A Real-Time Open Public Sources Text Analysis System

Chi Mai Nguyen[1], Phat Trien Thai[2], Van Tuan Nguyen[3], Duy Khang Lam[4]
Viettel High Technology Industries Corporation
Ho Chi Minh City, Vietnam

***Abstract*—With the emergence of digital newspapers and social media, one can easily suffer from information overload. The enormous amount of data they provide has created several new challenges for computational and data mining, especially in the natural language processing field. Many pieces of research focusing on the information extraction process, such as named entity recognition, entity linking, and text analysis methodologies, are available. However, there is a lack of development for a system to unify all these advanced techniques. The current state-of-the-art systems are either semi-automatic or can only handle short-text documents. Most of them are not real-time or have a long lag. Some of them are domain restricted. Many of them only focus on a single source: Twitter. In this work, we proposed a system that can automatically collect, extract, and analyze information from public source text documents, like news and tweets. The system can be used in different domains, such as scientific research, marketing, and security-related domains.**

***Keywords*—named entity recognition; entity linking; text analysis system; data mining; natural language processing**

I. Introduction

We live in an age of information overload. The explosive growth of digital newspapers and online social networks creates enormous amounts of text data daily. This situation creates new challenges for computational and natural language processing (NLP) [1], [2], [3], [4], [5], [6]. While many pieces of research in the information extraction process and text analysis methodologies are available, there is a lack of development for a system to unify all these. This work proposed a system that can automatically collect, extract, analyze, and monitor information from text documents, such as named entities. The system takes in unstructured text from open public sources, like digital newspapers and Twitter, employs some information extraction processes, and records the information analysis results. It is fully automatic and has a real-time monitoring feature.

The system is not limited to any data domain. It can be applied in the news report system to automatically support reporters in detecting popular keywords. The proposed system can detect trending topics from scientific papers for research purposes. Trends are also crucial for analysts, marketing professionals, and retailers who want to monitor their and competitors' online products. Also, some government-run organizations can benefit from the proposed system by having security information about hot events related to entities from digital news or social media.

Many existing systems have applications comparable to ours, but they still have a few limitations. Some do the text analysis task on a single Twitter data stream [7], [8], [9], [10], while others only serve on their specific domains [11], [12], [13]. Also, many systems are not real-time as Trend Miner, [10], or not fully automatic as LRA Crisis Tracker, [12]. In addition, several studies analyze documents using various entity recognition methods to extract entities [9], [14], [15]. Still, they do not mention how to address the issue of many different entities co-referring to the same real-world object. To overcome these drawbacks, we propose a system with the below contributions:

- It can automatically collect data from Twitter and digital newspapers.
- It can process both long and short text.
- It employs an extra Entity Linking after the Named Entity Recognition module to map the extracted entities to their unique identities.
- It is capable of real-time processing.

The proposed system comprises six modules deployed as micro-services. First, the Data Stream collects raw documents from Twitter and multiple news sources. Next, named entities are extracted from the text by the Named Entity Recognition module. Once entities have been mapped with their identity by Entity Linking, they are fed into the Entity Tracking module to count the frequency within a time step. The Trend Detection module will take the calculated frequency and perform a trending test computation. Finally, The Entity Monitoring module will keep track of all the entities' occurrence frequencies and notify users if abnormal rising trends are detected.

A more detailed description of the system is included in Section III. Section II presents a brief overview of some current state-of-the-art event monitoring systems. Experiment results and evaluation are shown in Section IV. The summary and future works are discussed in Section V.

II. Related Works

There are a significant number of systems that analyze document streams. The majority of previous works focus on social media, such as Twitter. These include TwitterStand [16], TwitterMonitor [17], and Jasmine [18], which focus on detecting trending keywords that correspond to global or local events.

LRA Crisis Tracker[1] is a system that tracks armed group activities and conflict-related incidents in the remote border region encompassing the northeastern Democratic Republic

[1] https://crisistracker.org/

of Congo and the eastern Central African Republic. It uses data from crowd-sourcing. Domain experts must re-examine the data before feeding it to the system. In contrast to the Crisis Tracker system, the proposed system is fully automatic.

Redites [7] is a system that monitors events based on Twitter's tweets. The system will decide if a new tweet contains a new event whenever it is streamed. The event is then classified. Only events in security categories are analyzed and monitored. Redites is fully automatic. However, it can only handle short text documents like tweets from Twitter. The proposed system can handle both long- and short-text documents.

Social Sensor[2] tracks and monitors predefined events on social media [19]. The proposed system automatically extracts the events from news and tweets. Moreover, Social Sensor can only handle short text documents like tweets.

Trend Miner[3] is similar to our proposed system; however, it does not focus on real-time aspects and can only handle short-text documents [20].

Cheng et al. [10] proposed an early warning system to detect COVID-19 outbreaks based on Twitter's data. The system has a 6-27 days lag and is only applicable for short-text analysis.

Epitweetr [11] is a system that detects outbreaks and public health threats using Twitter's data. The system is fully automatic and provides additional support for public health experts in detecting and identifying the geo-location of an outbreak. However, compared to the proposal, the system is restricted by data domain and can only handle short-text documents.

Goh et al. [12] proposed an approach using structured data and unstructured clinical notes to predict and diagnose sepsis. This approach is, however, restricted by the data domain and is not a fully automatic system.

A real-time system [9] was built to ingest the stream of all tweets and identify clusters of event-related entities on a minute-by-minute basis. The system first extracts named entities, hashtags and internal knowledge graph entities from each tweet. It keeps only trending entities detected by its internal Trend Detection system known as Twitter Trends. Then, a weighted graph is naturally constructed with the entities as nodes and their cosine similarities as edge weights. Community detection algorithms have been used to detect the sub-graphs (clusters) based on their links to others. To solve the Event Evolution problem, they add a layer of Cluster Linking to form cluster chains. Finally, clusters are ranked based on the aggregate popularity of their entities and persisted in internal stores for future use.

EveSense [8] is an Event Detection application that detects real-life events and related trending topics from the Twitter stream and allows users to find interesting events that have recently occurred.

News Monitor [15] is a scalable real-time framework for analyzing and exploring news articles. The system collects news from many RSS news sources and automatically extracts the main content and other metadata. News Monitor is analogous to our approach compared to many news services, such as Event Registry [21]. However, it cannot handle the entity ambiguity problem when data diversity increases.

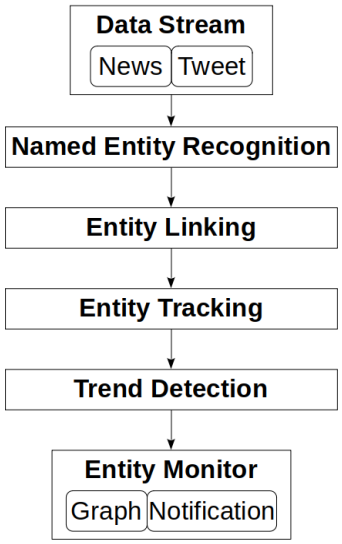

Fig. 1. System diagram

Our system can handle both long texts like news and short texts like tweets. It is also capable of analyzing these data sources simultaneously. The design, thus, can explore the potential relationship between the information flows to produce more consistent, accurate, and valuable information. Moreover, the system has an Entity Linking module to assign a unique identity to entities mentioned in the text, which helps robust join and union operations that can integrate information about entities [22].

III. Main Components

The system is engineered in modules and is extensible. Each module is a separate service that can be modified or removed without affecting the whole system. A new module can also be added easily. Fig. 1 presents an overview of the system.

A. Data Stream

The Data Stream module handles the task of streaming posts from social media (such as tweets from Twitter) and news from digital newspapers. The module uses Scrapy[4] and Automation Browser[5] for its streaming process. This module is updated accordingly to the newspapers and social media structures.

B. Named Entity Recognition

The Named Entity Recognition (NER) [23], [24] module handles the task of extracting and classifying named entities in

[2]http://socialsensor.iti.gr/
[3]http://www.trendminer-project.eu/
[4]https://scrapy.org/
[5]https://axiom.ai/

Fig. 2. Named entity recognition process

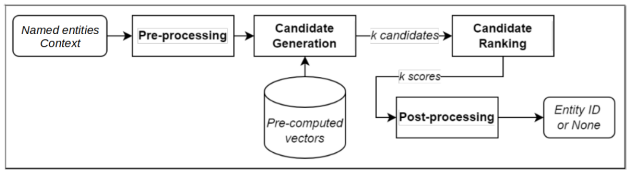

Fig. 3. Entity linking process

unstructured text into predefined categories. Two deep learning approaches, namely Bidirectional Encoder Representations from Transformers (BERT) [25], [26] and Bidirectional Long Short-Term Memory (BiLSTM) [27], [28], [29], were used in this module. Each sentence from an unstructured text document will be tokenized and vectorized using the BERT model. The result vectors are fed to the BiLSTM model for detecting and classifying the named entities and their categories, as shown in Fig. 2.

C. Entity Linking

The Entity Linking (EL)[30] module links the named entities extracted from the NER module to their corresponding predefined unique identity. For example, the entities "US", "USA", and "United States" have the same identity as "United States of America". Whereas "New York" is the name for both a city and a state in the United States of America, they have two separate identities. Fig. 3 shows how the entity linking process works in this module. The core of this module is BERT deep learning approach [31], [32].

D. Entity Tracking

The Entity Tracking module handles the task of tracing and tracking the named entities from each unstructured document. The module computes and keeps a record of the occurrence frequency of each named entity in a defined time step. The entities are then ranked by their occurrence frequencies. Top-ranked entities are fed to a trend detection module to examine if there is an abnormal increasing trend. This work is a real-time process. Data is streamed and processed simultaneously.

E. Trend Detection

The Trend Detection module uses the Mann-Kendall approach [33], [34], [35], [36] to detect an abnormal increasing trend in the occurrence frequency of an entity. The Entity Monitoring module will notify the users if a rising trend is seen.

F. Entity Monitoring

The Entity Monitoring module controls both the Entity Tracking module and the Trend Detection module. It handles the task of monitoring the entities and notifying users if there is an anomaly, as shown in Fig. 4.

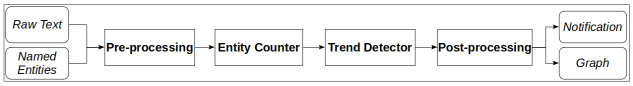

Fig. 4. Entity monitoring process

IV. Experiments and Evaluation

We have implemented and tested our system. We manually labelled 927 text documents with 20000 sentences and 14 different entity categories to train and evaluate the NER module. After training, we used a set consisting of 1500 sentences for testing the trained module. The final f1-score of the module is 91.41%. The EL module is trained and tested on a dataset consisting of 4475 entities, 1244 of which were manually labelled. The recall rate of the EL module is 97.2%. The system is capable of monitoring up to 3000 entities simultaneously.

Currently, the system can automatically collect and process news from 63 newspapers and public tweets in English from Twitter. We tested the system on a dataset of news articles and tweets collected from mentioned sources from July 15, 2022 to October 20, 2022. We chose July 21, 2022 as a virtual "current date" and let the system simulate the process of extracting and analyzing data. The results of the analysis are visualized as graphs.

The systems will automatically monitor the top-ranked entities when nothing is specified, as mentioned in Section III-D. Fig. 5 and Fig. 6 show the example graphical presentations of the tracking and the monitoring processes over multi entities, respectively. As a default, the system will monitor the top 100 entities; however, we chose to showcase only five of them in the figures for readability. Each subgraph in Fig. 5 presents the time series of the occurrence frequency of an entity, i.e., the number of news articles and tweets mentioning the entity during some predefined time step. Each subgraph in Fig. 6 presents the monitoring process of an entity. The blue line presents the occurrence frequency of an entity, the yellow line shows the plot of the normalized Mann-Kendall values of that entity, and the vertical red lines indicate the time when abnormal rising trends are detected.

The system also allows users to choose the monitored entities and sources freely. Fig. 7 shows an example graphical presentation of the monitoring process of one specific entity. The upper graph presents the time series of the occurrence frequency of the entity. In the lower graph, the yellow line shows the plot of the computed normalized Mann-Kendall values of the entity, and the vertical red lines indicate the time when abnormal rising trends are detected. Each red line presents an anomaly. Fig. 8 shows a monitoring process of three entities of interest on Twitter, and Fig. 9 shows the same process on digital newspapers.

Moreover, the system can detect if two (or more) entities simultaneously have an abnormal rising trend. The upper graph of Fig. 10 shows the plots of two occurrence frequency time series of the two entities in comparison. The lower graph's orange and blue lines present the computed normalized Mann-Kendall values of the two entities. At the same time, the vertical red lines indicate the time when both entities have

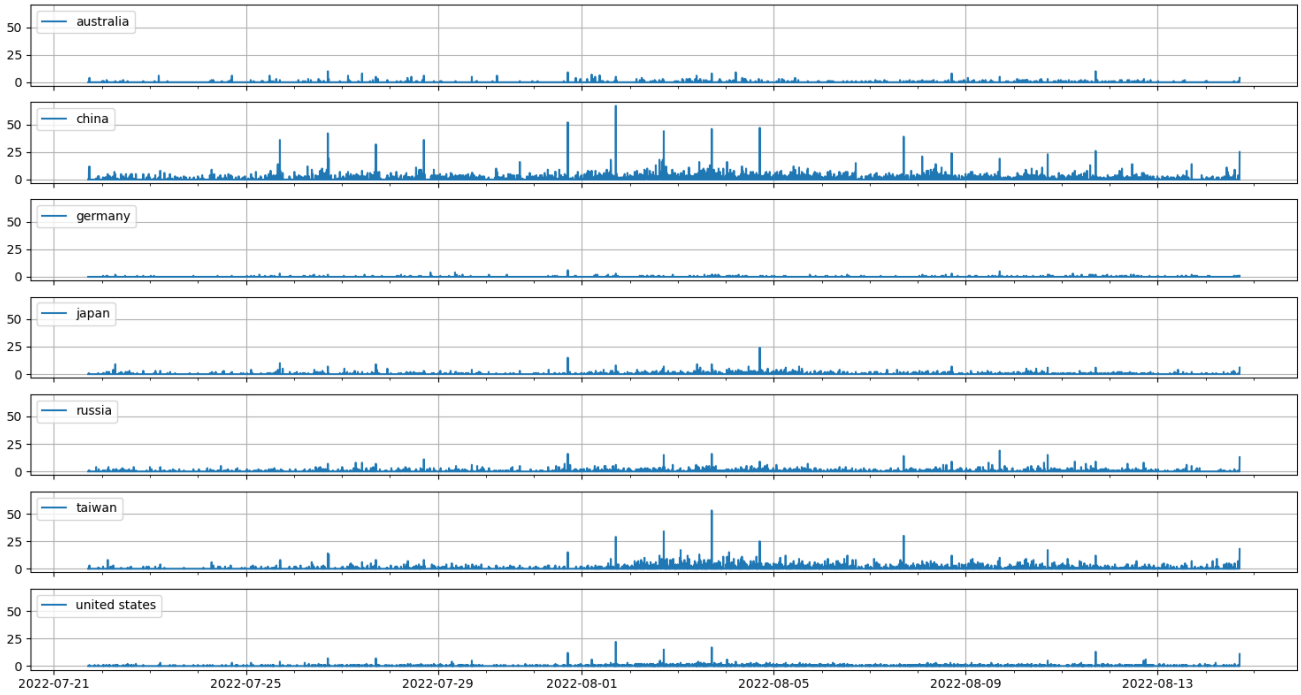

Fig. 5. Example of multi-entities tracking

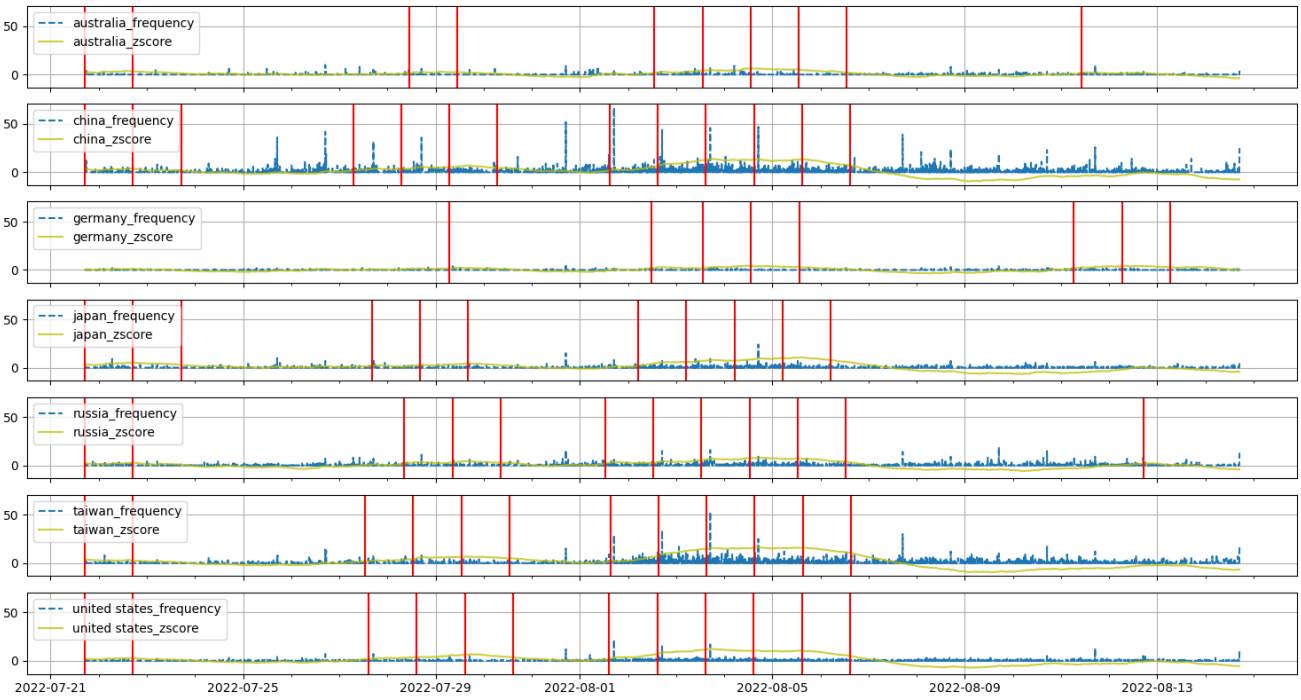

Fig. 6. Example of multi-entities monitoring

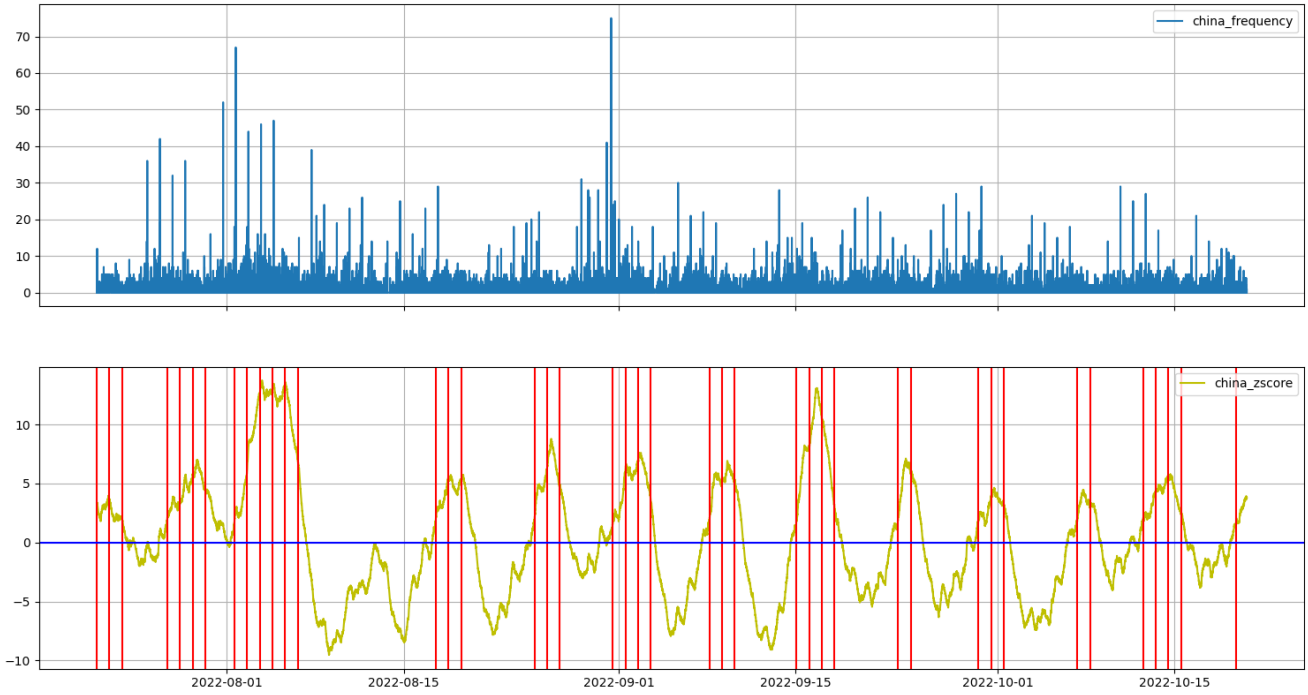

Fig. 7. Example of an entity monitoring process

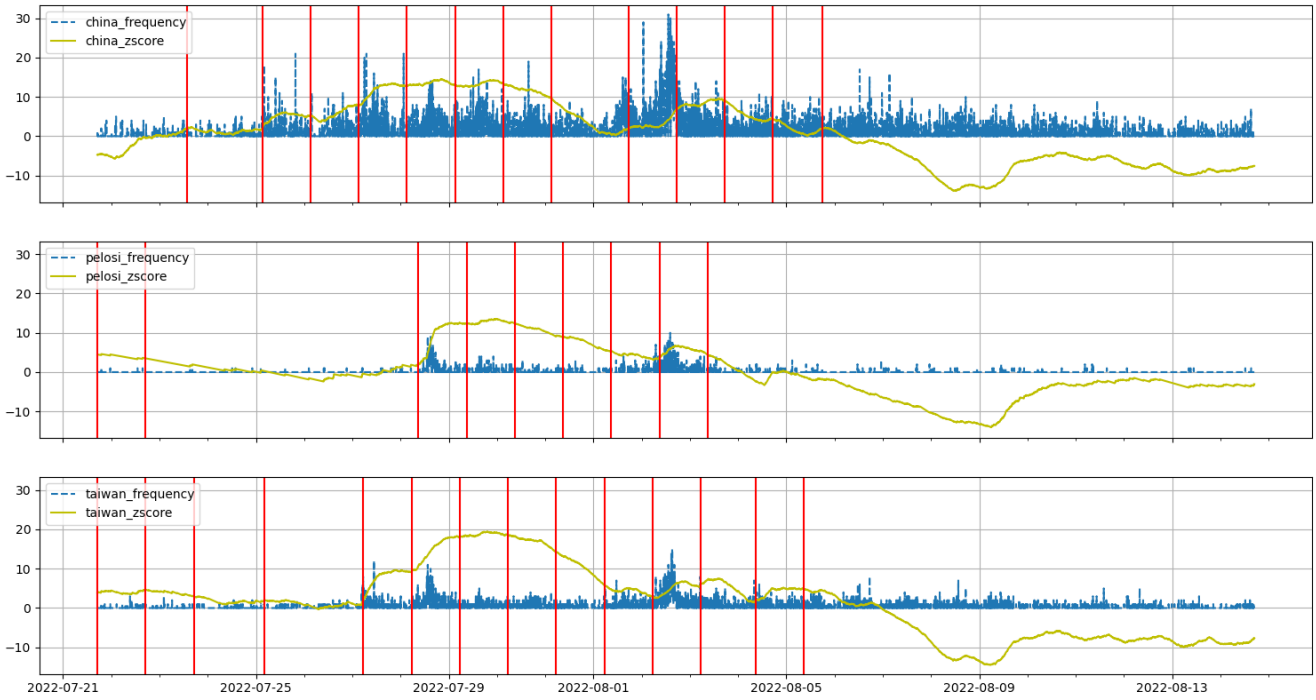

Fig. 8. Example of selected entities monitoring on twitter

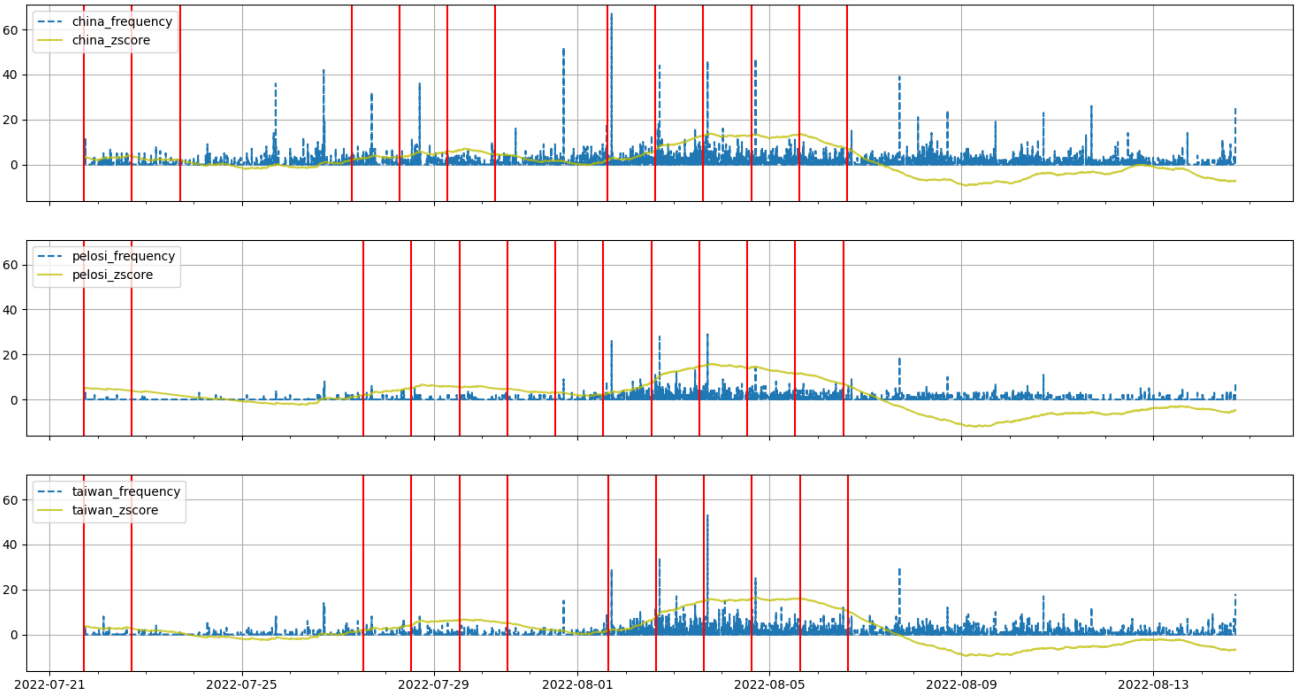

Fig. 9. Example of selected entities monitoring on news

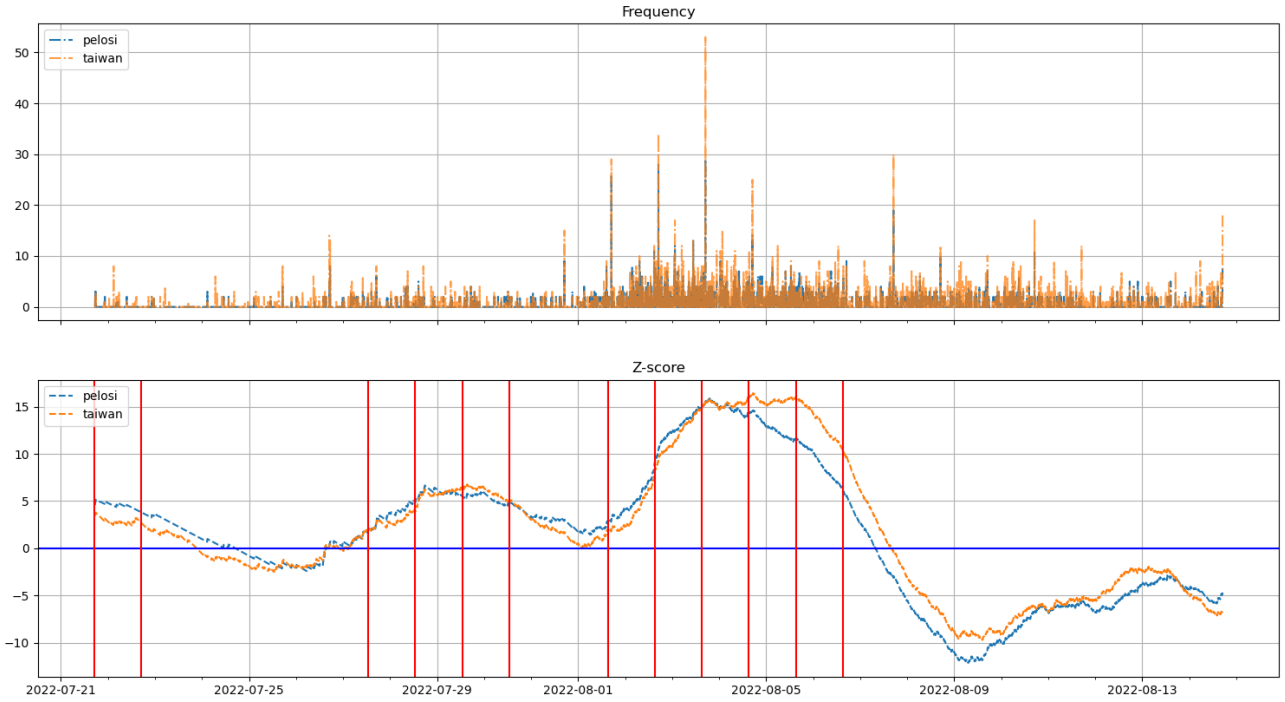

Fig. 10. Example of trend comparison between two entities

abnormal rising trends.

In all graphs, the y-axis presents values, and the x-axis is time.

V. Conclusions

We presented a system capable of automatically collecting, extracting, and analyzing information from open sources such as digital newspapers and Twitter. The system provides significant help for online exploratory analysis. Since the system is flexible, it can be used in different domains, such as scientific research, marketing, and security-related domains. As stated, the system is extensible. Possible future works may include but are not limited to (i) an event extraction module that detects and classifies possible events in which entities are involved, (ii) a sentiment analysis module that determines emotion about events, (iii) a language detection module that handles multi-lingual data, (iv) machine translator module that translates documents into a specific language, (v) a graphical user interface that is friendly for users in different domains, and (vi) a visualization improvement for the data and the outcome analysis.

References

[1] K. Jaseena, J. M. David *et al.*, "Issues, challenges, and solutions: big data mining," *CS & IT-CSCP*, vol. 4, no. 13, pp. 131–140, 2014.

[2] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[3] K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[4] H. A. Schwartz and L. H. Ungar, "Data-driven content analysis of social media: A systematic overview of automated methods," *The ANNALS of the American Academy of Political and Social Science*, vol. 659, no. 1, pp. 78–94, 2015.

[5] L. S. Lai and W. M. To, "Content analysis of social media: A grounded theory approach," *Journal of Electronic Commerce Research*, vol. 16, no. 2, p. 138, 2015.

[6] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools and Applications*, pp. 1–32, 2022.

[7] M. Osborne, S. Moran, R. McCreadie, A. Von Lunen, M. Sykora, E. Cano, N. Ireson, C. Macdonald, I. Ounis, Y. He *et al.*, "Real-time detection, tracking, and monitoring of automatically discovered events in social media," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 37–42.

[8] Z. Saeed, R. Abbasi, and I. Razzak, *EveSense: What Can You Sense from Twitter?*, 04 2020, pp. 491–495.

[9] M. Fedoryszak, B. Frederick, V. Rajaram, and C. Zhong, "Real-time event detection on social data streams," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, jul 2019.

[10] I. Cheng, J. Heyl, N. Lad, G. Facini, and Z. Grout, "Evaluation of twitter data for an emerging crisis: an application to the first wave of covid-19 in the uk," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[11] L. Espinosa, A. Wijermans, F. Orchard, M. Höhle, T. Czernichow, P. Coletti, L. Hermans, C. Faes, E. Kissling, and T. Mollet, "Epitweetr: Early warning of public health threats using twitter data," *Eurosurveillance*, vol. 27, no. 39, p. 2200177, 2022.

[12] K. H. Goh, L. Wang, A. Y. K. Yeow, H. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan, "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[13] B. Alkouz, Z. Al Aghbari, and J. H. Abawajy, "Tweetluenza: Predicting flu trends from twitter data," *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 273–287, 2019.

[14] A. Al-Laith and M. Shahbaz, "Tracking sentiment towards news entities from arabic news on social media," *Future Generation Computer Systems*, vol. 118, pp. 467–484, 2021.

[15] A. Saravanou, N. Panagiotou, and D. Gunopulos, *News Monitor: A Framework for Querying News in Real Time*, 03 2021, pp. 543–548.

[16] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling, "Twitterstand: News in tweets," 01 2009, pp. 42–51.

[17] M. Mathioudakis and N. Koudas, "Twittermonitor: Trend detection over the twitter stream," 06 2010, pp. 1155–1158.

[18] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs," 10 2011, pp. 2541–2544.

[19] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," *IEEE Transactions on multimedia*, vol. 15, no. 6, pp. 1268–1282, 2013.

[20] D. Preoţiuc-Pietro and T. Cohn, "A temporal model of text periodicities using gaussian processes," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 977–988.

[21] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry: learning about world events from news," 04 2014, pp. 107–110.

[22] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, pp. 443–460, 2015.

[23] B. Mohit, "Named entity recognition," in *Natural language processing of semitic languages*. Springer, 2014, pp. 221–245.

[24] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, and C. Zhang, "Bond: Bert-assisted open-domain named entity recognition with distant supervision," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1054–1064.

[27] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana, "Modelling radiological language with bidirectional long short-term memory networks," *arXiv preprint arXiv:1609.08409*, 2016.

[28] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification," in *Proceedings of the 29th Pacific Asia conference on language, information and computation*, 2015, pp. 73–78.

[29] G. Liu and J. Guo, "Bidirectional lstm with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[30] D. Rao, P. McNamee, and M. Dredze, "Entity linking: Finding extracted entities in a knowledge base," in *Multi-source, multilingual information extraction and summarization*. Springer, 2013, pp. 93–115.

[31] L. Wu, F. Petroni, M. Josifoski, S. Riedel, and L. Zettlemoyer, "Scalable zero-shot entity linking with dense entity retrieval," *arXiv preprint arXiv:1911.03814*, 2019.

[32] A. Hamdi, E. Linhares Pontes, E. Boros, T. T. H. Nguyen, G. Hackl, J. G. Moreno, and A. Doucet, "A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2328–2334.

[33] H. Mann, "Non-parametric test against trend. econometrical, 1945 (13): 245-259."

[34] M. G. Kendall, "Rank correlation methods." 1948.

[35] R. O. Gilbert, *Statistical methods for environmental pollution monitoring*. John Wiley & Sons, 1987.

[36] F. Wang, W. Shao, H. Yu, G. Kan, X. He, D. Zhang, M. Ren, and G. Wang, "Re-evaluation of the power of the mann-kendall test for detecting monotonic trends in hydrometeorological time series," *Frontiers in Earth Science*, vol. 8, p. 14, 2020.