Predicting Business Events from News Articles

Thibault Ehrhart[1,*,†], Raphaël Troncy[1,†], David Shapira[2] and Bertrand Limoges[2]

[1]EURECOM, Campus SophiaTech, France
[2]Pairing Technology Capital, France

Abstract
This paper presents a comparative study of different approaches for predicting business events from news articles. We evaluate the effectiveness of zero-shot classification models that use Large Language Models (LLM), methods relying on NLI, and a supervised approach using a fine-tuned BERT-based classifier. We also propose a novel ensemble method that combines spaCy, CamemBERT, and FlairNLP models to semantically annotate the news articles in terms of named entities. We discuss the strengths and limitations of each family of approaches that contribute to the development of tools for accurate event prediction from news articles. The public demonstration available at https://jde-predict.tools.eurecom.fr/ enables the user to submit news articles and to visualize the extracted and predicted semantic annotations. In addition, a SPARQL interface is exposed enabling to search through annotations of news articles.

Keywords
Business events prediction, Zero-shot classification methods, NLI models, ZeSTE, LLM (GPT-4, Claude),

1. Introduction

In today's fast-paced and information-driven world, the ability to extract valuable insights and predict business events from news articles is very important. We focus on business events such as mergers and acquisitions, investment, new sites or management in organizations, product launches, etc. which provide investors and decision-makers with a competitive edge and facilitate informed decision-making. In this work, we present a powerful tool for predicting business events from news articles sourced specifically from *Le Journal Des Entreprises*, an independant French press company oriented towards business news. We compare various methods for predicting named events, including, i) leveraging on (closed) Large Language Models (LLMs) such as ChatGPT, GPT-4 and Claude, ii) Natural Language Inference for Zero-Shot Topic Extraction (entail and ZeSTE methods), or iii) a supervised classification method that uses a fine-tuned BERT-based classifier which proves to obtain the best classification results.

2. Method

In this section, we explore three family of methods: Zero-Shot classification using Large Language Model, Zero-Shot classification leveraging Natural Language Inference (NLI) and supervised classification. We select representative algorithms of these different families as they

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece
thibault.ehrhart@eurecom.fr (T. Ehrhart); raphael.troncy@eurecom.fr (R. Troncy);
david.schapira@pairingtechnology.com (D. Shapira); bertrand.limoges@pairingtechnology.com (B. Limoges)
0000-0003-1377-8279 (T. Ehrhart); 0000-0003-0457-1436 (R. Troncy)

exhibit different advantages and limitations, namely: no need for labeling data for zero-shot classification approaches but with the risk of getting hallucinations for LLM or insufficient background knowledge in the case of NLI; the a priori most accurate classification with a supervised approach but with the limitation of getting labelled data in sufficient volume for under-represented classes.

Additionally, we extract named entities and topics using state-of-the-art models. The news articles are written in French. Experts have, beforehand, elicited 11 named business events of interest that are enumerated in Listing 1. This is a multiclass classification problem as one news article can be annotated in terms of multiple business events.

2.1. Predicting business events from news articles

2.1.1. Zero-Shot method using LLM

We first exploit the capabilities of pre-trained Large Language Models and we perform prompt engineering to design context-specific prompts that guide the LLMs towards predicting business events accurately. We experiment with four different LLMs: bloomz-176b, gpt-3.5-turbo, gpt-4, and claude 1.3.

```
Text to classify: {text}
Return up to 3 category numbers, comma separated, among the following choices
    , and only if explictly described:
1. Buyout / Transfer (Rachat / Cession)
2. Fundraising (Levée de fonds)
3. New site (Nouvelle implantation)
4. Change of Manager (Changement de Dirigeant)
5. Safeguard procedure (Procédure de sauvegarde)
6. Site closure (Fermeture de site)
7. Job creation / recruitment (Création d'emploi / recrutement)
8. Geographical extension (Extension géographique)
9. Investment (Investissement)
10. New activity / product (Nouvelle activité / produit)
11. Acquisition project (Projet d'acquisition)
Choice:
```

Listing 1: English translation of the prompt used for LLM-based predictions. The original French class names are written in parenthesis

In this prompt, the placeholder "text" represents the input news article. The LLMs are instructed to analyze the given text and select the relevant category numbers from the provided options. We restrict the models to return up to three category numbers and only if they are explicitly described in the text. By designing this specific prompt, we aim to guide the LLMs to focus on identifying and classifying business events related to the given set of categories. This prompt is the result of several attempts until receiving the best results.

2.1.2. Zero-Shot method leveraging NLI

ZeSTE (Zero-Shot Topic Extraction with Common-Sense Knowledge Graph) use ConceptNet's common-sense knowledge graph and embeddings to generate predictions without relying on training data [1]. By leveraging ConceptNet's vast knowledge graph, it computes the similarity between the encoded article and each term from the graph, selecting the closest match as the predicted category. ZeSTE's predictions are explainable, as they are grounded in the common-sense knowledge encoded in ConceptNet. This allows for better understanding and interpretation of the assigned category for the business events mentioned in the news articles. However, ZeSTE relies on a mapping between the concepts defined in ConceptNet and the targeted business event concepts which is approximative. Along with ZeSTE, we also experimented using other NLI models, namely camembert-base-xnli and xlm-roberta-large-xnli, available via HuggingFace.

2.1.3. Supervised classification method

We have finally fine-tuned a BERT-based classifier using a supervised approach. To train the classifier, we have created a labeled dataset consisting of 224 news articles paired with their corresponding business event labels. These articles have been selected by experts in order to cover the different type of business events we are looking for to extract. We split the dataset into a training set of 179 rows and a test set of 45 rows. During the dataset preparation, we observed a class imbalance issue, where some event classes had a significantly smaller number of instances compared to others. To ensure fair evaluation and prevent the models from being biased towards the majority classes, we performed a stratified split of the dataset.

2.2. Annotating business articles with named entities and topics

In addition to predicting business events, we annotate the news articles with mentions of named entities and topics. To accomplish this, we employ multiple Named Entity Recognition (NER) systems, including spaCy, Flair, and a pre-trained CamemBERT model. We ensemble the named entity annotations with various post-processing techniques to refine the results: a majority vote mechanism to determine the final entity annotations, and a harmonization of the entity type from the sentence to the entire document for the same mention.

We annotate news articles in terms of topics using again Large Language Model and casting the problem as a Zero-Shot classification problem where the topics were 8 pre-defined named classes among: 'Merger - Acquisition', 'CSR (Corporate Social Responsibility)', 'Human Resources', 'Employment', 'International', 'Portfolio', 'Investment', 'Project'.[1]

3. Experiments and Evaluation

3.1. Dataset

The dataset used for evaluation consists of 224 news articles collected from Le Journal Des Entreprises.[2] Each article was manually annotated based on a set of 11 distinct business event classes (Section 2.1.1). The articles were also annotated with 8 named topics (Section 2.2). To

[1]The original topics in French are: 'Fusion - Acquisition', 'RSE', 'Ressource Humaine', 'Emploi', 'International', 'Carnet', 'Investissement', 'Projet'.
[2]https://lejournaldesentreprises.com/

evaluate the performance of our models, we employ the Matthews Correlation Coefficient (MCC) as a metric, which assesses the quality of the binary classifications by taking into account true positives, true negatives, false positives, and false negatives as well as with accuracy scores.

3.2. Results

Table 1 presents the results of each method using a weighted score which considers the class distribution of the dataset. Among the evaluated methods, our fine-tuned BERT model displayed the highest level of performance, achieving a weighted accuracy of 0.9084 and a weighted MCC of 0.6069. This confirms prior findings where LLMs tend to lag behind fine-tuned transformer models for doing text classification [2].

Method	Weighted Accuracy	Weighted MCC
ZeSTE@3	0.6457	0.1509
camembert-base-xnli@3	0.6690	0.2289
xlm-roberta-large-xnli@3	0.6734	0.1754
bloomz-176b	0.8501	0.2891
gpt-3.5-turbo	0.7235	0.3870
gpt-4	0.7636	0.4419
claude-v1	0.7672	0.4423
bert-base-multilingual-cased	**0.9084**	**0.6069**

Table 1
Performances comparison for each prediction method

We also evaluate the different methods on a set of 15 French news articles coming from 3 different sources: Les Echos, Le Figaro and Le Monde. We obtained an overall accuracy of 87% for the prediction of the themes and 94.6% for the extraction of the named entities. However, while GPT-4 and Claude enable to predict business events with 86.6% accuracy, the performance of the BERT-based supervised classifier dropped to only 26.6% demonstrating the lack of generalization of the model when applied to news articles coming from another source.[3]

4. Demo

The demo is available at https://jde-predict.tools.eurecom.fr/. It consists of a backend API, a front-end interface, and it incorporates a caching method for enhanced performance. The backend API serves as the core engine of our system, handling the processing and prediction of business events from news articles. It is made in Python and utilizes the trained models and algorithms discussed in Section 2 to generate predictions based on the URL of an article provided in the request body. The API then returns the predicted business events along with their corresponding probabilities or confidence scores when available.

The front-end interface is created with Next.js and React. It provides a user-friendly web application for interacting with our business event prediction system. Users can input the URL of a news article through a web-based form. The interface displays the text of the news

[3]The details of the evaluation is available at https://github.com/D2KLab/jde-predict/issues/2

article along its metadata and the predicted events (Figure 1). To optimize performance and reduce computational overhead, we employ a caching method in our system with Redis to store previously processed articles and their corresponding predictions, allowing for faster retrieval of results when encountering similar or identical articles in subsequent requests.

Figure 1: Screenshot of the user interface showing the business events predicted by 4 algorithms, as well as the general themes and the named entities extracted from the article.

5. Conclusion and Future Work

We have developed a tool for predicting business events from news articles by leveraging state-of-the-art machine learning and natural language processing techniques. In the demo, we are visualizing four methods based on GPT 4, Claude 1.3, ZeSTE and BERT supervised. All annotations create a larger knowledge graph that can be used for business applications and queried from https://jde-predict.tools.eurecom.fr/kg/.

Acknowledgments

This work has been partially funded by Pairing Technology. We thank Le Journal des Entreprises for providing us with access to their news articles.

References

[1] I. Harrando, R. Troncy, Explainable Zero-Shot Topic Extraction Using a Common-Sense Knowledge Graph, in: 3rd Conference on Language, Data and Knowledge (LDK), 2021.

[2] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, X. Hu, Harnessing the Power of LLMs in Practice: A survey on ChatGPT and Beyond, arXiv:2304.13712, 2023.