GENERATING PRESS REVIEW ABOUT STARTUP INVESTMENTS

V.A. IVANENKO,
Taras Shevchenko National University of Kyiv,
Ukraine, Kyiv
viacheslav.iwanenko@gmail.com

Annotation. *This paper addresses the problem of generating press reviews about startup investments, which involves collecting data from various sources, validating news, Named Entity Recognition, and grouping texts by belonging to a particular investment. This task is very relevant today, as the market for investment in startups is rapidly growing, and the proposed approach can be actively used for the tasks of Business Intelligence in this area. The experimental results prove the effectiveness of the proposed approach.*
Keywords: *text mining, named entity recognition, text classification, innovation, venture capital, press review.*

1. Introduction

Innovative solutions are a guarantee of successful business dealing. It gives companies an advantage over their competitors, allows them to gain more market share, helps to adapt to modern conditions, allows optimizing and speeding up the production process, and as a result to increase profits. Therefore, innovative business models and modern technological solutions are powerful catalysts for business development [1].

Companies that use proven business models to minimize risks and avoid failure cannot be considered startups [2]. Since the implementation of innovations is an extremely risky and complex process, startups(from a financial point of view) are the most optimal and safe solution for implementing new ideas and testing their relevance.

One of the main indicators of a startup's success is the size and regularity of incoming investments. Every year more and more people and organizations start investing. According to Unicorn Nest company [3], in 2020 about $788 billion was invested in startup projects, which is 36% more than the previous year. Today, there are a large number of news sites that specialize in analyzing data about startup investment and venture

capital. This trend indicates the importance of investment market data. For these reasons, the task of automatic monitoring of changes in the investment market becomes more relevant. The most convenient and brief form of this monitoring is a press review of news about a certain round of startup investments. That is why during writing this article the pipeline of generating press reviews about a certain round of startup investments was researched and used.

2. Process workflow of generation press review

Generating a press review is a complex task in the world of Natural Language Processing (NLP) and Text Mining, that includes the stages of gathering text data collection, word processing, preparation of several marked datasets for text classification and named entity recognition (NER), neural network training for text classification and NER tasks, grouping news about one of investment rounds.

2.1 Phase News Mining

Data collection is one of the most important steps in press review generating. The results of all next steps depend on the quality of selected sites and collected texts. In general, this stage can be split into several tasks: selecting news websites, gathering links to posts, and collecting

data from the publication page.

Two popular news sites about venture capital were selected for the experiment: finsmes.com and pulse2.com. All posts from one month were collected from each site. As a result, a dataset of 1,085 unique publications was created. We will use these posts to generate all possible press reviews.

2.2 Phase Text Pre-Processing

To get more accurate results we need to preprocess the received texts before NLP models training begins. Unnecessary information that is not related to your topic may distort the results in the next steps, so all data should be filtered.

Usually, posts on the Internet may contain irrelevant links, symbols, hashtags, advertisements, HTML tags, and more. Often texts can contain specific formats of dates and numbers, there are a lot of prepositions and articles. Obviously, all of these things are not necessary for further

analysis and only increase the size of input data. For this reason, at the preprocessing stage, all unnecessary words and symbols have to be removed [4].

Also, to improve the quality of the text classifier and NER model training, a steaming procedure was used to match the words and to bring them to one morphological form. Researches show that removing stop words and stemming during the pre-processing stage can reduce the size of data by 20-30%, thereby speeding up the analyzers work in the next steps [4].

2.3 News Validation phase

In this step from the collected set of data, we need to select publications that are related to investing in startups. To do this, the binary text classifier, that can detect such news for further analysis, was trained.

Unicorn Nest company [3] provided the dataset for the research, which contains 10,000 financial news, half of them are marked as investment round news.

Using the FlairNLP library [6], a model of a binary text classifier was developed and taught, the architecture of which consists of:

- Token embedding layer
- Bidirectional Long Short-Term Memory Layer [8]

Token embedding layer uses a pre-trained transformer model BERT [10]. This model achieved an f-score of 98% on the test sample of 1,500 publications. Due to the fact that information about investment round may be present in several news items, news that was marked as invalid can be skipped.

2.4 News linking phase

To group news articles about the same round of investments, it is important to understand what startups and investors were mentioned in the text of the financial news. A neural network was trained to recognize names of companies in the news to solve the Named Entity Recognition problem. Dataset of 7,000 news from finsmes.com(that specializes only in news about Venture Capital) was used to train the neural network. This site is convenient because each news article about the round of investment has a "tagged with" list that contains the names of investors and startups that were involved. During dataset tagging, each phrase in news text, that

is equivalent to any name from the “tagged list”, was marked with entity "COMPANY_NAME".

Using the FlairNLP library [6], a model was developed to solve the Named Entity Recognition problem, whose architecture consists of:

- Token embedding layer (RoBERTa [11])
- Bidirectional Long Short-Term Memory Layer [8]
- Conditional Random Field Layer [9]

Based on empirical experiments and training results of other NER models, this architecture has proven to be one of the most effective. For example, it has the highest f-score training biomedical named entity recognition model [8].

Our results on the dataset have an F-score of 81%. This is a good result for this dataset. To improve the f-score, we have to increase the number of texts and make sure that the list "tagged with" contains all possible names of startups and investors used in the text of the news.

As a result of analyzing all news texts by our trained NER model, we can get a press review of the grouped news by proper names that are present in the news or by date of publication. The difference between the dates of publication of several news items for one round of investments should not exceed 3 months. Otherwise, this news, most likely, relates to another round of investments of a certain startup.

3. Conclusions

As a result of the research, a software solution was developed to generate press reviews about investments in startups, and an experiment was conducted to generate a press review of 1,085 news items that were collected in one month from two popular sites about venture capital. Only 821 valid news items were detected during the validation process. As a result of the NER model working and grouping by company name, 276 press reviews were generated, which is a good result and shows the potential of the proposed approach.

Acknowledgments

I would like to thank Unicorn Nest company for accessing the training dataset and assisting in my research.

References.

1. Baden-Fuller, Charles, and Stefan Haefliger. "Business models and technological innovation." *Long range planning* 46.6 (2013): 419-426.
2. Skala, Agnieszka. "The startup as a result of innovative entrepreneurship." *Digital Startups in Transition Economies*. Palgrave Pivot, Cham, 2019. 1-40.
3. Unicorn Nest: https://unicorn-nest.com/
4. Kannan, Subbu, et al. "Preprocessing techniques for text mining." *International Journal of Computer Science & Communication Networks* 5.1 (2014): 7-16.
5. Rana, Mazhar Iqbal, Shehzad Khalid, and Muhammad Usman Akbar. "News classification based on their headlines: A review." *17th IEEE International Multi Topic Conference 2014*. IEEE, 2014.
6. Akbik, Alan, et al. "FLAIR: An easy-to-use framework for state-of-the-art NLP." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 2019.
7. Weber, Leon, et al. "HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition." *Bioinformatics* 37.17 (2021): 2792-2794.
8. Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." IEEE transactions on Signal Processing 45.11 (1997): 2673-2681.
9. Luo, Ling, et al. "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition." Bioinformatics 34.8 (2018): 1381-1388.
10. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
11. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).