What will happen tomorrow? Predicting future event types for businesses

Tesia Šker
Jožef Stefan Institute
Ljubljana, Slovenia
tesia.sker@gmail.com

Jože M. Rožanec
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
joze.rozanec@ijs.si

Gregor Leban
Event Registry d.o.o.
Ljubljana, Slovenia
gregor@eventregistry.org

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

Strategic foresight helps organizations anticipate future challenges and opportunities, allowing them to handle uncertainty better. While strategic foresight is becoming more widely adopted across organizations, the process still heavily relies on expert knowledge, and little of it has been automated through artificial intelligence. In this research, we explore how media news events can be analyzed to forecast event types that will take place in the near future. In particular, we consider it a supervised machine learning problem with a well-defined set of event types and leverage graph representation of the media news events to create graph embeddings, train a classifier, and predict event types that will likely occur one day ahead. We validated our approach on a real-world dataset of an American multinational conglomerate operating in industry, worker safety, healthcare, and consumer goods.

KEYWORDS

strategic foresight, event prediction, machine learning, graphs

1 INTRODUCTION

Strategic foresight helps organizations anticipate future challenges and opportunities, allowing them to handle uncertainty better [9]. Therefore, predicting future event types as a part of strategic foresight became necessary for businesses to manage their operations without significant losses. Various events on a major scale, such as floods, earthquakes, internet failures, or pandemics, as we are witnessing recently, or on a minor scale, such as road closures due to sports events or promotions at fairs, can have a major impact on business operations. By predicting the next event type, businesses can adjust prices, reschedule staff, manage stocks, reschedule transportation routes to avoid delays, and more, and thus reduce losses or increase their sales and profits.

There is currently a massive number of articles written on Future Event Predictions. Based on Zhao [11], the event prediction methods can be classified in terms of goals into time prediction, location prediction, semantics prediction, and a combination of these. Each goal is divided into subgoals for which various techniques can be applied. According to the classification provided by Zhao, our technique can be classified as a semantic prediction.

In this research, we explore how graphs can be used to model media news events and to forecast event types in the near future. By doing so, we provide a valuable tool for decision-makers, offering them a clearer view of potential outcomes. Specifically, our research focuses on using a JSON dataset containing a variety of articles about a particular business company. We create a graph representation of the articles and use Graph2Vec to create embeddings that can be used downstream to fit other machine-learning models. Using this information, we apply a Random Forest Classifier to predict the categories of articles about the company for the following day.

In particular, we expect this to be useful to give organizations a competitive advantage in fast-changing markets [5]. While human expertise is valuable, it varies from person to person, leading to inconsistent predictions. Manually analyzing large datasets is also time-consuming and prone to errors. AI, however, can process vast amounts of data, spot patterns, and predict future event types more accurately.

This work is structured as follows. Section 2 presents related work that is relevant for this paper. Section 3 describes the data in the dataset, and the data extraction process. Section 4 introduces a new approach to predict future event types. Section 5 presents the results of this research. Section 6 concludes this work and proposes future improvements.

2 RELATED WORK

In recent decades there has been an increasing interest in strategic foresight in the academic field. According to Fergnani (2020) [2] this is because by "using corporate foresight, organisations can reconfigure their strategy based on the analysis of business opportunities suggested by future possibilities". Even in academia "one of the domains heavily impacted by Artificial Intelligence is innovation management and in this context especially the area of Strategic Foresight (SF)" as per Brandtner et. al (2021) [1].

However it seems that strategic foresight methods related to AI only end up being used by bigger companies with a larger number of resources. As noted by Kim and Seo (2023) [6], "except for AI start-ups and players in the consumer electronics and information and communication industry, small- and medium-sized enterprises (hereafter SMEs) in other industries do not demonstrate competence in AI." Therefore, effective implementation of AI solutions for strategic foresight in smaller and medium sized

Jože M. Rožanec and Tesia Šker are co-first authors with equal contribution and importance.
Corresponding author: Jože M. Rožanec: joze.rozanec@ijs.si.

Information Society 2024, 7–11 October 2023, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.24

companies would be one of the topics to be explored in future research.

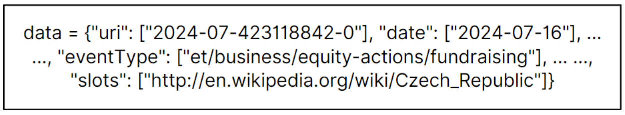
In this research however, we focus more on the general implementation of strategic foresight by means of next event prediction. Exploring similar fields, we found that there was already some research exploring the field of event predictions, which rather than focusing on businesses focused on other domains. In the field of sequential event prediction, several researchers are exploring diverse methods. Although the methods share some conceptual similarities with our research, they differ significantly in methodology and focus. Letham, Rudin, and Madigan (2013) [7] developed a model that predicts the next event using an ERM-based approach with logistic regression, focusing on the presence of events rather than their order. On the other hand, our work uses labeled article databases and considers the sequence of past events, using techniques like graph construction, random walks, and random forests. Yeon, Kim, and Jang (2015) [10] focus on predicting event flow through visual analytics, using LDA for topic extraction and emphasizing specific keywords, while our approach is entirely text-based and relies on graphs. On the other hand, Hu et al. (2017) [4] use LSTM networks for predicting future subevents, which offers an alternative method to our non-LSTM-based text analysis.

Although these studies provide useful insights and have offered significant improvement in sequential event prediction, they may face certain challenges. For instance, Letham, Rudin, and Madigan (2013) [7] emphasize event presence over sequence, potentially missing key temporal relationships, while Yeon, Kim, and Jang (2015) [10] depend heavily on keywords, overlooking broader context. Additionally, LSTM-based models like those used by Hu et al. (2017) [4] are powerful however they require significant computational power. In contrast, our work addresses these limitations by employing a graph-based approach that prioritizes event sequences and leverages standardized data from sources like DMOZ and Wikipedia. This enables us to make more accurate and efficient predictions, offering a practical and scalable solution that enhances predictive accuracy.

3 DATASET

3.1 Data Extraction Pipeline

The event detection pipeline processes about 300.000 English news articles per day. Each news article is first annotated using tools like entity linking, topic classification and sentiment detection. Each article is then split into sentences where each sentence retains it's annotations and other meta-data. For each pair of the entities in the sentence, an event classifier then determines if there is a particular relation of interest expressed in the sentence between the two entities. The predefined taxonomy currently includes 133 event types of interest, ranging from security, environment, natural disasters, accidents, politics, and other areas. To classify the events, a neural network transformer architecture with a pretrained encoder is used. The entire network, including the encoder, is trained on our supervised dataset using best practices like online hard example mining, class balancing, dropout, and consistency regularization. The sentences for which the classifier finds that it mentions a relation of interest are then stored in a database, together with the pair of associated entities and other available meta-data.



```
data = {"uri": ["2024-07-423118842-0"], "date": ["2024-07-16"], ...
..., "eventType": ["et/business/equity-actions/fundraising"], ... ...,
"slots": ["http://en.wikipedia.org/wiki/Czech_Republic"]}
```

Figure 1: Sample of relevant data considered when parsing an event type to build the dataset.

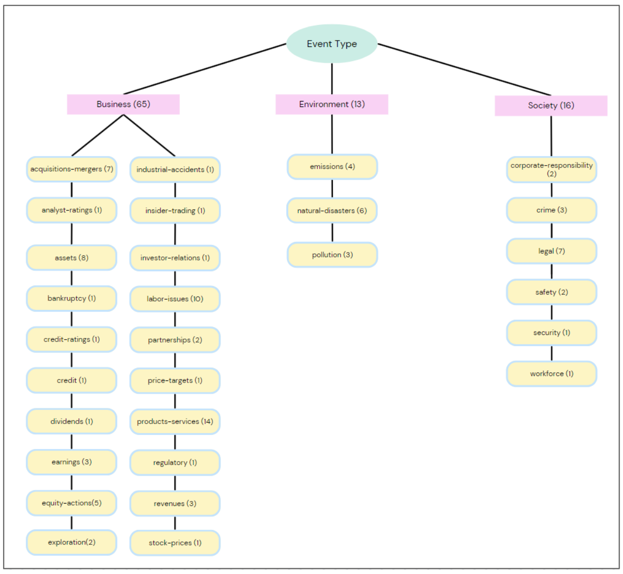


Figure 2: Event Type Taxonomy

3.2 Data Description

For our research, we used a dataset of events provided by Event Registry, with media events encoded in JSON format. Specifically, we analyzed 4,216 events related to the company 3M, recorded between June 23, 2021, and July 23, 2024. We used a URI to classify each event, drawing from DMOZ and Wikipedia categories (Fig. 1). These were selected because they provide standardized descriptions of the events being reported, which makes the data consistent and reliable. The events are categorized into 94 distinct types, which are further grouped into three primary domains: business, environment, and society. The business domain makes up the largest proportion of events, accounting for 65 types (69% of the total), while the environment and society domains contain 13 types (14%) and 16 types (16%), respectively. Within these domains, the event types are further divided into smaller subdomains, which can be aggregated into larger subdomain units as demonstrated in the event type taxonomy (Fig. 2).

4 METHODOLOGY

This study uses graph-based techniques to predict future event types from news articles about a specific company. The process starts by building a graph that maps relationships between event types and concepts from Wikipedia and DMOZ. Random walks are then performed on this graph to extract key information such as URIs, dates, and event types, which are then transformed into embeddings using Graph2Vec [8]. Next, the event types are encoded and adjusted through a process called target shifting. This step aligns the features to better forecast future outcomes based on previous data. The predictions are made using a Random Forest classifier, which is then validated through stratified k-fold

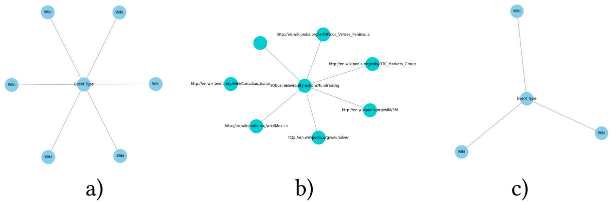

Wiki
Event Type
http://en.wikipedia.org/wiki/Palos_Verdes_Peninsula
http://en.wikipedia.org/wiki/OTC_Markets_Group
http://en.wikipedia.org/wiki/Canadian_dollar
et/business/equity-actions/fundraising
http://en.wikipedia.org/wiki/3M
http://en.wikipedia.org/wiki/Mexico
http://en.wikipedia.org/wiki/Silver
a)
b)
c)

Figure 3: Event Type Graphs

cross-validation for higher accuracy. The following sections will present each step of this process in more detail (see Fig. 4).

4.1 Graph Construction

For each article in the JSON dataset, a detailed graph G is generated using the NetworkX library [3]. The graph construction process starts by extracting key information such as the article's URI(unique identifier), as well as the date associated with the article and the event types, which are represented by specific URIs. In addition to these elements, each article also includes two important lists: 'slots' and 'categories'. The 'slots' list contains wiki and dmoz addresses that are directly related to the event described in the article, while the 'categories' list includes various classifications of the event. To complete the graph, labels are created by extracting URIs from the 'slots' list and filtering the 'categories' to focus on those with the "dmoz" prefix.

4.2 Random Walks for Feature Extraction

Once the graphs for each article are constructed, random walks are performed, starting at a given node (event type) and moving to adjacent nodes based on specific probabilities. Several random walks are generated for each node, forming the foundation for feature extraction processes. A single random walk begins by initializing the path with the starting node and iterating over a specified path length. At each step, a random number is compared with a probability p. If the number is less than p, the walker stays at the current node, otherwise it moves to a random neighbor. If no neighbors are available, the walk ends.

Generating multiple random walks for every node follows a similar approach, using p as the probability of staying at the current node (set at 0.05). The process involves creating an empty list to store all random walks and iterating through each node in the graph. For each node, the specified number of random walks is generated, and each walk is appended to the list.

4.3 Embedding Generation Using Graph2Vec

The random walks from the graphs are processed similarly to word sequences in a document. The 'embedding_data' function generates vector embeddings for graph data using the Doc2Vec model. It begins by converting each random walk into a TaggedDocument, storing these in 'documents_gensim'. The Doc2Vec model, with a vector size of 5 is trained on these documents, creating a vector space where similar sequences are positioned close together.

The function then processes each graph in the graphs dictionary, extracting uri, date, and event type, and generating additional random walks. These walks are converted into embeddings using the 'infer_vector' method, and the resulting vectors are averaged into one final embedding. This embedding is stored in a dictionary across 'embedding1' to 'embedding5', alongside the graph's metadata.

4.4 One Hot Encoding & Target Shifting

To transform the categorical event types into binary vectors, One hot encoding is applied. This allows the model to treat each event type as a separate class. After extracting relevant column names, the encoded target data is concatenated with the feature embeddings, creating a dataset for model training and evaluation. The dataset is then aggregated by averaging out the embeddings and calculating the maximum value of the encoded target columns for a given day. Finally the 'target' data is shifted by one day, which allows the embeddings to forecast the event types for the following day.

4.5 Random Forest Classification & Stratified K-Fold Cross Validation

To ensure an effective classification and prediction of the data, A Random Forest classifier is created. When employing this method, embeddings are used as features and the one-hot encoded event types are used as labels. The data itself is split into testing and training sets, followed by the incorporation of the Stratified K-Fold cross validation. This technique splits the data into 10 folds, while ensuring that the event type proportion in each fold remains equal. The model is then trained on 9 folds, with the remaining fold being used for validation. This ensures balanced representation of each class across the folds resulting in a more effective performance.

5 RESULTS

As mentioned above, the model was trained on a training set, and then evaluated on a test set. The training set included approximately 508 samples for each fold, and the test set included about 10% of the whole set, which amounted to 56 samples per fold. Using this, the model then predicted the probabilities for event types for each set. When training the model for each class, we noticed certain classes did not have enough occurrences to have at least one entry of such a class per dataset fold and were skipped. We, therefore, trained the model and predicted for a total of 45 classes.

To evaluate the discriminative performance of the model, the ROC AUC score was used. The results produced showed us how well the model distinguishes between different classes, as well as the model's ability to predict future event types. The ROC AUC score showed us that the average performance of the model was around 0.5674, and the median was close to it, with an AUC ROC score of 0.5559, with the highest score reaching 0.8194 and the lowest reaching a value of 0.3338. While the best scores demonstrate we can effectively forecast event types ahead of time, further work is required to enhance results, which in most cases remain close to 0.5.

6 CONCLUSIONS

This study was used to develop a graph-based approach to predicting event types in articles. In the process, we utilized random walks for feature extraction and Doc2Vec for embedding generation. Then, we trained the resulting model on a Random Forest classifier and evaluated it with a Stratified K-Fold Cross Validation. The model demonstrated solid performance with an average ROC AUC score of around 0.5674, reaching a peak at approximately 0.8194. This indicates the model's effectiveness in capturing relationships within the data and predicting future event types.

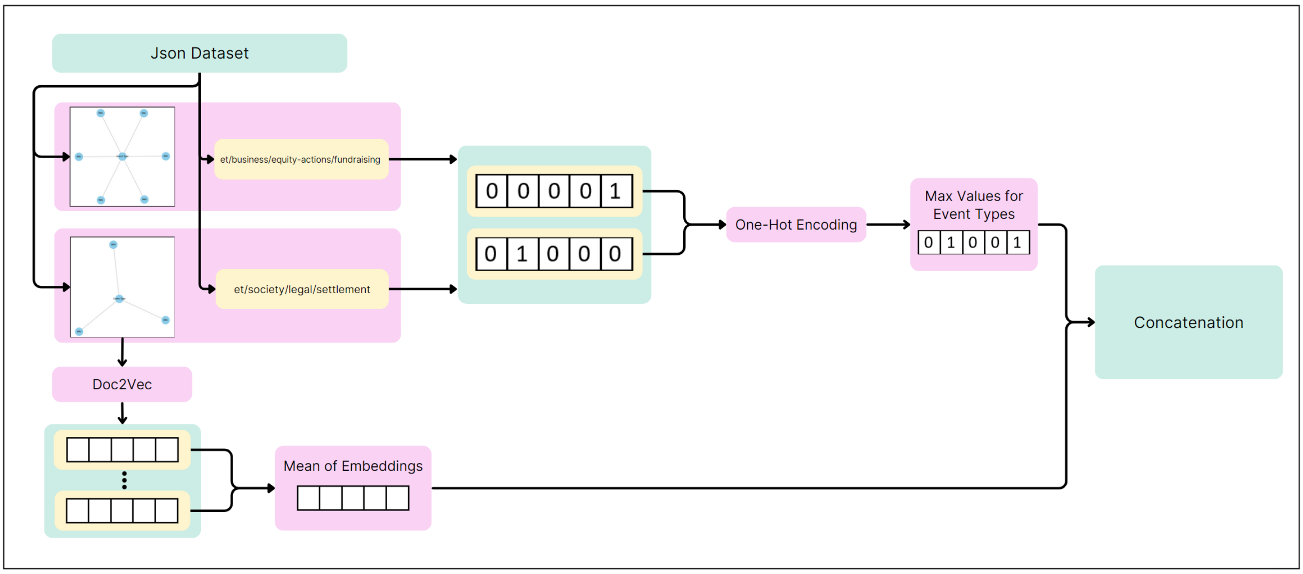


Figure 4: Data Extraction Pipeline

However, while the model performed well overall, occasional fluctuations in accuracy suggest space for further improvement. We are currently striving to find ways to make graphs more informative. In future work we could refine the feature extraction process by incorporating larger datasets, with a wider variety samples and a larger number of companies.

ACKNOWLEDGMENTS

The Slovenian Research Agency supported this work. This research was developed as part of the Graph-Massivizer project funded under the Horizon Europe research and innovation program of the European Union under grant agreement 101093202.

REFERENCES

[1] Patrick Brandtner and Marius Mates. 2021. Artificial intelligence in strategic foresight–Current practices and future application potentials: current practices and future application potentials. In *Proceedings of the 2021 12th International Conference on E-business, Management and Economics*. 75–81.

[2] Alex Fergnani, Andy Hines, Alessandro Lanteri, and Mark Esposito. 2020. Corporate foresight in an ever-turbulent era. *European business review* 25 (2020), 26–33.

[3] Aric Hagberg, Pieter J Swart, and Daniel A Schult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

[4] Linmei Hu, Juanzi Li, Liqiang Nie, Xiao-Li Li, and Chao Shao. 2017. What happens next? Future subevent prediction using contextual hierarchical LSTM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[5] Jon Iden, Leif B. Methlie, and Gunnar E. Christensen. 2017. The nature of strategic foresight research: A systematic literature review. *Technological Forecasting and Social Change* 116 (2017), 87–97. https://www.sciencedirect.com/science/article/pii/S0040162516306035

[6] Jong-Seok Kim and Dongsu Seo. 2023. Foresight and strategic decision-making framework from artificial intelligence technology development to utilization activities in small-and-medium-sized enterprises. *foresight* 25, 6 (2023), 769–787.

[7] Benjamin Letham, Cynthia Rudin, and David Madigan. 2013. Sequential event prediction. *Machine learning* 93 (2013), 357–380.

[8] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005* (2017).

[9] Freija van Duijne and Peter Bishop. 2018. Introduction to strategic foresight. *Future* 1 (2018), 67.

[10] Hanbyul Yeon, Seokyeon Kim, and Yun Jang. 2015. Visual Analytics using Topic Composition for Predicting Event Flow. *KIISE Transactions on Computing Practices* 21, 12 (2015), 768–773.

[11] Liang Zhao. 2021. Event Prediction in the Big Data Era. *Comput. Surveys* 54, 5 (2021), 1–37.