Data Preprocessing Algorithm for Decision Support Software Systems

In the context of Developing Socially Oriented Solutions

Artem Khovrat
dept. of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
artem.khovrat@gmail.com

Volodymyr Kobziev
dept. of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
volodymyr.kobziev@nure.ua

***Abstract*—The problem of forecasting social project indicators is becoming increasingly acute in uncertain environment, such as during social unrest. Existing software solutions either provide limited functionality in such situations or require significant investments. The paper is devoted to developing the data preprocessing algorithm for the most popular forecasting approaches, which would allow users to consider not only instability in the economy and society, but also a possible information reaction to the product being created. The findings of this study affirm the efficacy of the proposed approach and pave the path for subsequent real-world project implementations.**

Keywords—forecasting; natural language processing; neural networks with memory; risk management; vector autoregression

I. Introduction

A crucial step in managing projects across various fields involves identifying major risks and finding ways to deal with them effectively. In today's world, this task has become more complicated due to the large amount of data involved. This has led to the development and widespread use of decision support systems (DSS), which help gather information, make predictions, and sometimes offer solutions to problems. However, many DSSs aren't as effective when things are uncertain, especially for projects involving social groups. The emergence of social unrest in recent years, exemplified by the pandemic, as well as subsequent events such as the Russian-Ukrainian war and various localized armed disputes, underscores the urgency of enhancing algorithms that could serve as the foundation for the aforementioned systems. Concurrently, the imperative to develop intelligent solutions for people accentuates this challenge, particularly given the impact of misinformation on human behavior. For instance, the resistance against implementing a more advanced 5G-based internet, crucial for bolstering the dependability of cutting-edge infrastructure solutions, serves as a pertinent illustration [1].

A significant challenge lies in the fact that existed solutions either demand substantial cloud resources or lack the necessary speed, both of which can be pivotal under specific circumstances [2]. Considering these factors, it was decided as part of current study to enhance existing Decision Support System (DSS) algorithms to enhance their efficacy in risk management for socially oriented systems. This modification must not only consider the nature of social changes or audience behavior but also anticipate the potential impact of misinformation designed to discredit the developed solutions.

II. Domain analysis

The first step in finding the required preprocessing algorithm is to analyze the most popular DSSs. After reviewing feedback spanning the last three years and examining the official websites of the identified projects, the following characteristics have been observed [3]:

- prognostic algorithms primarily rely on neural networks and/or autoregression models;
- there is a limited consideration of the risks associated with the behaviors of the intended users of the developed systems (for checking this sentence was finding reviews related to COVID-19 pandemic);
- the target projects' information environments are developed is often overlooked, despite its critical importance in urban solution development.

The defined categories of fundamental predictive algorithms are relatively broad and necessitate further refinement. Specifically, in the case of neural networks CNN, RNN, LSTM and RCNN models can be used.

Among autoregressive models can be included:

- distributed lag autoregression and seasonal autoregression;
- autoregression of moving average and integrated moving average.

Further exploration of forecasting economic indicators through international papers enables the narrowing down of basic forecasting algorithms to a combined hybrid neural network and autoregression of an integrated moving average [4]. Despite being slower, these models offer superior accuracy compared to other options. Additionally, they support parallelization and do not mandate substantial hardware resources.

To address potential risks inherent in developing and deploying socially oriented software projects, it's crucial to establish a comprehensive set of key aggregated indicators:

- Social Shift Profile: quantifying the uncertainty associated with transitions in societal conditions into a numerical format;
- Target Audience Profile: summarizing the behaviours of the most influential individuals or groups targeted by the project;
- Business Environment Profile: describing the market-specific conditions that affect the project's implementation;
- Information Environment: reflecting the degree of influence exerted by fabricated information on the project's intended concept.

The following indicators were derived from analysis of contemporary scientific literature and expert assessments conducted among a 100 professionals including sociologists, engineers and managers from Kharkiv, Lviv, Dnipro, Kyiv, Lisbon, and Krakow.

Upon investigating the concept of social unrests, also referred to as "social disasters," it became evident that the most influential subindicators include the prevalence of the shift's source, its duration (from the moment of initial information emergence), field-specific characteristics, and the intensity level. The first two indicators are inherently objective numerical variables, while the latter two reflect subjective perceptions of the shift, necessitating additional algorithms for utilization in forecasting.

The profile of the target audience can be delineated by considering the audience size, the market's susceptibility to known neoclassical economic paradoxes, the level of trust, and societal characteristics. Similar to the first case, this entails a blend of objective numerical indicators and subjective textual indicators pertinent to the target project.

The profile of the business environment revolves around businesses' responses to both the proposed project's implementation and specific social disasters. To address this, emphasis is placed on indicators of economic stability (both globally and locally) and businesses' preparedness for emergencies. While the latter indicator facilitates aligning the objective assessment of a environment’s financial stability with the subjective perceptions of its internal stakeholders.

The information environment pertains to the intensity of false news dissemination concerning the project's topic, technological reforms, and related domains of knowledge. Assessing the veracity of information is not a novel field, as previously mentioned. Research conducted by various groups of European scientists on fake text news demonstrated that machine learning algorithms, constructed using neural networks, transformers, or autoencoders, necessitate substantial data volumes to attain accuracy exceeding 90% [5]. Nonetheless, this challenge can be addressed by employing a balanced dataset.

III. Overview of Selected Base Models

As previously mentioned, this study focuses on two fundamental algorithms: RCNN and vector autoregression of the moving average. The vectorizing of the latter is crucial for simultaneous processing of multiple indicators.

While a simple CNN model considers the environment of each element by passing through filters, the nature of the proposed indicators necessitates an understanding of a longer time span without a drastic leap into the future. Consequently, the pertinent context may lie beyond the scope of the CNN model's filters. To circumvent this issue, a combination of RNN and CNN was chosen. To comprehensively incorporate context, a decision was made to utilize a bidirectional recurrent neural network with support for long- and short-term memory, instead of a simple RNN architecture. This approach relies on hyperbolic tangents and sigmoids to address issues of explosive and vanishing gradients, thereby constraining the range of resultant values. Consequently, the RCNN architecture can be depicted as illustrated in Figure 1.

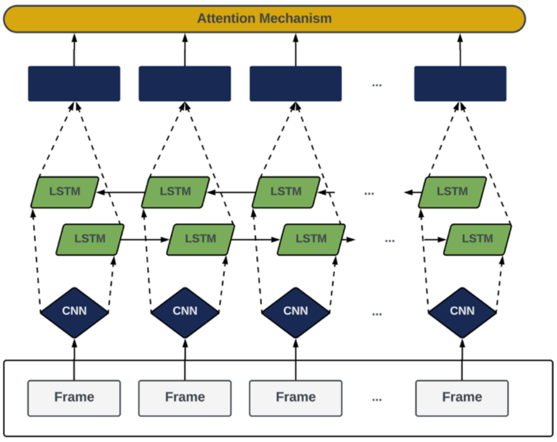

Fig. 1. Scheme of the chosen RCNN architecture.

Following cross-validation testing, the optimal key hyperparameters were identified as follows:

- the kernel size is set at 4
- the stride size is defined as 1;
- the option to add non-significant zeros based on the step size will not be applied;
- the displacement parameter will not be utilized;
- the filter dimension is established as 5x5x3 (the last value determined by the number of target indicators).

In traditional vector autoregression (VAR) models, only static variables are typically predicted. However, to incorporate exogenous indicators, researchers have adopted error correction modification (EC). A similar adjustment becomes imperative when multiple endogenous variables share a common stochastic trend. This situation applies to the specific problem being examined. As for the preprocessing algorithm, special attention should be paid to the sequence of steps for converting text into a numeric representation:

- removing non-significant lexically loaded words from the textual description;

- establishing a dictionary of essential lemmas and analyzing the frequency distribution of each word form;
- assessing the polarity of individual words and adjusting frequency values accordingly;
- aggregating and subsequently normalizing the gathered data within a scale from 0 to 1.

To enhance the efficiency of the target algorithm, was decided to use MapReduce technology, involving the segmentation of the original dataset into distinct nodes. Central to this approach are the mapping and reduction functions. While various implementations exist, those based on Spark and Hadoop are widely favored. In this study, the latter option was selected.

IV. Experimental environment

In the framework of the current work, the experimental environment will mean the combination of the data and efficiency function.

Datasets for examining the falsification of news were curated based on the processing of news articles pertaining to the deployment of e-ticketing in Kharkiv and the adoption of the "smart city" concept in Kyiv. Additionally, data for predictive analysis with project objectives were semi-automatically generated. Target indicators encompassed the dynamics of costs and revenues, the level of engagement of the target audience, and performance metrics of tasks.

After expert evaluation, the following indicators were designated as crucial performance criteria to build up the efficiency function:

- “accuracy”, assigned an importance factor of 16;
- time-saving for the target algorithm assigned an importance factor of 8;
- the required data to attain an “accuracy” exceeding 80%, assigned an importance factor of 4.

Since the problem is prediction rather than classification, the accuracy will be determined using the normalised inverse root mean square error. The saving of processing time will also consider the parallelisation proposed above, to level the loss that accompanies the use of the data reprocessing algorithm. Weighting coefficients for linear additive convolution are derived from these importance factors.

V. Results of the experiment

The simple EC-VARIMA algorithm emerges as the quickest, with the modified version following closely. This speed is attributed to parallelization across both the model and refinement steps. The detailed results for time-saving are following: simple RCNN – 0 s, modified RCNN – 5.8 s, simple EC-VARIMA – 13.8 s, modified EC-VARIMA – 10.1 s. In contrast, accuracy varies among algorithms, with the modified RCNN exhibiting the highest precision. However, it's worth noting the inherent instability of basic algorithms when external indicators are disregarded. The detailed results for accuracy are following: simple RCNN – 72%, modified RCNN – 95%, simple EC-VARIMA – 62%, modified EC-VARIMA – 93%. The ultimate metric to consider is data volume savings. Notably, the two baseline models failed to reach the desired minimum accuracy threshold. Consequently, the saving value for these algorithms is recorded as 0. Conversely, for the modified RCNN, the minimum acceptable value stands at 50,000 elements, while for the modified EC-VARIMA, it is set at 100,000 elements.

Based on the results obtained, the value of the linear additive convolution with weighting coefficients was calculated. For Simple RCNN, the value was 0.41, for Simple EC-VARIMA – 0.64, for Modified RCNN – 0.79, and for Modified EC-VARIMA – 0.87.

VI. Conclusions

The current paper aimed to modify the existing basic algorithms embedded in decision support systems for socially orientated systems under nondeterministic conditions.

For this purpose, the subject area related to DSSs was analysed and the specific approach for preprocessing data was built. Given the values obtained for linear additive convolution, the proposed approach to the parallelisation and reprocessing of external data gives the desired result, increasing the efficiency of the application of simple models. At the same time, the forecast accuracy is high for both modified algorithms. However, due to its simplicity, the modified EC-VARIMA is more effective given the selected set of indicators. The efficiency value allows to proceed to the next stage of testing the proposed solution in the real environment of the development of socially orientated urban projects and considering the problems of determining the fact of falsification of information and other external indicators.

References

[1] E. Flaherty, T. Sturm, E. Farries, “The conspiracy of Covid-19 and 5G: Spatial analysis fallacies in the age of data democratization”, Soc Sci Med. vol. 293, no. 114546, Jan. 2022. Accessed: May 6, 2024. [Online]. doi: 10.1016/j.socscimed.2021.114546.

[2] U. A. Butt, R. Amin, M. Mehmood, H. Aldabbas, M. T. Alharbi, N. Albaqami, “Cloud Security Threats and Solutions: A Survey”, Wireless Personal Communications, vol. 128, pp. 387-413, Sep. 2022. Accessed: May 6, 2024. [Online]. doi: 10.1007/s11277-022-09960-z.

[3] “IT Risk Management Software”, G2. https://www.g2.com/categories/it-risk-management (accessed May 7 2024).

[4] S. Yakovlev, A. Khovrat, V. Kobziev, “Using Parallelized Neural Networks to Detect Falsified Audio Information in Socially Oriented Systems”, in: Proceedings of the International Scientific Conference "Information Technology and Implementation", IT&I ’23, Kyiv, Ukraine, pp. 220–238, Jan. 2024 Accessed: May 6, 2024. [Online]. https://ceur-ws.org/Vol-3624/Paper_19.pdf.

[5] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, J. Vilares, “Sentiment Analysis for Fake News Detection”, Electronics, vol. 10 (11) no. 1348, Jun. 2021. Accessed: May 6, 2024. [Online]. doi: 10.3390/electronics1011134.