earlier, the outcome of the best triples selection process leads to the following situation: $C1$ remains intact, while in $C2$, redundant triples like ("*Covid*", "*infect*", "*patients*") and ("*Hospital*", "*treat*", "*cure*") are pruned. Afterwards the user can actively engage in the analysis by potentially removing triples, adding new ones to better contextualize each cluster's analysis, deleting clusters or even suggesting additional clusters for useful facets of the analysis overlooked by the system. Let's assume the user eliminates $C3$ as it is of little interest for user analysis purposes.

The final step involves transforming the identified SVO triples into prompts suitable for a Zero-shot classification system, achieved through the utilization of a Large Language Model (LLM) (Touvron et al., 2023; Jiang et al., 2023; OpenAI, 2023). For each cluster, all selected triples are fed into an LLM alongside a prompt designed to convert them into assertive forms that can define concepts. This process may involve synthesizing and aptly combining the contributions of all triples within the cluster. An important factor is the variable number of triples per cluster. However, the LLM can also be tasked with generating m distinct assertions, accommodating the diversity and breadth of information captured by the cluster's triples. This step of **Prompt Generation** is implemented by making a request to the LLM through prompts such as: "*Consider the input triples consisting of 3 terms. I need you to generate 3 sentences, where each sentence serves to 'describe' the triples. The sentences you generate must follow the format 'This news is about <complete>' and MUST NOT exceed 12 words in length. The triples are as follows:*". After analyzing LLMs such as LLAMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), and GPT-4 (OpenAI, 2023), we have chosen to employ GPT-4 for the quality of the generated prompts. Currently, our selection is solely based on empirical analyses to determine the "best" LLM: a more comprehensive examination is still underway. For instance, using the above triples from the example cluster C_2, statements such as "*This news is about hospitals admitting patients during the pandemic.*" or "*This news is about hospitals vaccinating patients to combat illnesses.*" or "*This news is about Covid infecting lungs, posing respiratory risks.*" are generated. A similar strategy is applied to provide a title for the cluster. The request "*Write a name that precisely describes the following set of word triples. Please respond with ONLY ONE name consisting of ONLY 2 or 3 WORDS, the triples are as follows:*" is adopted. Taking the example further, the names of the following emerging topics are generated starting, respectively, from the sets $C1, C2$: "COVID INFECTIONS", "COVID HEALTHCARE DYNAMICS".

3.3. Zero-shot Classification

In the final step, **Zero-shot Classification** is applied to match the amassed news articles with the identified topics, significantly enhancing the article metadata granularity associated with specific thematic facets. Inspired by (Yin et al., 2019), our approach employs a Zero-shot classifier built upon Large Language Models (LLMs) that requires no fine-tuning. This method aligns articles to topics by treating the task as one of natural language inference (NLI), following paradigms established in (Dagan et al., 2013; Bowman et al., 2015). Specifically, it assigns each article (treated as the premise in a classical textual entailment task) to a topic based on the degree to which the LLM deems the text to logically infer the topic's defining prompt (treated as the hypothesis corresponding to the premise). This process ensures articles are categorically aligned with topics through a logical inference mechanism, offering a precise and context-aware topic association. In our approach, we have employed a model based on BART (Lewis et al., 2020), trained on the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018), a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information across various genres, including news[3]. The process involves presenting a news article and, in turn, each prompt of each generated topic, calculating a score, in terms of probability of the "truth" of the entailment. This evaluation allows us to determine which prompts are activated by the news article along with the corresponding probability.

For instance, let's consider a scenario where we

[3] `https://huggingface.co/facebook/bart-large-mnli`

have a news item: "*Rome, Italy - As the Eternal City faces a concerning rise in COVID-19 cases, Rome's hospitals are stepping up their response with an aggressive vaccination campaign. The regional health authorities have reported a significant increase in hospital admissions due to COVID-19, prompting a swift reaction from the medical community. ...*" each prompt related to the overarching theme of 'COVID HEALTHCARE DYNAMICS" is coupled with the original news excerpt to form a composite input for the Zero-shot classifier. For instance, the news snippet "*Rome, Italy - As the Eternal City faces a concerning rise in COVID-19 cases ...*" is appended with a separator `[SEP]` followed by a prompt such as "*This news is about hospitals vaccinating patients to combat illnesses.*" This methodology allows the classifier to generate an internal representation of the combined input and evaluate it against predefined categories that determine whether the second statement is implied by the first. For each pairing, the classifier assigns an entailment probability score, reflecting the relevance of the prompt to the original text within the context of the selected theme. From the examples provided, the prompt stating "*This news is about hospitals vaccinating patients to combat illnesses.*" yielded an entailment score of 0.87, indicating a strong connection to the topic at hand. Conversely, the statement "*This news is about hospitals vaccinating patients to combat illnesses.*" received a lower score of 0.68, suggesting it is related but does not capture the core aspects of the news item as effectively. The prompt "*This news is about Covid infecting lungs, posing respiratory risks.*" regarding COVID-19 affecting the lungs with a score of 0.05 was not triggered, likely due to the absence of direct reference to lung infections in the news piece. From this small sample, it becomes clear that the prompt with a score of 0.87 is identified as a significant trigger for the corresponding theme cluster, effectively encapsulating the primary focus of the news item. The prompt with a score of 0.68, while relevant, may not fully capture the salient aspects of the scenario. In contrast, the prompt with a score of 0.01 is deemed unrelated, highlighting the classifier's ability to discern relevance based on the specificity of the information provided about the lungs.

Subsequently, an overall probability for the topic itself is based on a composition of each of those derived from the m individual prompts. Given m possible prompts and their corresponding entailment probabilities, we investigated the following policies: **Maximum Probability**: Assigns the maximum score observed across prompts to the topic. This greedy approach, however, is exposed to imprecise outliers or spikes; **Top-Best Average**: Returns the average score from the top b scores among the topic's prompts. This method is more robust against outliers; **Topic Average**: Computes the average score across all m prompts. The topic is activated only when, on average, the news article verifies all aspects described by the prompts.

Considering two topic clusters with different sets of entailment scores for their prompts—0.87, 0.68, 0.05 for the first cluster, and 0.88, 0.19, 0.02 for a second hypothetical cluster—we evaluate topic assignment policies. The Maximum Probability method, focusing on the highest score per cluster, initially suggests the second cluster (0.88) as more relevant than the first (0.87). Yet, this overlooks the composite thematic relevance of all prompts. Adopting the Top-Best Average method with $b = 2$, we find that the average of the top two scores offers a more nuanced perspective: 0.775 for the first cluster versus 0.535 for the second. This method, by mitigating outlier influence, indicates a stronger alignment of the first cluster with the news, demonstrating the importance of a broader assessment beyond single peak scores for more accurate topic relevance evaluation.

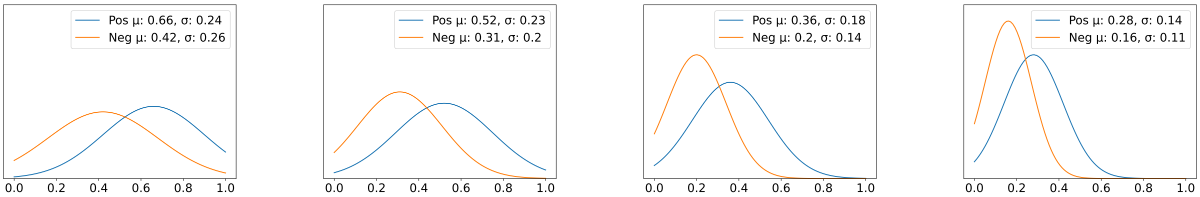
4. Experimental Evaluation

The experimental section is designed to evaluate the effectiveness of our workflow, which includes a series of complex steps ranging from web page collection, and topic generation, to the 0-shot classification of retrieved news articles. The core hypothesis of this experiment is that the model is effective if it can accurately re-process news articles previously categorized by the analysts under specific themes. It should regenerate consistent topics and associate the news articles to the topics coherently with the analysts' original choices. The workflow's success is measured according to the comparison between probabilities assigned to the method and those expected, as derived from the analyst annotation. First, the model is used to generate distinct topics based on the input analyst classes: in this way, the initial association between news articles and their original theme is known. Subsequently, if the system classifies the news articles according to the newly generated topics, and these topics align closely with the input themes, then the model is performing in harmony with the analyst's expectation.

Experimental setup. During the surveillance period from February 2020 to September 2022, analysts concentrated on monitoring COVID-19 outbreaks across various epidemiological settings. Within this timeframe, ISS experts manually categorized a total of 2,254 news articles, associated to "COVID VARIANTS" (313 news), "NURSING HOMES OUTBREAKS" (682), "HOSPITAL OUTBREAKS" (417), "SCHOOL OUTBREAKS" (574) and FAMILY/FRIEND

OUTBREAKS" (268). It's important to note that for their analysis, the analysts focused on a subset of topics at a time and could only associate a news article with a subset of those that manifested, such as a news piece discussing an outbreak in a hospital and then in a nursing home. To generate the topics and attempt to replicate the analysis performed on the documents, we selected a couple of seed words for each input theme: "COVID VARIANTS" corresponds to "*variant*" and "*English variant*", "*omicron variant*", "*delta variant*"; "NURSING HOMES" is represented by "*healthcare worker*" and "*elderly*", "*healthcare residence*"; for "HOSPITAL OUTBREAKS" the seeds are "*Hospital*", "*department*", "*patient*", "*contagion*"; "SCHOOL OUTBREAKS" is expressed by "*school*", "*remote learning*", "*student*", "*teacher*"; and "FAMILY/FRIEND OUTBREAKS" corresponds to "*parent*", "*family member*", "*condominium*", "*relative*". To generate the topics, seed words were selected for each input theme, and parameters were evaluated to model the process closely. For the linguistic triple generation step, we expanded the search to include a broad range of verbs and nouns, specifically setting $e_v = 150$ for verbs and $e_n = 100$ for nouns. The process was carried out for each topic individually, applying Triple Clustering with values of $k = 2, 3, 4$. This range was chosen to prevent an overwhelming proliferation of clusters while still capturing a diverse array of topics. To mimic the analytical phase typically performed by an analyst, clusters were selected based on which k values yielded the most coherent and relevant topics. The pruning process, as outlined in Algorithm 1, was guided by specific parameters for measuring novelty within and between the generated triples. For *Inner Novelty*, the parameters were set as $\beta^{\mathcal{SV}} = 0.25$, $\beta^{\mathcal{SO}} = 0.65$, and $\beta^{\mathcal{VO}} = 0.10$. These parameters helped assess how distinct each SVO triple argument was from the others, ensuring a richer semantic variety within the topics. *Outer Novelty*, which measures the diversity between triples, was regulated with $\gamma^{\mathcal{S}} = 0.25$, $\gamma^{\mathcal{O}} = 0.10$, and $\gamma^{\mathcal{V}} = 0.65$. The threshold $\epsilon = 0.3$ in the algorithm ensured that only the triples significantly different from those already selected were chosen, simulating an analyst's decision-making process in refining the topics. Ultimately, this approach led to the generation of 11 distinct clusters for 44 total prompts. For each of these generated topics, the system requested m prompts to facilitate the prompt generation phase. The detailed list of clusters and their respective prompts, which form the backbone of our topic generation and classification process, is provided in Appendix A. Finally, the individual news articles were associated with topics by applying the Zero-shot classifier, determining an association score for each of the 44 prompts.

Results. To assess the quality of the classification system in accurately associating news articles with the correct topics, aligning with the input themes, we examined the system's ability to assign a score to each news article for the single class identified by the analyst (referred to as the positive class) used to distinguish it through the probabilities associated with unrelated themes (referred to as negative classes). Notice that, as we cannot rely on the hypothesis that each news article belongs to only one correct class, studying the overall behaviors of the method requires studying its probability distributions. Probability scores depend on the three policies defined for Zero-shot classification. Another reference policy can be added, allowing a news article to be associated with the average value of all prompts generated from the seeds related to that input topic or associated class, thus called the **Class average** policy. In Figure 1, the distributions of scores associated with each generated topic are shown for each defined policy. The distributions of scores for positive (pos) and negative (neg) classes are depicted in blue and orange respectively, by assuming a normal distribution given by its mean (μ) and standard deviation (σ). It is interesting to note how, for each policy, the system is shown capable of separating the distributions, confirming its ability to re-associate news articles with topics consistently with the original analysts' classifications. Clearly, the smaller the intersection between the two curves, the greater the system's ability to replicate the work of the analysts. It is interesting to note that in the policy called Best Probability, selecting the prompt that maximizes the association probability by observing only one prompt in the class has a high mean for positives ($\mu = 0.66$ in Figure 1 (a)). Unfortunately, there are also several spikes for the discarded classes with a mean ($\mu = 0.42$). This could be due to other topics, generated for other input themes, being often activated, apparently in disagreement with the analyst. However, generally, analysts have only selected one major topic, and it is possible they discarded other topics that are part of the article discussion. Moving to Figure 1 (b), by taking the average of the first $b = 2$ in the Top-Best Average policy, the average values predictably tend to decrease, and interestingly, the standard deviation also decreases because the spikes are mitigated. This phenomenon is further evident in Figure 1 (c) where averaging all prompts in a generated topic forces a topic to cover practically all sub-themes discussed in the topic, according to the Topic Average policy. Pushing the situation further in Figure 1 (d), where all prompts of all topics generated for a class are averaged, it is evident that the means significantly decrease (dropping to $\mu = 0.28$ for positives) but the negatives are practically nullified (with a $\mu = 0.16$) and the standard deviations are greatly reduced, also reducing the



(a) Best probability (b) Top-Best Average (c) Topic Average (d) Class Average

Figure 1: Distributions of scores for generated topics per policy, showing positive and negative classes with the corresponding means (μ) and standard deviations (σ).

intersection area between the curves. This suggests the importance of having multiple prompts in individual topics to make the system more stable and avoid associations due to possible spikes.

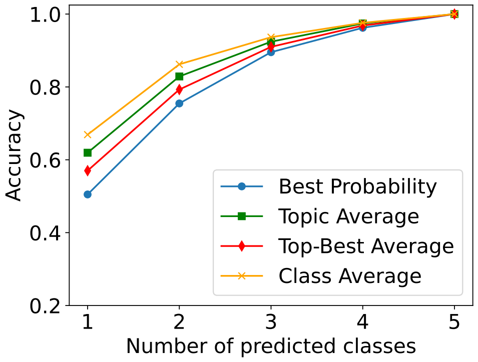


Figure 2: Accuracy

We then assessed the system's ability to reclassify individual news articles, ranking the classes based on the scores suggested by the system's topics, as shown in Figure 2: the y-axis represents accuracy, while the x-axis lists the classes. As mentioned, it's not necessarily the case that if the first class proposed by the system is incorrect, the system's handling is flawed. Therefore, we calculated accuracy for correctly identifying just the first class (c=1), but also for c=2, then c=3, up to c=5 (where, by construction, accuracy is 1.0). The Best Probability policy confirms its vulnerability to spikes, with accuracy at c=1 of 0.51 and c=2 of 0.75. This is significantly compensated by the various policies, as evidenced by the distributions in Figure 1, with Top-best Avg at c=1 having an accuracy of 0.57, Topic Average at c=1 rising to 0.62, and then Class Average at c=1 reaching 0.67. Interestingly, more than 88% of the news articles are reassociated with the analysts' topics if only the first two system suggestions are considered under the Class Average policy. The experimental results emphasize the benefits of our approach, in accurately mapping news articles to relevant topics, demonstrating the workflow's potential to streamline Epidemic Intelligence processes. An error analysis is reported in Appendix B.

5. Conclusion

This study presented a novel framework utilizing Large Language Models (LLMs) to enhance Epidemic Intelligence through automated topic discovery and 0-shot classification. We aimed to address in this way the challenge of effectively identifying and categorizing potential health hazards, with a focus on the COVID-19 pandemic. Our methodology diverges from traditional probabilistic models by offering explicit analytical support through the generation of actionable topic statements, thereby facilitating a Zero-shot classification mechanism that accurately matches news articles to defined topics without resorting to fine-tuning. Our methodology, integrating a decoder LLM, faces potential limitations highlighted by (Huang et al., 2023), such as susceptibility to hallucinations. This affects the generation of prompts and topic names, which could lead to inaccurate descriptions of emerging arguments. Our approach, however, is not designed as an inflexible, fully automated system that diminishes the role of analysts. Rather, it is intended to enhance and enrich their analytical capabilities, promoting an interactive and collaborative exploration of data. Analysts maintain the capacity to modify outputs at any stage, enabling them to select relevant topics or refine prompts for Zero-shot classification, thereby ensuring a more accurate and insightful analysis.

The results from our experimental evaluation highlight the robustness and effectiveness of our workflow in aligning with the analytic processes traditionally employed by experts in Epidemic Intelligence. The implementation of multiple classification policies has demonstrated a significant improvement in the system's ability to accurately associate news articles with relevant topics. This advancement is evident in the increased accuracy rates and the reduction of classification errors, underscoring the system's capacity to handle complex thematic categorizations reliably.

Acknowledgements

The project realized with the technical and financial support of the Ministry of Health - CCM.

Bibliographical References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Stanislaw Adaszewski, Pascal Kuner, and Ralf J Jaeger. 2021. Automatic pharma news categorization. *arXiv preprint arXiv:2201.00688*.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, and Abeed Sarker. 2022. The role of natural language processing during the covid-19 pandemic: Health applications, opportunities, and challenges. *Healthcare*, 10(11).

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

David M Blei and John D Lafferty. 2009. Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7456–7463. AAAI Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Kevin Matthe Caramancion. 2023. News verifiers showdown: a comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. *arXiv preprint arXiv:2306.17176*.

Ambrish Choudhary, Mamatha Alugubelly, and Rupal Bhargava. 2023. A comparative study on transformer-based news summarization. In *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, pages 256–261. IEEE.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, et al. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Martina Del Manso, Daniele Petrone, Matteo Spuri, Chiara Sacco, Alberto Mateo Urdiales, Roberto Croci, Stefania Giannitelli, Patrizio Pezzotti, Daniele Mipatrini, Francesco Maraglino, et al. 2022. Il sistema di sorveglianza basato su eventi in italia dal 2009 al 2021: verso una intelligence di sanità pubblica. *Bollettino epidemiologico nazionale*.

Lin Deping, Wang Hongjuan, Liu Mengyang, and Li Pei. 2021. News text classification based on bidirectional encoder representation from transformers. In *2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, pages 137–140. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jerome Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38:367–378.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232.

Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method.

Ramanathan Guha, Rob McCool, and Eric Miller. 2003. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709. ACM.

Erkan Gunes and Christoffer Koch Florczak. 2023. Multiclass classification of policy documents with large language models. *arXiv preprint arXiv:2310.08167*.

Anushka Gupta, Diksha Chugh, Anjum, and Rahul Katarya. 2022. Automated news summarization using transformers. In *Sustainable Advanced Computing: Select Proceedings of ICSAC 2021*, pages 249–259. Springer.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction*, 2 edition. Springer.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Nancy Ide and Jean Véronis. 1990. Very large neural networks for word sense disambiguation. In *9th European Conference on Artificial Intelligence, ECAI 1990, Stockholm, Sweden, 1990*, pages 366–368.

Network Intelligence and Alessandro Miglietta. 2022. Istituto superiore di sanitá il sistema di sorveglianza basato su eventi in italia dal 2009 al 2021: verso una intelligence di sanitá pubblica. *Scientific reports of the Istituto superiore di sanitá*, pages 19–28.

Frederick Jelinek. 1998. *Statistical Methods for Speech Recognition (Language, Speech, and Communication)*. The MIT Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Daniel Jurafsky and James H. Martin. 2009. *Speech and language processing*, 2. ed., [pearson international edition] edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, London [u.a.].

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. 2010. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20.

Saima Khosa, Arif Mehmood, and Muhammad Rizwan. 2023. Unifying sentence transformer embedding and softmax voting ensemble for accurate news category prediction. *Computers*, 12(7):137.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, number 2 in Proceedings of Machine Learning Research, pages 1188–1196, Bejing, China. PMLR.

Michael Lebowitz. 1988. The use of memory in text processing. *Commun. ACM*, 31:1483–1502.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*, 3:111–132.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

Yuandong Luan and Shaofu Lin. 2019. Research on text classification based on cnn and lstm. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)*, pages 352–355. IEEE.

G. Madhu, Dr. A. Govardhan, and Dr. T. V. Rajinikanth. 2011. Intelligent semantic web search engines: A brief survey.

C.D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Jesse O'Shea. 2017. Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International journal of medical informatics*, 101:15–22.

C Paquet, D Coulombier, R Kaiser, and M Ciotti. 2006. Epidemic intelligence: a new framework for strengthening disease surveillance in europe. *Eurosurveillance*, 11(12):5–6.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. `https://github.com/musixmatchresearch/umberto`.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2020. Hopfield networks is all you need. Cite arxiv:2008.02217Comment: 10 pages (+ appendix); 12 figures; Blog: https://ml-jku.github.io/hopfield-layers/; GitHub: https://github.com/ml-jku/hopfield-layers.

Shaina Raza and Brian Schwartz. 2023. Constructing a disease database and using natural language processing to capture and standardize free text clinical information. *Scientific Reports*, 13(1):8591.

Shaina Raza, Brian Schwartz, and Laura C Rosella. 2022. Coquad: a covid-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC bioinformatics*, 23(1):1–28.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Flavia Riccardo, Martina Del Manso, Maria Grazia Caporali, Christian Napoli, Jens P Linge, Eleonora Mantica, Marco Verile, Alessandra Piatti, Maria Grazia Pompa, Loredana Vellucci, Virgilio Costanzo, Anan Judina Bastiampillai, Eugenia Gabrielli, Maria Gramegna, and Silvia Declich. 2016. Event-Based surveillance during EXPO milan 2015: Rationale, tools, procedures, and initial results. *Health Secur*, 14(3):161–172.

Flavia Riccardo, Mika Shigematsu, Chow Catherine, Mcknight Jason, Jens Linge, Brian Doherty, Maria Dente, Silvia Declich, Barker Mike, Barboza Philippe, Laetitia Vaillant, Donachie Alastair, Mawudeku Abla, Blench Michael, and Arthur Ray. 2014. Interfacing a biosurveillance portal and an international network of institutional analysts to detect biological threats. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, 12:325–36.

Agnès Rortais, Jenya Belyaeva, Monica Gemo, Erik Van der Goot, and Jens P Linge. 2010.

Medisys: An early-warning system for the detection of (re-) emerging food-and feed-borne hazards. *Food Research International*, 43(5):1553–1556.

Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.

Isabel N Santana, Raphael S Oliveira, and Erick GS Nascimento. 2022. Text classification of news using transformer-based models for portuguese. *Journal of Systemics, Cybernetics and Informatics*, 20(5):33–59.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2023. Zero-shot multi-label topic inference with sentence encoders and llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16218–16233.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. Cite arxiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yu Wang, Yuan Wang, Zhenwan Peng, Feifan Zhang, Luyao Zhou, and Fei Yang. 2023. Medical text classification based on the discriminative pre-training model and prompt-tuning. *Digital Health*, 9:20552076231193213.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

World Health Organization. 2008. Communicable disease alert and response for mass gatherings. In *Technical Workshop. Geneva, Switzerland*, pages 29–30.

World Health Organization. 2014. Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version. Technical report, World Health Organization.

Roman Yangarber and Ralf Steinberger. 2009. Automatic epidemiological surveillance from on-line news in medisys and puls. In *IMED-2009: International Meeting on Emerging Diseases and Surveillance (2009)*.

K Yasaswi, Vijaya Kumar Kambala, P Sai Pavan, M Sreya, and V Jasmika. 2022. News classification using natural language processing. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, pages 63–67. IEEE.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models.

A. Emerging Topics and prompts.

In this section, we present the topics and prompts that were generated and used in the experimental evaluation detailed in Section 4. For each cluster, we list the original prompts used in the experimentation and, for the convenience of the reader, their translation into English.

A.1. Topic: Covid Variants

C1: "*Diffusione delle varianti*" / "*Spread of variants*"

- "*Questa notizia riguarda una variante che richiede un accurato sequenziamento per monitorarne l'evoluzione.*" / "*This news is about a variant that requires careful sequencing to monitor its evolution.*"
- "*Questa notizia riguarda la capacità di sequenziare le varianti in casi complessi.*" / "*This news is about the ability to sequence variants in complex cases.*"
- "*Questa notizia riguarda come le varianti possono mutare il virus nel tempo.*" / "*This news is about how variants can mutate the virus over time.*"
- "*Questa notizia riguarda la diffusione delle varianti.*" / "*This news is about the spread of variants.*"

C2: "*Diffusione variante inglese*" / "*Spread of the English variant*"

- "*Questa notizia riguarda la vaccinazione per non incorrere in aumento di casi positivi.*" / "*This news is about vaccination to avoid an increase in positive cases.*"
- "*Questa notizia riguarda la necessità di vaccinare la popolazione contro la variante inglese.*" / "*This news is about the need to vaccinate the population against the English variant.*"
- "*Questa notizia riguarda la scoperta di un caso positivo relativo alla variante inglese.*" / "*This news is about the discovery of a positive case related to the English variant.*"
- "*Questa notizia riguarda la diffusione della variante inglese.*" / "*This news is about the spread of the English variant.*"

A.2. Topic: Nursing Homes Outbreaks

C3: "*Cura degli ospiti nelle RSA*" / "*Care of residents in nursing homes*"

- "*Questa notizia riguarda la necessità di ricoverare pazienti in una residenza sanitaria per anziani (RSA).*" / "*This news is about the need to hospitalize patients in a nursing home for the elderly.*"
- "*Questa notizia riguarda le procedure necessarie per accogliere un ospite nella residenza sanitaria per anziani (RSA).*" / "*This news is about the necessary procedures to welcome a guest into the nursing home for the elderly.*"
- "*Questa notizia riguarda le misure prese per isolare un ospite infetto nella residenza sanitaria per anziani (RSA).*" / "*This news is about the measures taken to isolate an infected guest in the nursing home for the elderly.*"
- "*Questa notizia riguarda la cura degli ospiti nelle RSA.*" / "*This news is about caring for guests in nursing homes for the elderly.*"

C4: "*Covid negli ospizi*" / "*Covid in nursing homes*"

- "*Questa notizia riguarda l'importante compito di vaccinare gli anziani contro il COVID.*" / "*This news is about the important task of vaccinating the elderly against COVID.*"
- "*Questa notizia riguarda l'importanza di isolare le residenze sanitarie per anziani (RSA) per prevenire focolai di COVID.*" / "*This news is about the importance of isolating nursing homes for the elderly to prevent COVID outbreaks.*"
- "*Questa notizia riguarda la necessità di ricoverare i pazienti COVID nelle residenze sanitarie per anziani (RSA) per cure adeguate.*" / "*This news is about the need to hospitalize COVID patients in nursing homes for the elderly for proper care.*"
- "*Questa notizia riguarda il covid negli ospizi.*" / "*This news is about covid in nursing homes.*"

A.3. Topic: Hospital Outbreaks

C5: "*Gestione dell'emergenza in ospedale*" / "*Emergency management in the hospital*"

- "*Questa notizia riguarda il paziente che risultò positivo al test presso la struttura ospedaliera.*" / "*This news is about the patient who tested positive at the hospital facility.*"
- "*Questa notizia riguarda il reparto di terapia intensiva all'interno dell'ospedale.*" / "*This news is about the intensive care unit within the hospital.*"
- "*Questa notizia riguarda un reparto dell'ospedale pronto a soccorrere ogni paziente.*" / "*This news is about a hospital department ready to assist every patient.*"
- "*Questa notizia riguarda la gestione dell'emergenza in ospedale.*" / "*This news is about the emergency management in the hospital.*"

C6: "*Impatto dell'epidemia: Ricovero ospedaliero*" / "*Impact of the epidemic: Hospitalization*"

- "*Questa notizia riguarda un ospedale che offre servizi di ricovero per i residenti.*" / "*This news is about a hospital that provides hospitalization services for residents.*"

- “*Questa notizia riguarda l’ospedale che si occupa di ricoverare i casi di COVID.*” / “*This news is about the hospital that takes care of hospitalizing COVID cases.*”
- “*Questa notizia riguarda un ospedale in una città, dove vengono ricoverate persone malate.*” / “*This news is about a hospital in a city where sick people are hospitalized.*”
- “*Questa notizia riguarda il ricovero ospedaliero dei casi.*” / “*This news is about the hospitalization of cases.*”

C7: “*Impatto dell’epidemia: Contagio in ospedale*” / “*Impact of the epidemic: Hospital contagion*”

- “*Questa notizia riguarda un paziente che risulta positivo al virus in ospedale.*” / “*This news is about a patient who tested positive for the virus in the hospital.*”
- “*Questa notizia riguarda il paziente che è risultato negativo al test.*” / “*This news is about the patient who tested negative.*”
- “*Questa notizia riguarda il contagio in ospedale che si può prevenire vaccinando ogni caso.*” / “*This news is about hospital contagion that can be prevented by vaccinating each case.*”
- “*Questa notizia riguarda i contagi che avvengono in ospedale.*” / “*This news is about the contagions that occur in the hospital.*”

A.4. Topic: School Outbreak

C8: “*Impatto della didattica a distanza*” / “*Impact of distance learning*”

- “*Questa notizia riguarda la chiusura della scuola conseguente all’attivazione della didattica a distanza.*” / “*This news is about the school closure following the activation of distance learning.*”
- “*Questa notizia riguarda l’importanza di rispettare le misure di sicurezza a scuola.*” / “*This news is about the importance of respecting safety measures at school.*”
- “*Questa notizia riguarda gli studenti che frequentano le scuole durante l’epidemia.*” / “*This news is about the students attending schools during the epidemic.*”
- “*Questa notizia riguarda l’impatto della didattica a distanza sulla scuola.*” / “*This news is about the impact of distance learning on school.*”

C9: “*Conseguenze dell’epidemia nella scuola*” / “*Consequences of the epidemic in schools*”

- “*Questa notizia riguarda l’insegnante che ha contribuito a contagiare un focolaio a scuola.*” / “*This news is about the teacher who contributed to infecting a outbreak at school.*”
- “*Questa notizia riguarda lo studente che risulta detenere il virus positivo.*” / “*This news is about the student who tested positive for the virus.*”
- “*Questa notizia riguarda le conseguenze dell’epidemia nella scuola.*” / “*This news is about the consequences of the epidemic in schools.*”

A.5. Topic: Family Outbreaks

C10: “*Impatto familiare della pandemia*” / “*Family impact of the pandemic*”

- “*Questa notizia riguarda un familiare che è stato contagiato in un focolaio.*” / “*This news is about a family member who was infected in an outbreak.*”
- “*Questa notizia riguarda un genitore che è stato vaccinato ma risulta positivo.*” / “*This news is about a parent who has been vaccinated but tested positive.*”
- “*Questa notizia riguarda un focolaio familiare che ha contagiato molte persone.*” / “*This news is about a family outbreak that has infected many people.*”
- “*Questa notizia riguarda l’impatto sulle famiglie della pandemia.*” / “*This news is about the impact on families of the pandemic.*”

C11: “*Vaccinazione familiare COVID*” / “*Family COVID vaccination*”

- “*Questa notizia riguarda la vaccinazione di gruppi di persone conoscenti.*” / “*This news is about the vaccination of groups of acquaintances.*”
- “*Questa notizia riguarda il timore di un familiare di contagiare altre persone con il coronavirus.*” / “*This news is about the fear of a family member of infecting other people with the coronavirus.*”
- “*Questa notizia riguarda l’importanza di vaccinare per proteggere la salute di parenti e amici.*” / “*This news is about the importance of vaccinating to protect the health of relatives and friends.*”
- “*Questa notizia riguarda la vaccinazione di famiglie contro il COVID.*” / “*This news is about the vaccination of families against COVID.*”

B. Error analysis

In the experimental evaluation reported in Section 4, the system manages to classify 67% of the news articles when a single class is proposed, and this figure rises to more than 88% if only the first two system suggestions are considered under the Class Average policy.

To understand the reason behind the discrepancy in these results, we examined the confusion matrix presented in Figure 3, which shows which classes were confused with each other. In the matrix, the analysts’ annotations are on the rows, and the system’s proposed associations when only one class is proposed are on the columns. Most of the classifications considered incorrect when only the first class is proposed (but corrected when two are proposed, indicating the second is correct) are news articles originally associated with the input topic Nursing Homes Outbreaks or School Outbreaks, which are classified as Family/Friend Outbreaks, or Nursing Homes Outbreaks classified as Hospital Outbreaks.

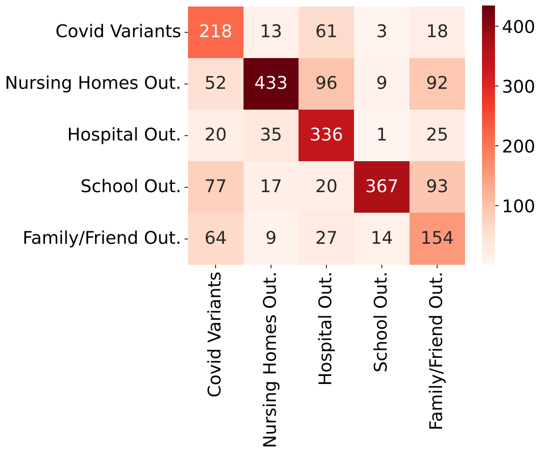


Figure 3: Confusion matrix

This observation implies that discussions about nursing homes may naturally reference families rather than hospitals, as these topics are inherently related. Motivated by this, we conducted a manual analysis of cases that would have been deemed errors. This deeper examination may reveal nuances in the data that automated classification initially overlooked, underscoring the complex interplay between seemingly distinct topics and the importance of contextual understanding in accurately categorizing news articles.

For example, consider the following news article:

> "*«La variante inglese era presente in un caso su cinque, ma nelle ultime due settimane la diffusione è molto aumentata». Così parlava tre giorni fa la dottoressa Antonia Ricci, direttrice generale dell'istituto zooprofilattico di Legnaro. Ora gli effetti si vedono e fanno paura: un focolaio è stato registrato in un gruppo di bambini di San Martino di Lupari e poi il contagio si è propagato in diverse zone dell'Alta Padovana costringendo l'Ulss ad una decisione drastica: quattro scuole chiuse. Il sindaco è stato invitato ad avviare la didattica a distanza per l'asilo Campagnalta e per la scuola elementare Sauro, dove è stato registrato il primo cluster. Stesso provvedimento anche per la materna Almarech di Villa del Conte e soprattutto per il liceo Tito Lucrezio Caro di Cittadella.*"

The system predicted the class as Covid Variants, primarily due to the strong activation of the prompt "*This news is about the discovery of a positive case related to the English variant*" which received a confidence score of 0.98. The sentence "*The prevalence of the English variant has significantly increased in the past two weeks, transitioning...*", would suggest a specific relevance of the mentioned prompt to the news at hand. Conversely, the original classification was School Outbreak, a category significantly represented by the prompt "*This news is about the school closure following the activation of distance learning*", achieving a confidence score of 0.84. In this case, it is worth noting how the prompt's score can be justified by the statement "*The mayor has been invited to initiate distance learning for the Campagnalta kindergarten and Sauro elementary school...*". Therefore, the system's prediction of Covid Variants as the primary class, with a higher confidence score for the prompt related to the English variant, reflects the significant mention of the variant in the article. However, the original class School Outbreak is also strongly represented, especially given the specific actions taken in response to the outbreaks in schools. This discrepancy suggests that while the variant's presence is a crucial aspect of the news, the article's core subject revolves around the implications of this presence on local schools. This case exemplifies the nuanced understanding required in classifying news articles, where multiple relevant themes can coexist, emphasizing the importance of considering all potential topics when classifying complex news stories.

Let us consider another example:

> "*Coronavirus. Ortona. Focolaio nella casa di riposo 'Don Bosco': 43 positivi. Contagi in ospedale nella struttura, che occupa i locali dell'ex Istituto salesiano, sono stati riscontrati 43 casi di Covid-19, dopo che sono stati effettuati tamponi a tappeto. Gli ospiti contagiati sono 33 e sono 10 coloro che, tra addetti e personale, hanno contratto l'infezione. "Situazione costantemente monitorata dalla Asl e da noi", dice il sindaco Leo Castiglione. Sette degli anziani positivi, quelli che presentano sintomi, sono stati trasferiti in ospedale a Chieti. Attenzione anche sull'ospedale "Bernabeo", dove il reparto di Lungodegenza si è trasformato in un focolaio, con 10 pazienti positivi. Erano otto ma nelle ultime ore i casi sono aumentati. Nel reparto al momento stop a ricoveri e a dimissioni. Nel Centro di procreazione medicalmente assistita, invece, sono cinque gli operatori sanitari che hanno preso il virus.*"

The article led to the system predicting Hospital Outbreaks as the primary class, significantly influenced by the activation of the prompt "*This news is about the hospital that takes care of hospitalizing COVID cases.*" which received a high confidence score of 0.96. This prediction underscores the focus on hospital-related aspects of the outbreak, particularly the transfer of symptomatic elderly patients to a hospital and the emergence of a cluster within the hospital's long-term care department. As is evident, for instance, in the sentence "*Seven of the elderly individuals who tested positive, those exhibiting symptoms, have been transferred to the hospital in Chieti*". However, the original classification was NURSING HOMES OUTBREAKS, which also finds strong representation through the activation of the prompt "*This news is about covid in nursing homes.*" with a confidence score of 0.89. This classification captures the article's primary focus on a COVID-19 outbreak in a nursing home, including the infection of residents and staff, which constitutes the core event described.

Finally, let us consider:

> "*Ladispoli, coronavirus nuovo focolaio alla Rsa Gonzaga. Parte il primo drive in della zona Finora 13 positivi nella struttura, tra degenti e operatori. Altro focolaio dopo una festa tra bambini: una mamma aveva il virus. Tre operai contagiati anche all'Fca di Cassino.*"

In this instance, the system classified the news article under FAMILY OUTBREAKS, predominantly due to the prompt "*This news is about a family member who was infected in an outbreak.*" being highly activated with a confidence score of 0.83. This classification highlights the mention of a family-related outbreak following a children's party within the article, which could explain the system's inclination towards the FAMILY OUTBREAKS class.

However, the intended classification was NURSING HOMES OUTBREAKS, which is significantly less represented in the system's evaluation, demonstrated by the most activated prompt, "*This news is about caring for guests in nursing homes for the elderly.*" receiving a lower confidence score of 0.34. This outcome indicates that while the article does mention a new outbreak at a nursing home ("*Rsa Gonzaga*") and provides a count of infected individuals, the mention of a family-related incident might have skewed the system's prioritization towards FAMILY OUTBREAKS.