ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 9, Number 2
december 2023

aA Accademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255000945

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_9_2

IJCoL

Volume 9, Number 2
december 2023

CONTENTS

Intelligent Natural Language Processing for Epidemic Intelligence

Danilo Croce, Federico Borazio, Giorgio Gambosi, Roberto Basili*
Università di Roma, Tor Vergata

Daniele Margiotta, Antonio Scaiella**
Reveal s.r.l.

Martina Del Manso, Daniele Petrone, Andrea Cannone, Alberto Mateo Urdiales, Chiara Sacco, Patrizio Pezzotti, Flavia Riccardo†
Istituto Superiore di Sanità

Daniele Mipatrini, Federica Ferraro, Sobha Pilati‡
Ministry of Health

Epidemic Intelligence activities depend significantly on analysts' ability to locate and aggregate heterogeneous and complex information promptly. The level of novelty of the targeted information is a challenge. The earlier events of interest are located the larger the benefit: more accurate and timely warnings can be made available by the analysts. In this work, the role of Natural Language Processing technologies is investigated. In particular, transformer-based encoding of Web documents (such as newspaper articles as well as epidemic bulletins) for the automatic recognition of events and relevant epidemic information is adopted and evaluated. The resulting framework is configured as a domain-specific meta-search methodology and as a possible basis for a novel generation of Web search environments supporting the Epidemic Intelligence analyst.

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).

This concept has evolved over time broadening to Public Health Intelligence defined by the World Health Organization (WHO as *"...a core public health function re-*

* Dept. of Enterprise Engineering - Via del Politecnico 1, 00133 Rome, Italy.
E-mail: {croce,gambosi,basili}@info.uniroma2.it E-mail: borazio@ing.uniroma2.it
** Reveal s.r.l., Via Kenia, 00142 Rome, Italy. E-mail: {margiotta,scaiella}@revealsrl.it
† Infectious diseases department - Istituto Superiore della Sanità - Viale Regina Elena, 299 - 00161 Roma.
E-mail: {martina.delmanso,daniele.petrone,andrea.cannone,alberto.mateourdiales,
chiara.sacco,patrizio.pezzotti,flavia.riccardo}@iss.it
‡ General Directorate for health prevention - Ministry of Health - Viale Giorgio Ribotta, 5 - 00144 Roma.
E-mail: {d.mipatrini,fe.ferraro,s.pilati}@sanita.it
1 International Health Regulations (2005):
https://iris.who.int/bitstream/handle/10665/43883/9789241580410_eng.pdf

sponsible for identifying, collecting, connecting, synthesizing, analyzing, assessing, interpreting and generating a wide range of information for actionable insights and disseminating these for informed and effective decision-making to protect and improve the health of the population."[2].

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 assess existing capacities and areas where additional implementation would be needed.

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. sourced from official websites or 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[3]).

Following several pilots to design and test this national event-based surveillance component of epidemic intelligence, Italy chose to follow the implementation model contextually developed and sustainably 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. Each nominated analyst, before being included in the network, is required to accomplish theoretical and practical training coordinated by the ISS and the Italian Ministry of Health. Since 2023, the Italian Network of Epidemic Intelligence has been part of the community of practice of Epidemic Intelligence from Open Source - EIOS - an initiative of the WHO.

To date, the Italian Network of Epidemic Intelligence comprises 48 analysts across the country operating under a formal surveillance framework established by the Min-

2 Annual global report on public health intelligence (2021): `https://cdn.who.int/media/docs/default-source/documents/emergencies/phi_report_2021_production_final_web.pdf`

3 ECDC: `https://www.ecdc.europa.eu/en/news-events/e-learning-course-epidemic-intelligence-ei`

istry of Health[4] and with a formal role in threat detection and risk assessment in the national pandemic preparedness plan[5].

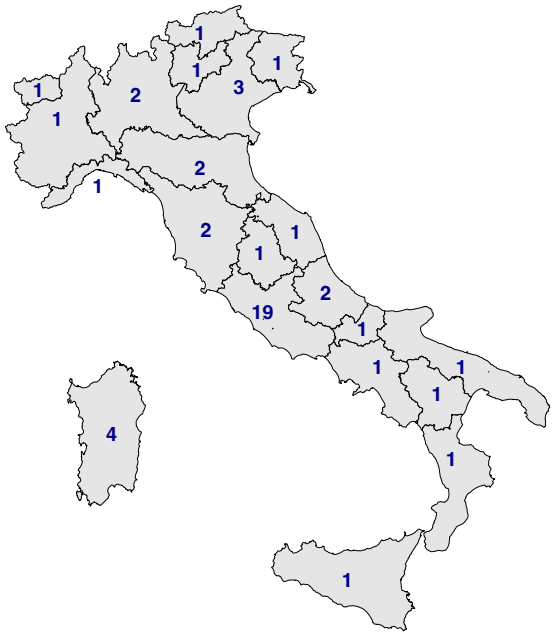

Figure 1
Regional distribution of analysts

The Italian event-based surveillance organizational model was activated for the first time on a continuous basis in 2015 to support surveillance and early warning of the EXPO mass gathering[6]. On this occasion, it was evaluated and found to be sustainable and effective.

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.

The event-based surveillance system operates on three activation levels:

- *Level 0*: situation awareness - signals captured by international and possible national monitoring are shared in a restricted group in real-time

4 Istituzione della sorveglianza basata su eventi in Italia: `https://www.quotidianosanita.it/allegati/allegato7643832.pdf`
5 PanFlu (2021-2023): `https://www.salute.gov.it/imgs/C_17_pubblicazioni_3005_allegato.pdf`
6 EXPO (2015): `https://www.politicheagricole.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/8682`

through instant messaging. The focus is national (on cases or clusters of any infectious disease in Italy) and on any international signals of interest for Italy. This activity is conducted in collaboration with other epidemic intelligence networks, such as ECDC and GHSAG-EAR.

- *Level 1*: activated once an event occurring internationally has been detected. The analyst performs the activities related to level 0 and in addition follows the identified event in time. Signals related to this event are shared in a restricted team in real-time through instant messaging.
- *Level 2*: activated once an event occurring or with the potential to occur in Italy is identified. In addition to activities performed under level 0, and if active level 1, analysts conduct an in depth screening and assessment of terms related to the identified level 2 event and publish an ad hoc thematic bulleting that, depending on the situation, can be daily or weekly. Instant messaging can be also applied if relevant.

While Level 0 is by default always active, Levels 1 and 2 are subject to activation and subsequent de-activation based on an assessment made by the ISS and the Italian Ministry of Health. Based on the workflow described, 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 application of Natural Language Processing techniques is beneficial in this scenario to optimize information analysis and monitoring. The main goal is to enhance effectiveness in the following ways:

- *Increased research capabilities*: Thanks to its ability to process large amounts of data, the system can identify hidden infectious diseases and potential biological threats. This helps analysts discover relevant information that might otherwise go unnoticed.
- *Reduced monitoring time*: Artificial intelligence can automate some monitoring tasks, freeing analysts from repetitive duties or saving time. This allows them to focus on more complex and strategic tasks as the assessment.
- *Improved analysis*: The system can help analysts make more informed decisions in the evaluation process and quickly spot any anomalies or threats.

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).

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 on the topic received a boost with the outbreak of the COVID-19 pandemic (Al-Garadi, Yang, and Sarker 2022), (Raza, Schwartz, and Rosella 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. 2021), (Choudhary, Alugubelly, and Bhargava 2023) as well as document classification (Li et al. 2022), (Lin et al. 2021), (Kaliyar, Goswami, and Narang 2021).

The use of transformer architectures as a basis for Large Language Models, to news classification is an active research area (see for example (Khosa, Mehmood, and Rizwan 2023), (Lin et al. 2021), (Santana, Oliveira, and Nascimento 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, Kuner, and Jaeger 2021) and it is still in its infancy.

This study investigates the impact of automatic event classification in Epidemic Intelligence by leveraging the advanced capabilities of Transformer-based models, specifically exploring the application of these models for automatic news classification relevant to epidemic monitoring. The research is motivated by the need for efficient and accurate tools to navigate the vast amount of information available, aiming to improve the identification of pertinent epidemic signals. This is accomplished by employing robust Transformer models, such as BERT-based models (Devlin et al. 2019), to classify articles from the ISS surveillance of infectious disease events in Italy.

The methodology employed in this study is grounded in the practical application of the UmBERTo model (Parisi, Francia, and Magnani 2020), a Transformer-based classifier tailored to the Italian language, to analyze and classify articles. The experiment is designed to provide a detailed examination of the model's performance in classifying the content according to predetermined themes related to epidemic intelligence. The results of this classification process are meticulously documented, providing insights into the effectiveness of Transformer-based models in a real-world application scenario where comprehensive coverage and accuracy are paramount.

In addition to the classification task, this study also investigates the integration of the classified data into a sophisticated semantic search system. This system is designed to go beyond simple information retrieval, offering dynamic search capabilities that consider various metadata attributes, including date and region. The introduction of such a system represents a significant enhancement in the field of epidemic intelligence, offering a more efficient and focused mechanism for accessing and analyzing news related to epidemics.

Overall, this research highlights the potential of Transformer-based models to support the field of epidemic intelligence by providing tools that significantly improve the speed and accuracy of event classification. This advancement is crucial for public health officials and analysts who rely on timely and accurate data to make informed decisions in response to epidemic threats. The study's findings underscore the applicability of machine learning algorithms in enhancing the capabilities of epidemic intelligence systems, thereby contributing to the broader effort to safeguard public health.

In the rest of the paper, we explore neural AI in language processing (2), demonstrate automating event classification for Epidemic Intelligence (3), and detail integrating this into a semantic search system for epidemiological analysis (4).

2. Language Processing with Neural AI models

Machine Learning approaches to Language Understanding and Text Processing tasks date back to the 80's (Lebowitz 1988), and have been revived during the 90's by the research activities on Statistical Natural Language Processing (Jelinek 1998; Manning

and Schütze 2001). The rise of neural approaches to NLP, although very old in its inspiration (Ide and Véronis 1990; Yuan et al. 2016), has inspired important contributions to distributional models of the lexicon (e.g., (Bengio, Ducharme, and Vincent 2000), (Mikolov et al. 2013)) lately applied to sentence (Le and Mikolov 2014) and text encoding tasks (Devlin et al. 2019; Reimers and Gurevych 2019). In this work, we emphasize the adoption of advanced Text Encoding techniques as a basic mechanism to capture event information which is relevant for Epidemic Intelligence.

Transfer learning, i.e. pre-training a neural network on data available for T1, and then using the model M_1 induced for T1 as the basis for a fine-tuning stage on the data available for the task T2, has been widely applied in the recent years, and has been shown beneficial in many contexts, among others computer vision. For example, the pre-training of convolutional neural networks on the ImageNet dataset is commonly used to support the later fine-tuning stage of the resulting pre-trained network, in order to obtain a new optimized task-specific model, e.g., (Girshick et al. 2013).

The approach proposed in (Devlin et al. 2019), namely *Bidirectional Encoder Representations from Transformers* (BERT) embodies exactly this approach, applied in NLP. It provides a very effective model to pre-train a deep and complex neural network over very large corpora of unannotated text. This makes the resulting network suitable for application to a large variety of NLP tasks: it can be simply extended to new tasks by fine-tuning the entire architecture with (possibly small) problem-specific datasets.

BERT exploits the *Transformer* architecture, an attention-based mechanism able to learn contextual relations between words (or sub-words, i.e. word pieces, (Schuster and Nakajima 2012)) in a text. In its original form, proposed in (Vaswani et al. 2017), a transformer includes two separate components, an *encoder* that reads the text input and a *decoder* that produces a prediction for a targeted rewriting (e.g., Machine Translation) tasks. Notice that the BERT input is handled by incorporating positional embeddings: individual symbols in the input sequence, such as words or wordpieces, are coupled with their specific positional information, through a dedicated embedding based on sinusoidal functionals, designed to represent the position of each word within a sentence. Notice that positional embeddings aim at inducing and expressing the contextual and syntactic information inherently carried out by the word order in a sentence. By encoding the position of each word, BERT can be made sensitive to the role and relationship of words within the entire sentence.

As BERT may depend on more than 110 million parameters, the key to its success lies in the concept of *pre-training*. Network's weights are initialized through several tasks, called pre-training tasks, that, while potentially unrelated to the target network's primary task, are linguistic in nature and help the model generalize its understanding of language use. According to Wittgenstein (Wittgenstein 1953), language meaning arises as a side-effect of its use by native speakers. Language use is thus the crucial source of information about syntactic and lexical semantics phenomena in natural language. Pre-training in transformers aims at capturing exactly such universal properties of natural languages *before* attempting the training aimed at specific linguistic inferences (e.g., machine translation or question answering).

These tasks are thus carried out by applying the network to extensive document collections, often consisting of billions of tokens, supporting a large-scale exposure to diverse linguistic phenomena. BERT is thus allowed to develop a nuanced understanding of natural language, akin to linguistic acquisition among humans.

As reported in Figure 2 (on the left), *during pre-training* the Transformer encoder takes in input entire sequences of words at once, in parallel. It acquires a language model by learning to reconstruct an original sentence from its corrupted version: it

randomly masks some of the tokens from the input, and the objective is to predict every original masked word based only on its context. The model derived through such pre-training objective is called MLM (*Masked Language Model*). In addition to the masked language modeling task, BERT also uses a second task, i.e. *next sentence prediction*, that jointly pre-trains over sentence-pair representations. This last objective is crucial to improve the network capability of modeling relational information between text pairs, which is particularly important in tasks such as QA (Devlin et al. 2019) in order to relate an answer to a question.

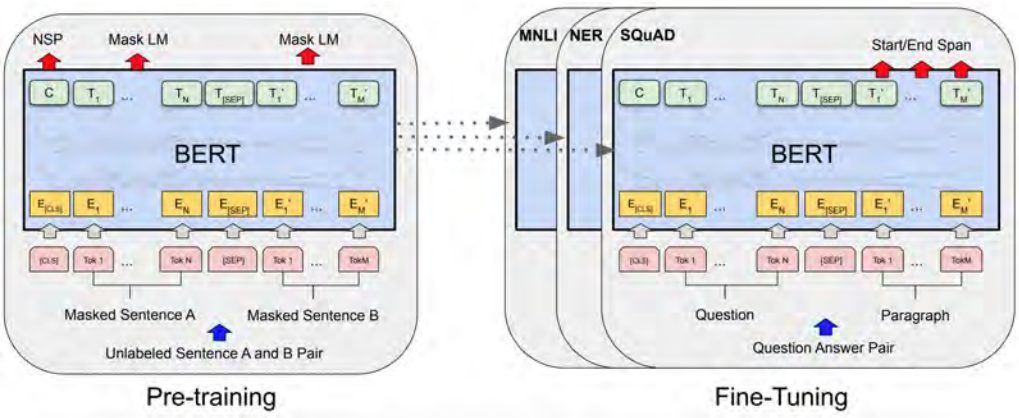

Figure 2
Pre-training and fine-tuning procedures for BERT. In the architecture, all the layers, except the output ones, are used (e.g., optimized) in both pre-training and fine-tuning. The pre-trained model parameters are used to initialize the different models for independent downstream tasks (e.g., Question Answering over SQuAD). [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token between sentences (e.g., separating questions/answers).

After the language model has been trained over a generic document collection, the BERT architecture allows encoding (i) specific words belonging to a sentence, (ii) the entire sentence, and (iii) sentence pairs with dedicated embeddings. These can be used in input to further deep architectures to solve sentence classification, sequence labeling, or relational learning tasks by simply adding simple layers and fine-tuning the entire architecture (Bouraoui, Camacho-Collados, and Schockaert 2020). On top of such embeddings, *fine-tuning* is applied by adding task-specific and simple layers on top of the architecture acquiring the language model. In a nutshell, this layer introduces a minimal number of task-specific parameters and is trained on the targeted tasks by simply fine-tuning all pre-trained parameters, optimizing the performance of the specific problem.

The adoption of BERT encodings in Text Classification. BERT can be considered a sentence-enconding model that takes an input sequence s and generates a vector representation $h = M_{BERT}(s)$ for s. In the context of this work, our primary interest is in classification tasks. For a given input sequence s, BERT can be viewed as generating a vector from the first symbol of the sequence. In BERT's architecture, this first symbol is the artificial token [CLS], and h is a dense vector of d dimensions (e.g., d =768).

BERT can be employed for classification tasks, as its output h, a 768-dimensional vector, is further processed through a neural classification network, able to map this

high-dimensional vector to a space of c dimensions, where c is the number of targeted classes. The output of this classifier, y, is a one-hot vector representing the class probabilities. Mathematically:

$$y = M_{Class} \circ M_{BERT}(s)$$

where $M_{BERT}(s)$ is the BERT model applied to the input sequence s, and M_{Class} is a dedicated multilayer perceptron with a given number of hidden layers (usually one) and one output layer with c neurons. The domain of the last weight matrix is thus $\mathbb{R}^{(d+1)\times c}$, with $d = 768$ being the dimension of the BERT output. The classifier model M_{Class} is trained through backpropagation, usually by referring to the *cross-entropy loss* function

$$\text{Loss} = -\sum_{i=1}^{c} y_i \log(p_i)$$

Here, y_i is the true class label in one-hot encoded form, and p_i is the predicted probability of each class. The minimization of this loss function on training data optimizes the model parameters, ensuring the accurate classification of input sequences. The straightforward application of BERT has shown better results than previous state-of-the-art models on a wide spectrum of natural language processing tasks (Lin et al. 2022).

3. Automating Event Classification in Epidemic Intelligence: from keywords based retrieval to Automatic Topic Recognition.

Epidemic Intelligence (EI) is a wide process that involves information spread across several sources ranging from governmental health information systems, Clinical Data at the National and International level as well as Web and Media sources such as newspaper articles, technical reports, and social media posts. The involved profiles are analysts who are focused on the collection of evidence about epidemic phenomena as documented by eventualities (i.e. facts, events) and the geographical locations where they occur. Notice that the quality and the timeliness of the collected information are very important for the analysts as they are faced with impressive data volumes to be analyzed, (mostly) filtered out, and then summarized in health-related reports and infographics. The ability to locate as timely as possible the information of interest across the different and heterogeneous sources is crucial for supporting and augmenting the capability of the EI analysts.

Usually, Web search is central to the above process, as the gathering of the suitable but unknown sources should be able to "*scan*" at least any publically accessible distributed information source. In this scenario, the Information Retrieval techniques currently employed are based on the Web metasearch engines fed with traditional keyword-based searches. Keywords allow on the one-side the analysts to directly specify the target phenomena through lexical expressions (names or phrases recalling events, virus names, or biological expressions related to potential epidemic phenomena) and guarantee a certain level of precision in the resulting retrieved materials. On the other side, keywords are usually adopted with similar semantics across different engines and sources, in order to make documents derived from different sources and

engines semantically coherent and comparable with respect to the target epidemic phenomena.

It is clear in the above scenario that keyword-based searches cannot always be a guarantee for consistent and precise information gathering so they may increase the risk of gathering large volumes of irrelevant information. This is mainly due to ambiguity phenomena affecting individual keywords. For example, clinical studies may well refer to uses of the word *flood* that are health-related, as in the case of epidemics, e.g.,

... flood of novel clinical cases are reported across the entire area ...

However, when other meanings are adopted as in

Forests provide protection against floods and drought.
After you recognize it, real remorse floods your soul..

Notice that every time a wrong use of a keyword is captured by a search engine the corresponding documents have a very high probability of resulting irrelevant to the experts analysts activity. This burden becomes prohibitive if an improper usage of a keyword is made by analysts who are surely non-expert in Web indexing and searching. They are thus not keen to adopt complex languages to express keyword logical compositions. Moreover, leaving the burden of correct, consistent, and efficient usage of keywords to analysts is a strong limitation of the overall impact of Web search. It makes the result semantically limited and not scalable so that the ability to improve the coverage (recall) of the targeted epidemic phenomena is severely impacted. In fact, the variety of phenomena of interest for analysts is left to its choice of keywords, a step that limits the semantic scope of the entire Web search process: *information* obtained by the analysts through the retrieved texts is fully enclosed by the phenomena that he contributed to define. No generalization or semantic inference about the targeted topics is made possible and the resulting *information novelty* is poor. With this impoverished level of technology, any proactive support to the analyst is minimally realized. The consequence is an evident reduction in the potentially reachable *recall*, i.e. a limitation of the involved Web search technology, as a concrete support in augmenting the scope of the analyst search as well as the novelty of the gathered Web information.

The above limitations strongly depend on the fact that keyword-based searches, usually close to the bag-of-word approaches, directly represent text semantics just through word sets and their descriptive statistics. No generalization nor linguistic interpretation of the retrieved texts is attempted either during indexing or during retrieval. The entire literature on lexical semantics and distributional methods of induction of word and text semantics from corpora is neglected by keyword-based approaches. Although different semantic approaches to Web search have been proposed and discussed since the 90's (Guha, McCool, and Miller 2003; Madhu, Govardhan, and Rajinikanth 2011), the Epidemic Intelligence community is not yet able to benefit from any advantage from them. In this work, we will pursue the idea that text interpretation is useful in EI for several dimensions

- Allows the analysts to expand the keywords with semantically related terms in the health domain, useful to augment the overall IR recall and to better constraint the interpretation (and ranking) by the search engine
- Use semantic interpretation of keywords and texts to filter out the truly irrelevant retrieved documents and automatically classify the incoming interesting data

- Support the analysts within an overall knowledge framework able to understand document content and helpful in better formulating target topics, in composing queries about specific phenomena as well as in supporting conceptual rather than textual navigation across the emerging document networks where associations are justified on semantic grounds

In the experimental section of our paper, we describe the methodology employed for classifying news articles using advanced natural language processing techniques. Data for this study was sourced from the ISS Epidemic Intelligence - event-based surveillance (EBS) activity. During the second phase of the pandemic, an ad hoc EBS was implemented by ISS to support the monitoring of the risk of spread during the COVID-19 pandemic, under the Ministerial Decree of 30 April 2020 (World Health Organization 2008) and (Riccardo et al. 2014). 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 3,245 news articles. Of these, 25% were sourced from national newspapers, reflecting the broader public health discourse at a country level. The remaining 75% were derived from local news outlets, providing insights into the pandemic's impact on specific communities and regions.

The annotation process was carried out by a team of four analysts at the ISS, comprising two medical doctors and two prevention technicians, all domain experts with specific knowledge on the spread of viral diseases such as COVID-19. These individuals meticulously categorized each article, ensuring that the classification reflected the most significant outbreak settings observed during the pandemic. The focus was particularly on outbreaks involving vulnerable groups such as hospital patients and the elderly, as well as those highlighting high rates of social interactions, including events involving children who, during the pandemic, were identified as having a high number of contacts and, by extension, their families. This detailed categorization aimed to sharpen the analysis on specific outbreak scenarios critical for public health surveillance and response strategies.

These categories thus reflect the primary theme of each news piece and include: "SCHOOL OUTBREAKS", FAMILY/FRIEND OUTBREAKS", "NURSING HOMES/LONG-TERM CARE FACILITIES", "HOSPITAL OUTBREAKS", "OUTBREAKS IN OTHER SETTINGS" and "OTHER VARIANTS". Notably, the OUTBREAKS IN OTHER SETTINGS category corresponds to a miscellaneous class, specifically useful in an open-world and exploratory scenario. It corresponds to a less semantically characterized area, whereas different topics (e.g., regional issues, environmental health-related phenomena as well as politics) are mixed. The distribution of examples is reported in Figure 3.

For the task of classification, we employed BERT (Devlin et al. 2019), by focusing on the very first excerpt of each text: news articles exceeding 512 wordpieces, the maximum size manageable by BERT, are thus truncated. We specifically adopted UmBERTo (Understandable BERT for Italian), a RoBERTa-based language model, distinctively pretrained on large Italian corpora (Parisi, Francia, and Magnani 2020). We employed UmBERTo-Commoncrawl-Cased[7], which leverages the Italian subcorpus of OSCAR as its training set: this comprises approximately 70 GB of plain text. Our implementation was carried out in Python, relying on the model versions available on Huggingface.

7 https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1

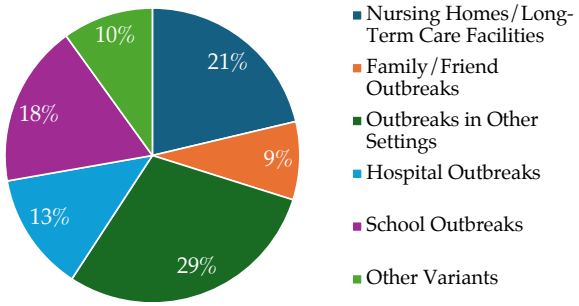

Figure 3
Class distribution

In the model fine-tuning, a dropout rate of 0.1 was applied to the last layer. We adopted a 5-fold classification schema for our experiments. In each fold, 8 were designated for training data, 1 for development of parameters, and 1 for testing. After repeating these experiments 5 times, results were evaluated in terms of average accuracy, which is the percentage of examples correctly reassigned to their respective categories. To gauge the model's performance across different categories, metrics such as Precision, Recall, and F1 Score were also calculated for each class.

Experimental Results. The results are presented in Table 1. The first column lists the different categories, for each of which Precision, Recall, and F1 scores are reported. The last row provides the arithmetic mean across all classes. Overall, the results are quite promising. A reference baseline model that assigns the most frequent class to any incoming document would just achieve an accuracy of 29.3%. The overall accuracy reached by our model is 86%, which confirms its robustness across all classes. The average F1 score obtained is 85%. Some classes are particularly well categorized (e.g., NURSING HOMES/LONG-TERM CARE FACILITIES or SCHOOL OUTBREAKS) with F1-scores over 0.90. The category with the lowest results is "FAMILY/FRIEND OUTBREAKS", with an F1 score of 0.63, which is also the least represented, accounting for less than 8% of the data (278 out of 3245 documents).

Table 1
Performance evaluations of the automatic categorization task in terms of Precision, Recall, and F1 measures: the overall average Accuracy is 86%.

Category	Precision	Recall	F1-Score
NURSING HOMES/LONG-TERM CARE FACILITIES	0.88	0.93	0.91
FAMILY/FRIEND OUTBREAKS	0.59	0.68	0.63
OUTBREAKS IN OTHER SETTINGS	0.81	0.84	0.83
HOSPITAL OUTBREAKS	0.96	0.80	0.87
SCHOOL OUTBREAKS	0.91	0.91	0.91
OTHER VARIANTS	0.96	0.89	0.92
Global Micro Average	*0.85*	*0.84*	*0.85*

To evaluate our Transformer-based model's robustness, we compared its performance against established baseline models commonly used in topic classification tasks. We selected traditional machine learning approaches for this purpose: Logistic Regression (Byrd et al. 1995), Gradient Boosting (Hastie, Tibshirani, and Friedman 2009), and Multinomial Naive Bayes (Manning, Raghavan, and Schütze 2008)[8]. The comparative results are presented in Table 2, detailing the Accuracy, Precision, Recall, and F1 scores for each model. Our Transformer-based model outperforms these baselines significantly, exhibiting at least an 18% increase in accuracy. Moreover, its F1 score exceeds that of the baselines by a minimum of 8 percentage points, showcasing its superior balance between precision and recall. This equilibrium is crucial for efficiently capturing true positives while minimizing false positives and negatives. It is essential to note that the baseline models, based on Bag-of-Words approaches, lack the contextual understanding inherent to BERT. While Logistic Regression and Gradient Boosting demonstrate commendable precision, they fall short in recall. Conversely, Naive Bayes shows high recall but lower precision. Our BERT-based system, benefiting from linguistic generalization acquired during pre-training and a sophisticated contextual representation mechanism, adeptly balances both aspects, achieving an F1 score of 85%. This performance underscores the transformative potential of Transformer models in handling complex classification tasks.

Table 2
Comparative Performance Evaluation of Different Models on the Automatic Categorization Task. Accuracy, Precision, Recall, and F1 Scores are reported for the baselines and the Transformer-based model, more comprehensively presented in Table 1.

Model	Accuracy	Precision	Recall	F1-Score
LOGISTIC REGRESSION	0.68	0.84	0.71	0.77
GRADIENT BOOSTING CLASSIFIER	0.63	0.86	0.68	0.76
MULTINOMIAL NAIVE BAYES	0.66	0.71	0.82	0.76
UMBERTO (this paper)	**0.86**	**0.85**	**0.84**	**0.85**

Notice that the nature of the dataset is representative of different complex phenomena, such as news discussing multiple topics or ambiguous texts referring to some borderline events not related to just one class. As, in the target reports, the analysts are asked to decide a unique category for each news item, subjective choices representative of just partial views on some news are well possible. Unfortunately, as no inter-annotator agreement had been measured during the development of the corpus, it has not been possible to have a quantitative estimation of these phenomena.

Error Analysis. The confusion matrix can be helpful in this sense, and Table 3 displays the results for one fold, whereas system outputs are compared in columns against the annotated (gold) classes reported in rows. From the confusion matrix, it is evident that the most overlapping classes are "OTHER SETTINGS" and "OTHER VARIANTS", as they represent broader, collective, and mixed categories. Error analysis of specific news is important to better understand the nuances of misclassifications and gain insights into the model's performance in search of potential areas for improvement. Let's consider the following detailed example:

8 Implementations for these methods were obtained from the *scikit-learn.org* library.

Table 3
Confusion Matrix of the Classification outcomes of the UmBERTO model: rows represent the gold classes while columns report the system's assignments.

Confusion Matrix	NURSING HOMES	FAMILY FRIEND	OTHER SETT.	HOSP.	SCHOOL	OTHER VARIANTS
NURSING HOMES	**183**	0	11	0	2	0
FAMILY/FRIEND	1	**51**	19	0	4	0
OTHER SETTINGS	7	25	**232**	4	6	2
HOSPITAL	16	3	7	**111**	0	2
SCHOOL	0	5	9	1	**160**	0
OTHER VARIANTS	0	3	7	0	3	**100**

"*Cluster in una scuola calcio a Tor Bella Monaca, ora chiusa dalla Asl Roma 2. I bambini positivi, divisi in due squadre, sono allo stato attuale quattro, ma in quarantena ci sono decine di contatti nelle elementari limitrofe. All'incirca dieci classi . . .*[9]".

The gold (manual) annotation of the above news is "OUTBREAKS IN OTHER SETTINGS" while the system prediction corresponds to "SCHOOL OUTBREAKS".

In this example, the system's misclassification can be attributed to the various expressions in the news that typically refer to the school environments. The event is about a *soccer school* but it also makes a significant reference to a school outbreak. The reference to a "*scuola calcio*" (*soccer school*) and the mention of "*bambini*" (*children*) and "*classi*" (classes) in the context of the quarantine are likely misleading information for the model pushing for preferring a "SCHOOL OUTBREAK" class. notice how the actual context of the news focuses on the sports activity (*soccer schoo*l) and this is what matters to the expert as the main aspect of this news item. It is thus correctly annotated as "OUTBREAKS IN OTHER SETTINGS".

Another instance of misclassification offers further insights into the challenges faced by the model. Let's delve into this specific example:

"*Sempre al Santa Maria del capoluogo ci sono quattro ricoverati di nazionalitá cinese: nella comunitá perugina si è infatti acceso un focolaio importante che é sotto stretto controllo dei sanitari.*[10]"

Here the gold annotation is "OUTBREAKS IN OTHER SETTINGS" while the system predicts "HOSPITAL OUTBREAKS". In this case, the misclassification appears to stem from the discussion about the typical hospital setting. The mention of "*ricoverati*" (*hospitalized individuals*) and the name "*Santa Maria*" (in fact, a hospital), along with an outbreak under medical supervision, have led the system to categorize this instance incorrectly as a "HOSPITAL OUTBREAK". However, the focus of the analyst seems to be the broader context of the community-wide outbreak in the *Perugian community*. The gold classification here, i.e. "OUTBREAKS IN OTHER SETTINGS", is a typical example of a subjective choice by the analysts. It could be motivated by the fact that other news were used to

9 In English: *Cluster in a soccer school at Tor Bella Monaca, now closed by Asl Roma 2. The positive children are divided into two teams, currently including four pupils, but there are dozens of contacts in quarantine in the neighboring elementary schools. Approximately ten classes . . .*

10 In English: *At the Santa Maria in the capital, there are four hospitalized individuals of Chinese nationality: a significant outbreak has indeed flared up in the Perugian community, which is under close medical supervision.*

account for the *Santa Matria* event, or the emphasis on the geographical contribution of this news, that makes explicit reference to a specific area.

Another notable mistaken instance sheds some light on the impact of the context on the text understanding implicit in our text classification process. Consider the following case:

"*Cava de' Tirreni. Sono saliti a 33 i casi di contagio da coronavirus nel reparto femminile di una casa di cura privata di Cava de' Tirreni. La situazione, dunque, si complica e diventa molto piú allarmante rispetto alle scorse settimane. Lo riporta La Cittá di Salerno. I primi casi di positivitá tra le ospiti del centro di riabilitazione . . .*[11]"

Here the gold annotation is "HOSPITAL OUTBREAKS" and the prediction is "NURSING HOMES/LONG-TERM CARE FACILITIES". In this scenario, the misclassification can be attributed to the description of the setting as a "*private care home*" and a "*rehabilitation center*". These terms typically resonate with the environment of nursing homes or long-term care facilities, leading the model to such a prediction. While, the context of the news, specifically the mention of a "*female ward*" and the nature of the outbreak, also aligns with the category of "HOSPITAL OUTBREAKS", the news content positions at the borderline of both categories. This emphasizes the challenge of distinguishing between closely related healthcare settings: the availability of clear distinctions, that are shared among annotators/analysts, and representative sets of examples is here crucial for the underlying NLP modeling.

The above different classifications illustrate the challenges for the categorization task, especially the limitations of keyword-based approaches. Terms in fact may overlap across categories as well as exhibit senses that are related to the news context in complex manners. This experience highlights the importance of an accurate contextual understanding of incoming news, thus outlining the need for high-quality training with clear definitions for the target classes. Notice that in Epidemic Intelligence classes may arise in an unexpected fashion and the development of large and accurate training sets is often not viable. This emphasizes the role of a collaborative framework where the system's proactive data gathering and filtering abilities can be made available to the analyst, in order to alleviate the uninteresting part of its effort. This is why an overall integrated Natural Language Processing and Information Retrieval approach, as described in the next section, can be considered a relevant support to trigger proper Machine Learning functionalities for the analyst, through a fully engineered collaborative environment.

4. Integrating News Classification into a Semantic Search System for Enhanced Epidemiological Analysis

The direct application of the accurate news classification into Epidemic Intelligence (EI) relevant thematic classes concerning COVID, as described in the previous section, allows to devise of a fully integrated Semantic Search system. The system is designed to support EI analysts in thematic event analysis as they can be observed in online documents and news streams, that reflect real-world phenomena.

The system under discussion aims to enable analysts to search for web news and then focus on themes related to specific epidemiological analyses. In essence, it acts as

11 In English: *Cava de' Tirreni. The number of coronavirus infection cases has risen to 33 in the female ward of a private care home in Cava de' Tirreni. The situation, therefore, becomes more complicated and much more alarming than in previous weeks. This is reported by La Cittá di Salerno. The first positive cases among the guests of the rehabilitation center . . .*

a filter for Web news, allowing analysts to concentrate on information pertinent to their research or monitoring needs.

Once the news articles are located, uploaded, and indexed, the system provides search functionalities. Analysts can use keyword-based or natural language queries to retrieve news articles. The thematic classification previously described plays a critical role in this process. It allows for the filtering of results based on the themes of interest to the analyst. For example, a query like "*Suspicious death*" filtered using the class "*Nursing Homes/Long-Term Care Facilities*" can enable analysts to delve deeply into this topic in a focused manner.

The thematic classification aims at refining the search results but also adds contextual information about news for interpreting its relevance to the query. When categorized news articles are retrieved by a query the system ensures that search results can be organized into classes, such as into classes like "*School Outbreaks*", "*Hospital Outbreaks*", or "*Other Settings*". Classes can be used as an enrichment of news and can be used as filters for targeting the specific information needs of epidemiological research.

News articles are also enriched with other editorial metadata, such as the publication date or the geographical region of origin. In this way, standard Business Intelligence (BI) filters and aggregations can be directly employed to cluster conceptually retrieved data, compute their statistics, and analyze trends. The thematic classification into EI classes of interest can be combined with these metadata elements. This enables dynamic insights and allows to examination of emerging patterns within a broad spectrum of web-based news. For example, a simple query towards a specific class (e.g., "*Hospital Outbreaks*") and filtered according to metadata matching a specific region can be used to focus on specific news articles whose statistics are direct evidence of potential warnings, e.g., regional increase in hospital-related COVID cases. This can directly prompt more detailed investigations, or immediate actions, based on other region-specific data. Additionally, tracking changes over time, made possible by temporal metadata, allows for trend analysis across time, and offers vital insights into the progression, or decline of specific events or alerts across different geographical areas.

The fully automatic data-driven approach suggested here, leveraging natural language understanding and thematic classification provides enriched metadata and transforms the way news is analyzed. It is a concrete way to straightforwardly promote informed decision-making and strategic planning, in public health and epidemiological research.

In Figure 4, the landing page of the Semantic Search system is showcased. This interface is designed to facilitate advanced search capabilities for users, particularly in the context of epidemiological analysis. Beyond the standard text field for natural language query input, the system provides additional options that allow users to refine their search based on various criteria related to publication time, geographical information, or EI topics related to news articles. The key features include:

- **Document Type:** This option enables users to select news based on information about their source and technology: here articles (such as news from Google) or tweets are distinguished. It allows filtering the information by its medium, by offering thus diverse perspectives or source perimeters.
- **Team:** This feature allows users to apply filters corresponding to specific team-defined criteria. For instance, users can filter data based on different

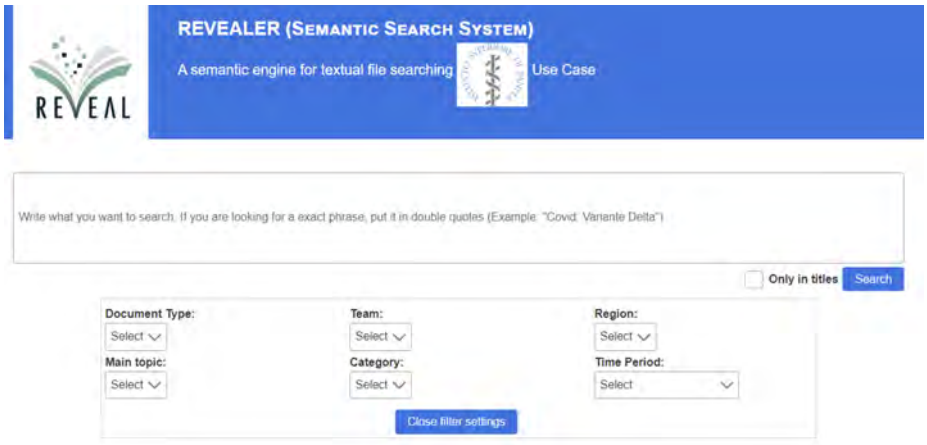

Figure 4
Landing page of the Semantic Search System, showcasing advanced search options. The interface includes fields for natural language queries, document type selection, team-specific filters, regional focus, main topic categorization, detailed category choices, and a time period selection tool, all designed to enhance the precision and relevance of epidemiological research and analysis.

analyst teams, as these are focused on independent aspects of epidemiological research or data curation.

- **Region:** With this feature, users can select a subset of news articles assigned during the download phase to one or more specific geographical districts or regions. This is particularly useful for localizing EI events, e.g., hospital outbreaks, to specific areas like Piedmont or Lombardy.
- **Main Topic:** Each analysis session addresses a specific set of EI events called "main topic", such as $COVID$, $MonkeyPox$, or $Dengue$. Articles downloaded during an event campaign are labeled with a specific "main topic" property: articles downloaded within different campaigns may carry multiple labels. When active, this feature allows for support requests related to specific epidemiological analyses (e.g., $COVID$ vs. $MonkeyPox$) by also enabling a cross-thematic analysis on all downloaded news.
- **Category:** This pertains to the categories assigned to each main topic, as discussed in the previous section. This functionality helps recognize events and separate them by categories. While $COVID$ related categories are defined on historical data, the collaborative work of analyst teams is currently underway to define more granular categories also for other topics, such as $MonkeyPox$.
- **Time Period:** This functionality allows users to specify a time frame for the returned news articles. Users can define this period using a calendar interface, thus enabling the analysis to focus on news from a particular period.

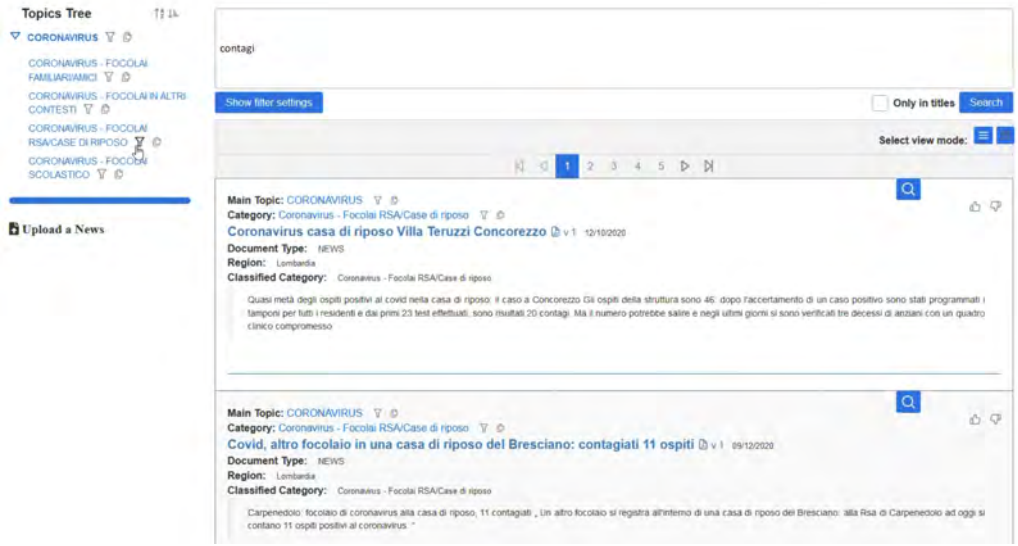

Figure 5
Results screen of the Semantic Search System following a query for "*contagi*" filtered by 2020 news articles from the Lombardy region under the main topic "*COVID*." The interface showcases categorized results and a synthetic display of the top news articles and their relevant metadata. User interaction functions support query expansion, re-ranking, and search for similar news. Users can also provide feedback on each article's relevance, in order to support retraining over time.

In Figure 5, we display the response to the query for "*contagi*" (*infections*) that also filters specifically news articles about the year 2020 and from the Lombardy region, under the main topic "*Coronavirus*". Unlike a traditional Web retrieval engine that might have returned tens of thousands of news items, the current system retrieves only 220 news articles, that precisely meet the analyst's needs.

Notice that, on the left side of the interface, the system displays the set of EI categories limited to those related to the retrieved news. These are ordered with respect to the number of retrieved news labeled accordingly: the largest is the news set of a category (e.g., HOSPITAL) the higher is its ranking on the left. This feature provides the analysts with a first informative topic map related to its query, that gives a semantic characterization of the query outcome. Selecting a specific thematic area, such as NURSING HOMES/LONG-TERM CARE FACILITIES[12], by a click on the left category tree, allows to further filter news accordingly, reducing the material to be read.

At the bottom of Figure 5, the top two relevant news articles are displayed, with their main metadata (such as the Lombardy region and year of publication), along with a news snippet preview.

The system employs Word Embedding (Mikolov et al. 2013) to generalize the content of the texts. The domain-specific lexical embeddings were acquired by analyzing all downloaded news articles, supplemented with pages from the Italian Wikipedia

12 NURSING HOMES/LONG-TERM CARE FACILITIES is *CORONAVIRUS - FOCOLAI RSA CASE DI RIPOSO* in Fig. 5

connected to the *Health Sciences* category[13]. Each news article is represented as a linear combination of the word embeddings for nouns and verbs in the news (title and first two paragraphs, based on the hypothesis that news generally focuses attention in the first portion of the text). This enables two functionalities:

1. *Query Expansion:* Every entered query term triggers the system selection of k words in the domain lexicon whose embeddings are most similar (i.e. those maximizing the cosine similarity with the vector obtained through the linear combination of already entered query terms).
2. *Re-ranking:* Once the news articles are retrieved using a SOLR-based search engine, which adopts a BM-25 model for ranking results, the system offers the possibility to reorder the news semantically more connected to the query, i.e., those whose representation in the word space maximizes the cosine similarity with the query representation.
3. *Similar News Search:* This function can be activated using the magnifying glass icon present for each news article. The vector representation of a news item can be used, similar to modern RAG systems, to retrieve news semantically most related to a query news item.

Currently, the system is operational under validation by the analysts: as shown in Figure 5, users can judge the quality of the news retrieved by the system (ThumbUp/ThumbDown symbols) to measure the search engine's quality and provide examples for the application of re-ranking methods (Liu 2009).

5. Conclusions

This study has explored the application of Natural Language Processing (NLP) and Machine Learning (ML) in the context of Epidemic Intelligence (EI), with a focus on developing a Semantic Search system that integrates advanced NLP techniques. The findings and developments presented aims at demonstrating the utility of these technologies in enhancing epidemiological analysis. The major outcomes of this work include the use of Transformer-based models for classifying web news into thematically relevant categories for EI, showing promising results. This approach suggests the effectiveness of language-specific models in handling complex linguistic data for health-related information processing. Moreover, integrating classified information into a Semantic Search system seems beneficial in streamlining the information retrieval process in EI. This system enables efficient thematic searches, focusing on epidemiological relevance and offering a more targeted approach to data analysis. Finally, the ability of the system to utilize thematic classification and metadata enriches the analysis process. By applying Business Intelligence techniques for data aggregation and trend analysis, the system aims at offering a nuanced understanding of epidemiological patterns and trends. In summary, this research contributes to the field of EI by showcasing the potential of intelligent NLP and ML applications. The development of the Semantic Search system highlights the significance of technology-driven approaches in public health and epidemiological research, particularly in terms of processing and analyzing large-scale, linguistically complex datasets.

13 https://it.wikipedia.org/wiki/Categoria:Scienze_della_salute

Our ongoing and future work involves a comprehensive approach to enhancing and expanding the Semantic Search system developed in this study, with a focus on its application in Epidemic Intelligence (EI). The immediate ongoing task is the qualitative and quantitative evaluation of the system. This involves in-depth user feedback analysis and usability studies, combined with a quantitative assessment of the system's accuracy and efficiency in information retrieval and classification. These evaluations are critical for identifying areas that require improvement and ensuring the system's efficacy and reliability for EI applications.

Looking toward future developments, a significant area of expansion is the adaptation of the system to cover a broader range of diseases beyond COVID-19. Other types of diseases, such as the dengue that is transmitted by mosquitoes, present specific and new challenges in terms of information needs and contexts. Adapting the system to these challenges will involve refining classification schemas and renewing training and search algorithms to make them seemingly effective to the unique aspects of further diseases. This is why, the exploration the application of unsupervised learning methods, particularly leveraging Large Language Models, for the classification of news. A promising direction in this regard is the implementation of zero-shot learning techniques, where the model is capable of categorizing news into relevant themes without being explicitly trained on those categories. This approach could significantly improve the adaptability and flexibility of the system, allowing it to respond effectively to the detection of threats and risks to public health detect threats and risks to public health. Another innovative direction for future research includes the unsupervised extraction of classes from news data. This method aims at automatically identifying new and emerging themes and categories in epidemiological news, thus enabling the system to stay updated with evolving health threats and trends.

Acknowledgments

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

Special thanks to: Stefania Giannitelli, Angela Ancona, Xanthi Andrianou, Federica Attanasi, Debora Ballarin, Elisa Bernini, Sandro Bonfigli, Bianca Borrini, Marco Cristofori, Silvia Dari, Elisa Di Maggio, Maria Paola Di Sebastiano, Mariapaola Farinelli , Pina Iannelli, Michele Labianca, Aurora Luciani, Dario Macchioni, Valentina Marras, Stefano Marro, Elena Mascia, Noemi Maria Mereu, Alessandro Miglietta, Nadia Olimpi, Daniele Paramatti, Ilaria Pati, Vincenzo Restivo, Alessandra Rossodivita, Michela Sabbatucci, Angelo Salzo, Sarah Samez, Monica Sane Schepisi, Francesca Sanità, Roberto Santoru, Simonetta Santus, Stefania Scaltriti, Irene Schenone, Marco Serale, Matteo Sponga, Francesco Vairo, Donatella Visentin, Francesca Zanella.

References

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

Al-Garadi, Mohammed Ali, 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).

Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, Denver, CO, USA. MIT Press.

Bouraoui, Zied, 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, New York, USA. AAAI Press.

Byrd, Richard H., 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.

Choudhary, Ambrish, Mamatha Alugubelly, and Rupal Bhargava. 2023. A comparative study on transformer-based news summarization. In *15th International Conference on Developments in eSystems Engineering, DeSE 2023, Baghdad & Anbar, Iraq, January 9-12, 2023*, pages 256–261. IEEE.

Del Manso, Martina, 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*.

Devlin, Jacob, 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, June. Association for Computational Linguistics.

Gillioz, Anthony, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the transformer-based models for NLP tasks. In Maria Ganzha, Leszek A. Maciaszek, and Marcin Paprzycki, editors, *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, Sofia, Bulgaria, September 6-9, 2020*, volume 21 of *Annals of Computer Science and Information Systems*, pages 179–183.

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

Guha, Ramanathan V., Rob McCool, and Eric Miller. 2003. Semantic search. In Gusztáv Hencsey, Bebo White, Yih-Farn Robin Chen, László Kovács, and Steve Lawrence, editors, *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, May 20-24, 2003*, pages 700–709. ACM.

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

Gupta, Anushka, Diksha Chugh, Anjum, and Rahul Katarya. 2021. Automated news summarization using transformers. *CoRR*, abs/2108.01064.

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

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

Intelligence, Network 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, 01.

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

Kaliyar, Rohit Kumar, 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.

Khan, Aurangzeb, 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.

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

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

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

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

Li, Qian, 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.
Lin, Deping, Hongjuan Wang, Mengyang Liu, and Pei Li. 2021. News text classification based on bidirectional encoder representation from transformers. In *2021 International Conference on Artificial Intelligence, Big Data and Algorithms, CAIBDA 2021, Xi'an, China, May 28-30, 2021*, pages 137–140. IEEE.
Lin, Tianyang, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*, 3:111–132.
Liu, Tie-Yan. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.
Luan, Yuandong 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 2019, Dalian, China, 29-31 March, 2019*, pages 352–355, Dalian, China.
Madhu, Golla, A. Govardhan, and T. V. Rajinikanth. 2011. Intelligent semantic web search engines: A brief survey. *CoRR*, abs/1102.0831.
Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.
Manning, Christopher D. and Hinrich Schütze. 2001. *Foundations of statistical natural language processing*. MIT Press.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.
Minaee, Shervin, 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.
O'Shea, Jesse. 2017. Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International journal of medical informatics*, 101:15–22.
Paquet, Clara, Denis Coulombier, Reinhard Kaiser, and Massimo Ciotti. 2006. Epidemic intelligence: a new framework for strengthening disease surveillance in europe. *Eurosurveillance*, 11(12):5–6.
Parisi, Loreto, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking.
`https://github.com/musixmatchresearch/umberto`.
Raza, Shaina 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.
Raza, Shaina, 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.
Reimers, Nils and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.
Riccardo, Flavia, 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, 12.
Santana, Isabel N, 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.
Schuster, Mike and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, pages 5149–5152, Kyoto, Japan. IEEE.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.
Wang, Yu, 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.
Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.
World Health Organization. 2008. Communicable disease alert and response for mass gatherings. In *Epidemic and Pandemic Alert and Reponse*, pages 29–30, Geneva, Switzerland, April.

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