ISSN 2224-087X. Електроніка та інформаційні технології. 2020. Випуск 14. С. 3–21
Electronics and information technologies. 2020. Issue 14. P. 3–21

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В НАУКОВИХ ДОСЛІДЖЕННЯХ

УДК 004.67:004.89 DOI: https://doi.org/10.30970/eli.14.1

MODELING COVID-19 SPREAD AND ITS IMPACT ON STOCK MARKET USING DIFFERENT TYPES OF DATA

Bohdan M. Pavlyshenko

Ivan Franko National University of Lviv,
50 Drahomanov St., 79005, Lviv, Ukraine
b.pavlyshenko@gmail.com

The paper studies different regression approaches for modeling COVID-19 spread and its impact on the stock market. The logistic curve model was used with Bayesian regression for predictive analytics of the COVID-19 spread. Bayesian approach makes it possible to use informative prior distributions formed by experts that allows considering the results as a compromise between historical data and expert opinion. The obtained results show that different crises with different reasons have different impact on the same stocks. It is important to analyze their impact separately. Bayesian inference makes it possible to analyze the uncertainty of crisis impacts. The impact of COVID-19 on the stock market using time series of visits on Wikipedia pages related to coronavirus was studied. Regression approach for modeling COVID-19 crises and other crises impact on stock market were investigated. The analysis of semantic structure of tweets related to coronavirus using graph theory and frequent itemsets and association rules theory was carried out.

Keywords: coronavirus, COVID-19, Bayesian regression, stock market, predictive analytics.

1. Introduction

At present time, there are different methods, approaches and data sets for modeling the COVID-19 spread [1, 2, 3, 4, 5, 6]. The approach based on Bayesian inference allows us to receive a posterior distribution of model parameters using conditional likelihood and prior distribution. In the Bayesian inference, we can use informative prior distributions which can be set up by an expert. So, the result can be considered as a compromise between historical data and expert opinion. It is important in the cases when we have a small number of historical data. In [7, 8, 9, 10] we consider different approaches of using Bayesian regression. For solving Bayesian models, numerical Monte-Carlo methods are used. Gibbs and Hamiltonian sampling are the popular methods of finding posterior distributions for the parameters of probabilistic mode [11, 12, 13]. Bayesian inference makes it possible to obtain probability density functions for model parameters and estimate the uncertainty that is important in risk assessment analytics. Different problems caused by COVID-19 are being widely considered in social networks especially Twitter. So, the analysis of tweet trends can reveal the semantic structure of users' opinions related to COVID-19.

In this paper, we consider the use of Bayesian regression for modeling the COVID-19 spread. We also consider the impact of COVID-19 on the stock market using time series of

visits on Wikipedia pages related to coronavirus. We study a regression approach for modeling the impact of different crises on the stock market. The usage of frequent itemsets and association rules theory for analyzing tweet sets was considered.

2 Bayesian Model for COVID-19 Spread Prediction

For the predictive analytics of the COVID-19 spread, we used a logistic curve model. Such model is very popular nowadays. To estimate model parameters, we used Bayesian regression [11, 12, 13]. A logistic curve model with Bayesian regression approach can be written as follows:

$$n \sim N(\mu, \sigma),$$
$$\mu = \frac{\alpha}{1+\exp(-\beta(t-t_0))} \cdot 10^5,$$
$$t = t_{weeks}(Date - Date_0),$$

where $Date_0$ is a start day for observations in the historical data set, it is measured in weeks, α parameter denotes maximum cases of coronavirus, β parameter is an empirical coefficient which denotes the rate of coronavirus spread. The data for the analysis were taken from [2]. For Bayesian inference calculations, we used a 'pystan' package for Stan platform for statistical modeling [13]. Figure 1 shows the box plots for calculated β parameters of the coronavirus spread model for different countries.

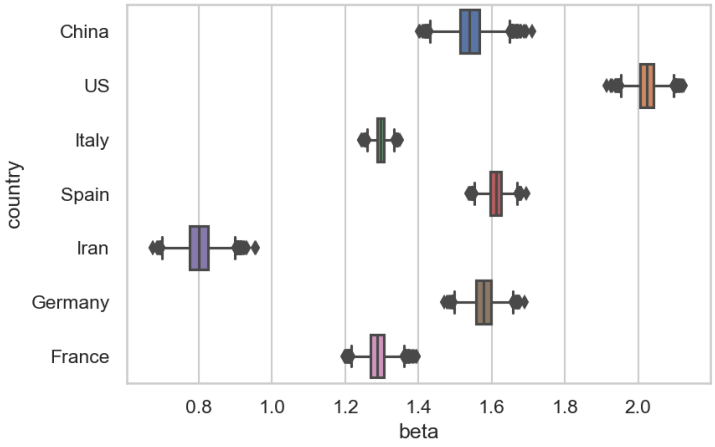

Figure 1: Box plots for beta coefficients of coronavirus spread model for different countries.

Fig. 2, 3 show the predictions for coronavirus spread cases using current historical data. In practical analytics, it is important to find the maximum of coronavirus cases per day, this point means the estimated half time of coronavirus spread in the region under investigation. New historical data will correct the distributions for model parameters and forecasting results. The results show that the Bayesian regression model using logistic curve can be effectively

used for predictive analytics of the COVID-19 spread. In Bayesian regression approach, we can take into account expert opinions via information prior distribution, so the results can be treated as a compromise between historical data and expert opinion that is important in the case of small amount of historical data or in the case of a non- stationary process. It is important to mention that new data and expert prior distribution for model can essentially correct previously received results.

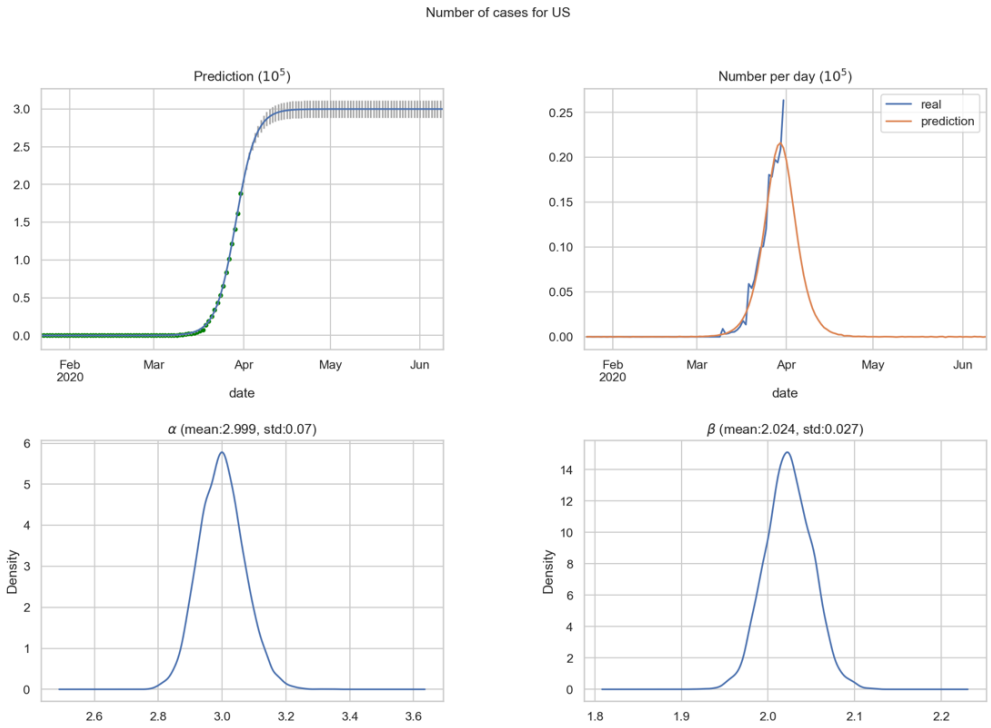

Figure 2: Modeling of COVID-19 spread for USA.

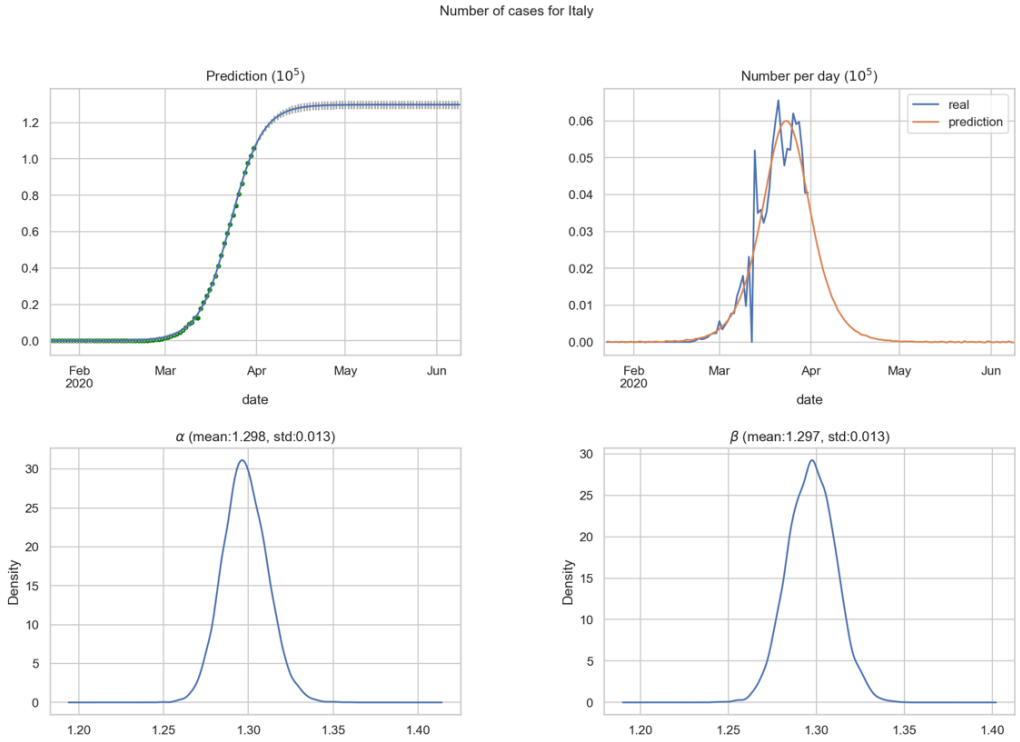

Figure 3: Modeling of COVID-19 spread for Italy.

Using Wikipedia Pages Visits Time Series for Prediction Stock Market Movements

Let us consider the influence of the COVID-19 spread on the stock market movement. The influence of impact factors can be described by alternative data, such as characteristics of search trends, users' activity in social networks, etc. Figure 4 shows the stock market indexes and some stock price time series in the period of the biggest impact of COVID-19 on the stock market. In this study, as alternative data, we consider the time series of visits to Wikipedia pages which are related to COVID-19.

The figure 5 shows the time series of numbers of visits to Wikipedia pages. For our analysis, we consider the time period of ['2020-02-15','2020-05-01']. For Bayesian regression, we used Stan platform for statistical modeling [13]. As features, we used z-scores of the time series of Wikipedia page visit numbers. As a target variable, we used z-scores of S&P-500 index. We applied the constraints that linear regression coefficients cannot be positive. Figure 6 shows the mean values and 0.01, 0.99 quantiles of probability density function of predictions for S&P-500 index. Figure 7 shows the boxplots for PDF of linear regression coefficients. Figure 8 shows variation coefficient which is equal to the ratio between the standard deviation and absolute mean values of regression coefficients. These coefficients describe the uncertainty of regression features. The obtained results show that different features have different impact and uncertainty with respect to the target variable. The most impactful and the least volatile among the considered features was the feature of the number of visits to the Wikipedia page about the vaccine.

Figure 4: Stock market indexes and stock price time series.

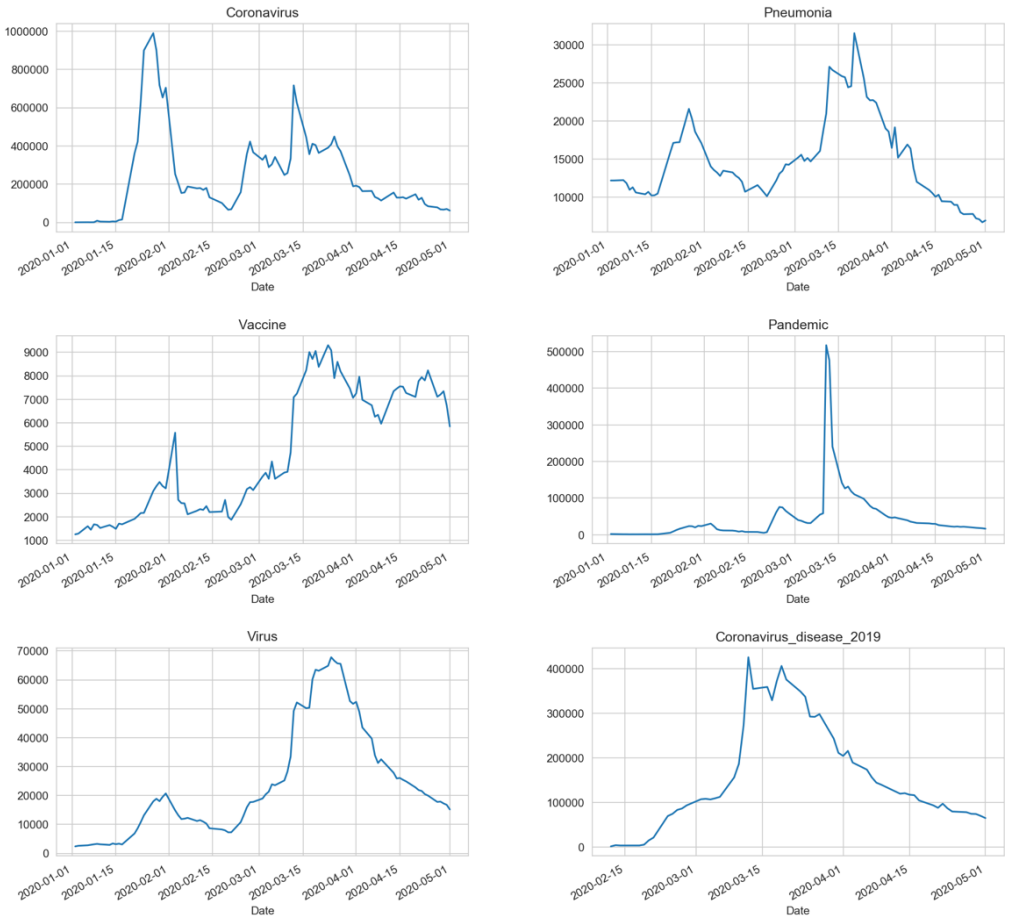

Figure 5: Time series of numbers of visits to Wikipedia pages.

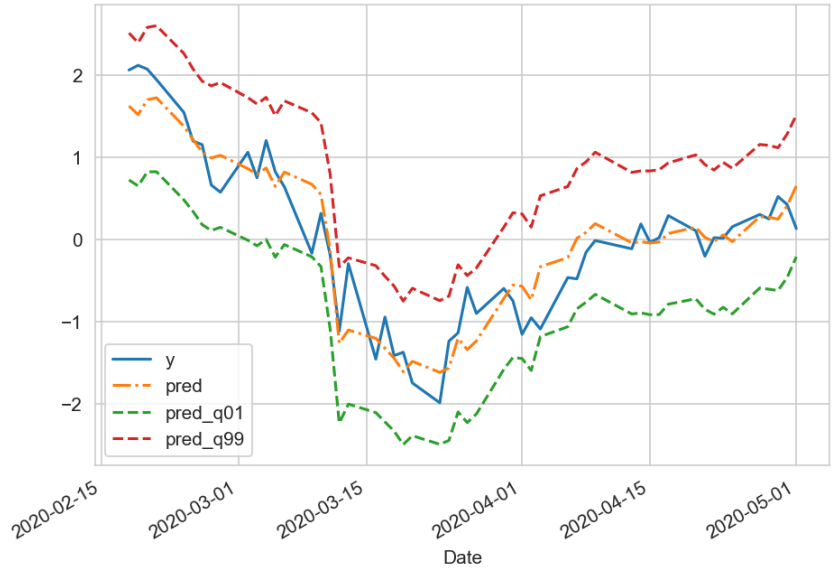

Figure 6: Mean values and 0.01, 0.99 quantiles of probability density function of predictions for S&P-500 index.

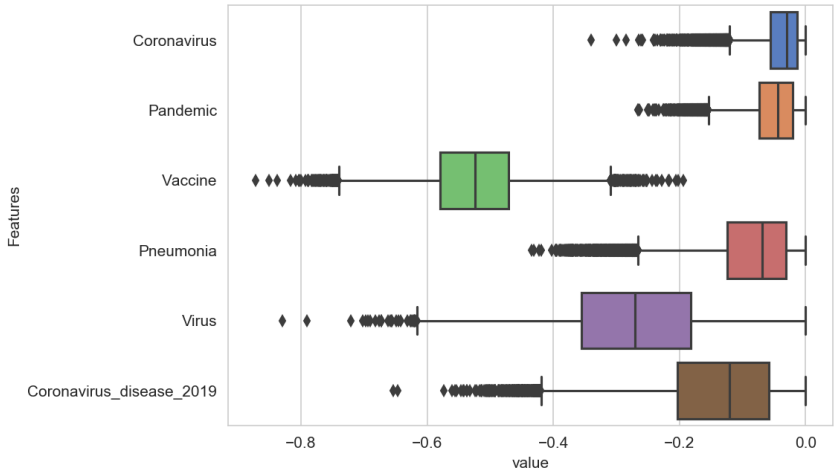

Figure 7: Boxplots for PDF of linear regression coefficients.

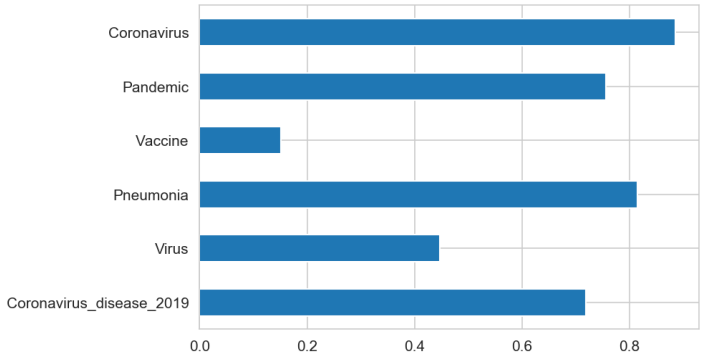

Figure 8: Variation coefficient for features.

Regression Approach for Modeling the Impact of Different Crises on the Stock **Market**

The coronavirus outbreak has a huge impact on the stock market. It is very important, e.g. for forming stable portfolios, to understand how different crises impact stock prices and the stock market as a whole. We are going to consider the impact of coronavirus crisis on the stock market and compare it to the crisis of 2008 and market downturn of 2018. For this, we can use the regression approach using ordinary least squares (OLS) regression and Bayesian regression. Bayesian inference makes it possible to obtain probability density functions for coefficients of the factors under investigation and estimate the uncertainty that is important in the risk assessment analytics. In Bayesian regression approach, we can analyze extreme target variable values using non-Gaussian distributions with fat tails. We took the following time periods for each of crises - crisis_2008: [2008-01-01,2009-01-31], down_turn_2018: [2018- 10-01,2019-01-03], coronavirus: [2020-02-18,2020-03-25]. For each of the above mentioned crises, we created a regression variable which is equal to 1 in the crisis time period and 0 in other cases. Figure 9 shows the time series for S&P500 composite index.

Figure 9: Time series for S&P500 composite index.

As a target variable, we consider the daily price return. Knowing the daily price return, changes in crises periods, one can estimate the ability of investors to understand trends and recalculate portfolios. These results were received using Bayesian inference. For Bayesian inference calculations, we used Python ’pystan’ package for Stan platform for statistical modeling [13]. Figure 10 shows the box plots of impact weights of each crisis on S&P500 composite index. The wider box for coronavirus weight can be caused by shorter time period of investigation comparing with other crises and consequently larger uncertainty.

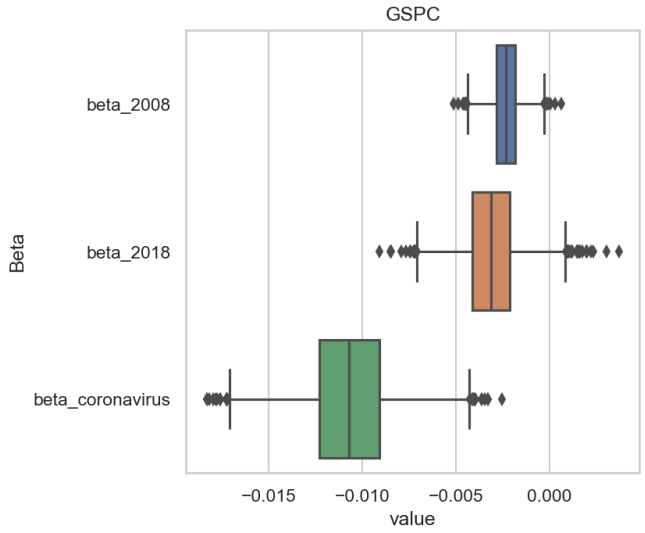

Figure 10: Box plots of impact weights of each crisis on S&P composite index.

For our investigations, we took a random set of tickers from S&P list. Figure 11 shows top negative price returns in coronavirus crises. Figure 12 shows the tickers with positive price return in coronavirus crisis. Figure 13 shows the weights for different crises for arbitrarily chosen stocks. We calculated the distributions for crises weights using Bayesian inference. Figure 14 shows the box plots for crisis weights for different stocks.

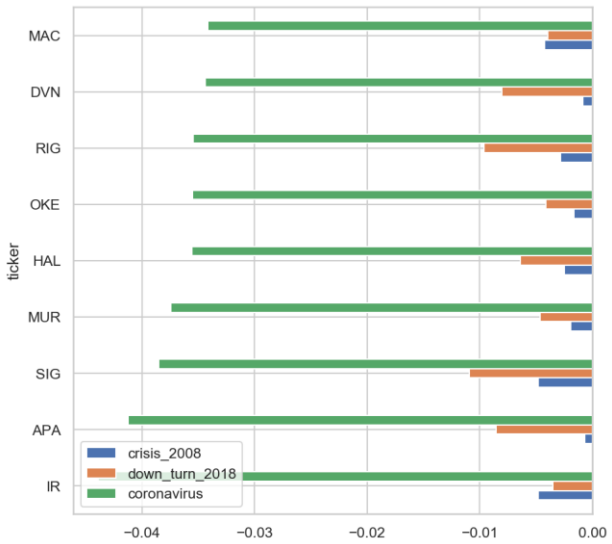

Figure 11: Top negative price returns in COVID-19 crises.

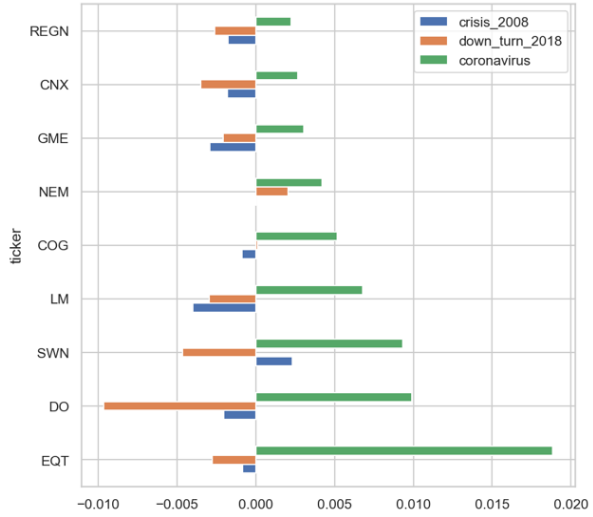

Figure 12: Tickers with positive price returns in COVID-19 crises.

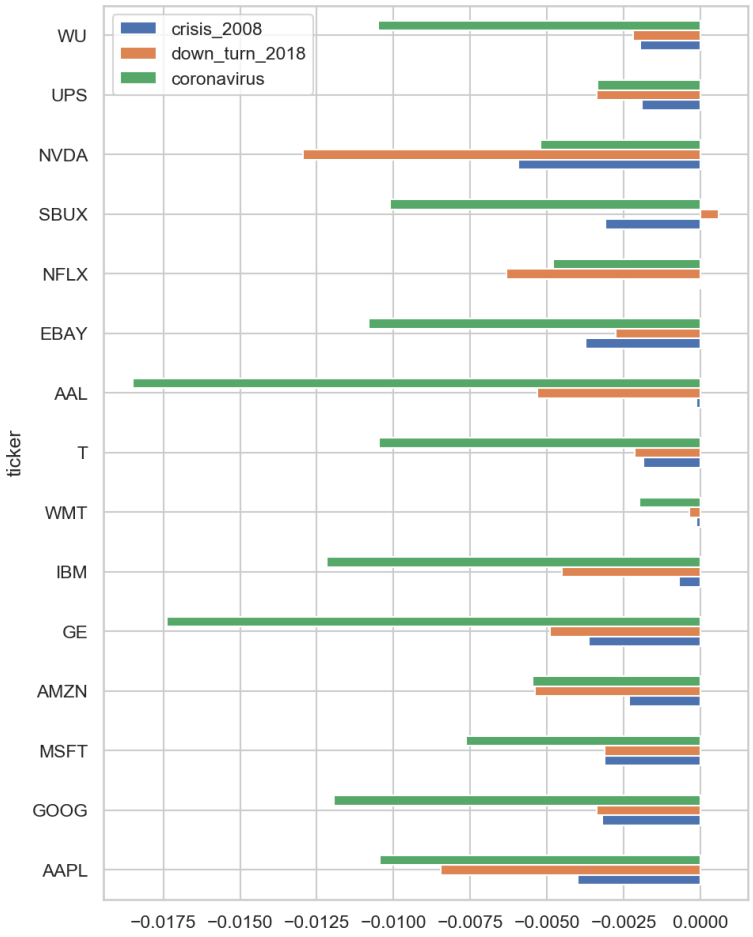

Figure 13: The weights for different crises for arbitrarily chosen stocks.

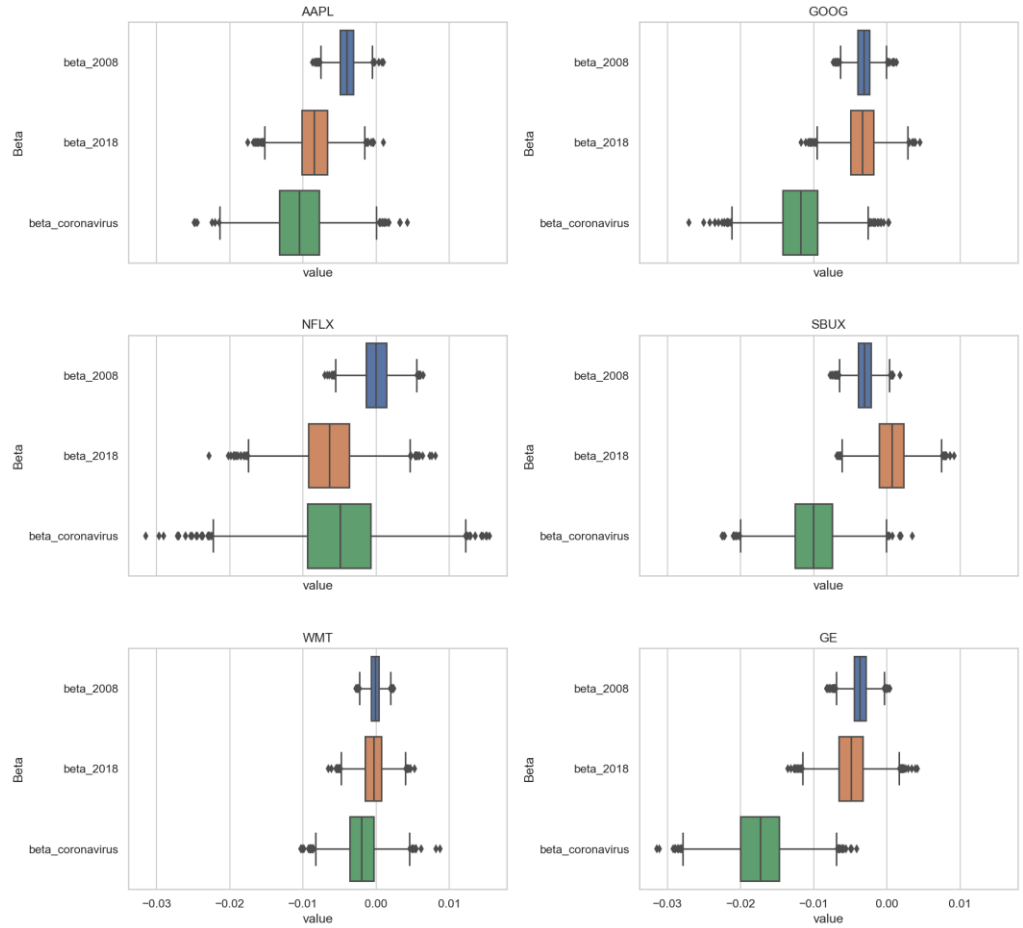

Figure 14: Box plots for crisis weights for different stocks.

Analysis of Tweets Related to COVID-19

Let us consider the approaches in the analysis of the tweet set related to COVID-19. Users' connections can be represented by a graph where vertices represent users and edges stand for their connections. Using graph mining algorithms, we can detect users' communities and find the top lists of users by different vertice scores as Hub, Authority, PageRank, Betweenness scores. For community detection, we used a Community Walktrap Algorithm [15] from 'igraph' package [16]. For the visualization, we used Fruchterman-Reingold [17] from this package. The tweets with 'coronavirus' keyword loaded in January, 2020 were used for analysis. Figure 15 shows revealed users' community for the subset of tweets.

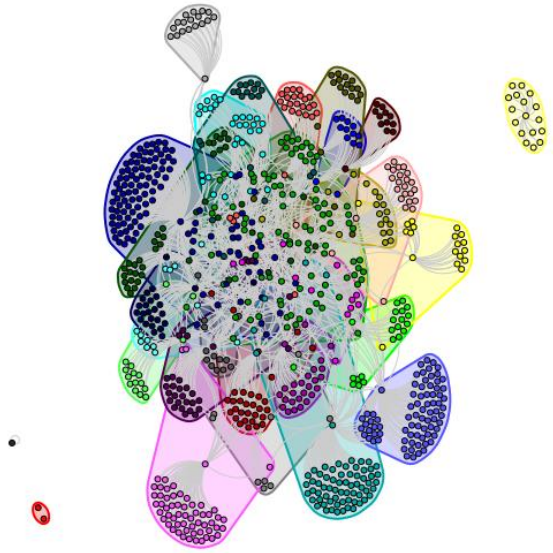

Figure 15: Graph of users' connections and users' communities.

In tweet trends, the different users' communities with different points of view are are formed. The analysis of trends and points of view of different communities reveals structure of communication trends in tweets related to COVID-19. In the analysis of semistructured data, e.g. texts, the theory of frequent itemsets and association rules can be used [18, 19, 20, 21, 22, 23, 24, 25]. Using frequent itemsets and association rules, we can find a semantic structure in specified semantic fields of lexemes. Different methods and approaches of using frequent itemsets and association rules we considered in [26, 27, 28]. For our analysis, we created a thematic field which consists of important keywords related to COVID-19. Figure 16 shows the text frequencies for chosen keywords. Some keywords denote semantic fields which are a set of semantically related words, e.g. semantic field 'fear' comprises such words as 'fear', 'worried', 'panic', 'apocalyptic', etc. Figures 17-19 show the semantic frequent sets found in the tweet set. Figures 20 shows found grouped association rules related to coronavirus. Found frequent itemsets and association rules reveal a semantic structure of tweets related to COVID-19. The quantitative characteristics of frequent itemsets and association rules, e.g. value of support, can be used as features in the predictive analysis.

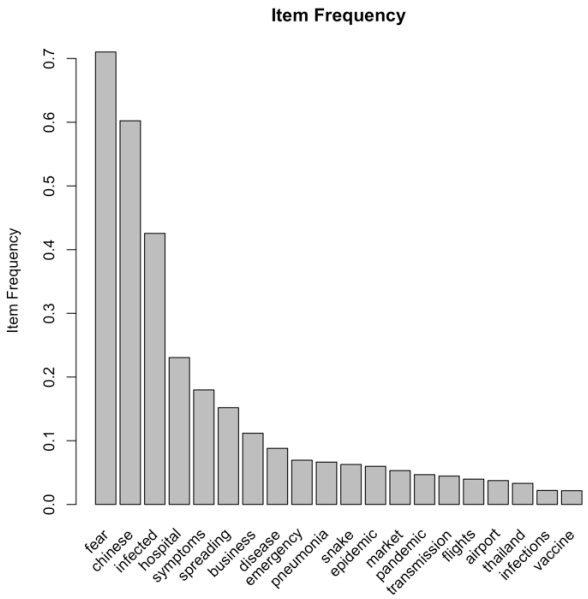

Figure 16: Frequencies for keywords.

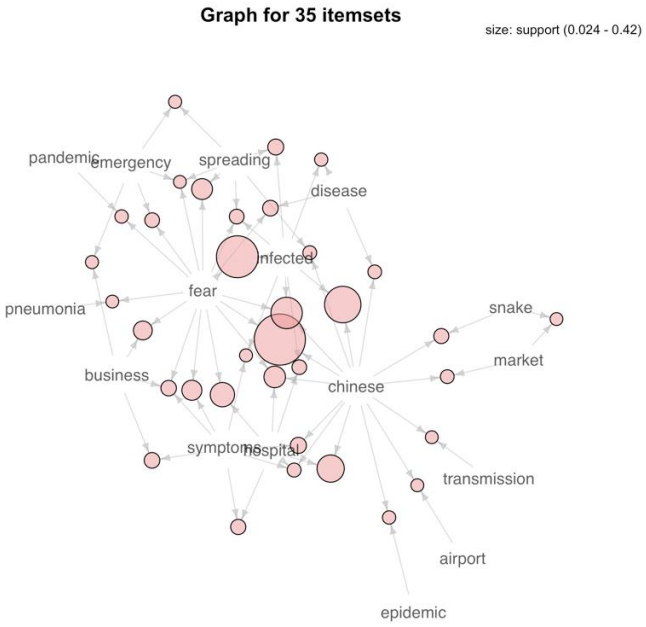

Figure 17: Frequent itemsets.

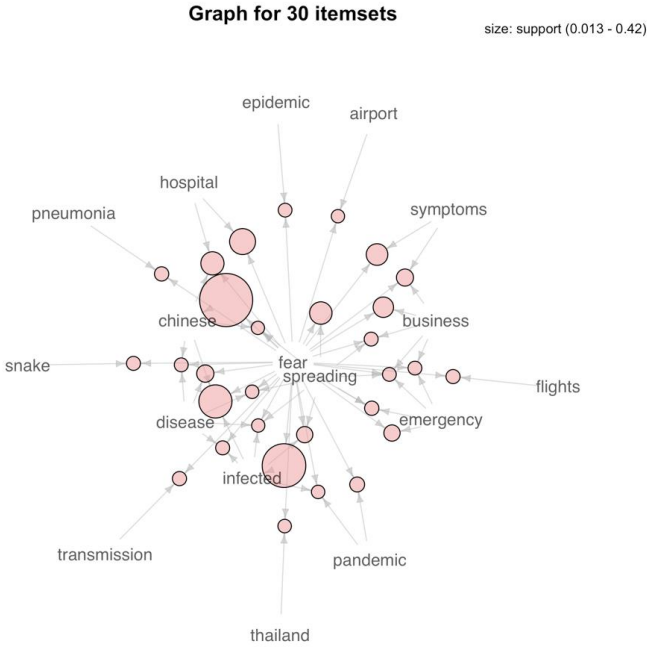

Figure 18: Frequent itemsets with the semantic field ’fear’.

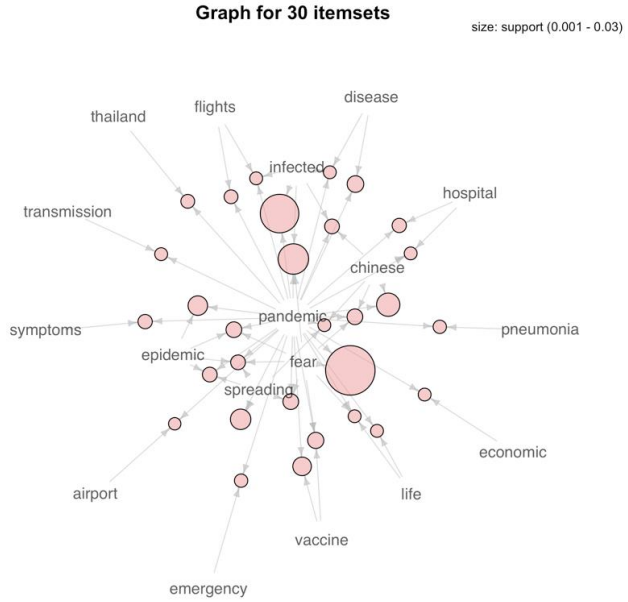

Figure 19: Frequent itemsets with the keyword ’pandemic’.

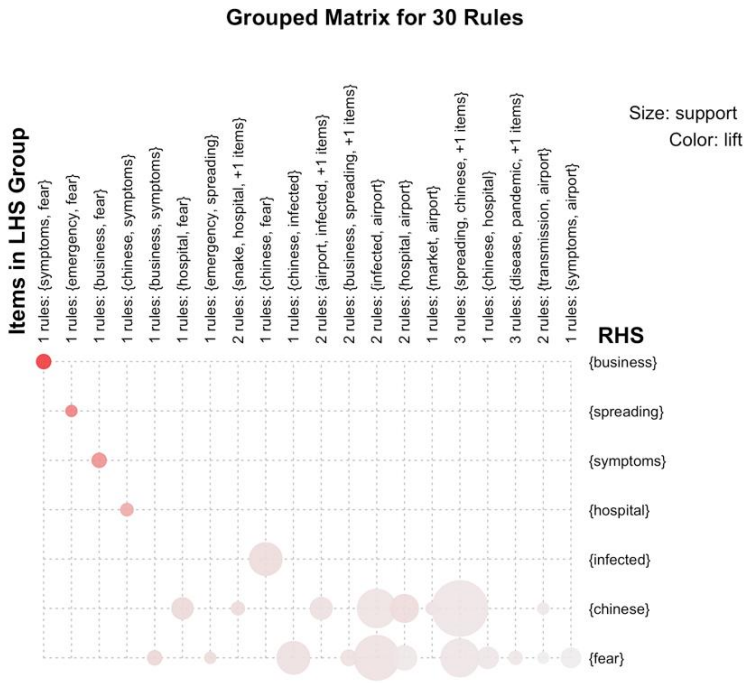

Figure 20: Grouped association rules.

Conclusion

The received results show that the logistic curve model can be used with Bayesian regression for the predictive analytics of the COVID-19 spread. Such a model can be effective when the exponential growth of coronavirus confirmed cases takes place. In practical analytics, it is important to find the maximum of coronavirus cases per day, this point means the estimated half time of coronavirus spread in the region under investigation. New historical data will correct the distributions for model parameters and forecasting results. For conducting the modeling of COVID-19, we developed the 'Bayesian Model for COVID-19 spread Prediction' Python package, which can be loaded at [14] for free use. In Bayesian regression approach, we can take into account expert opinions via information prior distribution, so the results can be treated as a compromise between the historical data and expert opinion that is important in the case of small amount of historical data or in the case of a non-stationary process. It is important to mention that new data and expert prior distribution for model can essentially correct previously received results. The impact of COVID-19 on the stock market using time series of visits on Wikipedia pages related to coronavirus was studied. The obtained results show that different features have different impact and uncertainty with respect to the target variable. The most impactful and the least volatile among the considered features was the feature of the number of visits to the Wikipedia page about the vaccine. The obtained results show that different crises with different reasons have different impact on the same stocks. Bayesian

inference makes it possible to analyze the uncertainty of crisis impacts. The uncertainty of crisis impact weights can be measured as a standard deviation for weight probability density functions. The uncertainty of coronavirus crisis is larger than other crises that can be caused by shorter analysis time. Knowing the uncertainty, allows risk assessment for portfolios and other financial and business processes. Using the graph theory, the users' communities and influencers can be revealed given vertices quantitative characteristics. The analysis of tweets related to COVID-19 was carried out using frequent itemsets and association rules. Found frequent itemsets and association rules reveal the semantic structure of tweets related to COVID-19. The quantitative characteristics of frequent itemsets and association rules, e.g. value of support, can be used as features in the predictive analysis.

REFERENCES

1. IHME COVID, Christopher JL Murray, et al. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1.
2. COVID19 Global Forecasting (Week 2). Kaggle.Com. https://www.kaggle.com/c/ covid19-global-forecasting-week-2.
3. CSSE COVID-19 Dataset. GitHub.Com. https://github.com/CSSEGISandData/ COVID-19/tree/master/csse_covid_19_data.
4. Coronavirus (Covid-19) Data in the United States. GitHub.Com. https://github. com/nytimes/covid-19-data.
5. COVID-19 reports. https://www.imperial.ac.uk/ mrc-global-infectious-disease-analysis/covid-19/.
6. COVID-19 Kaggle community contributions. Kaggle.Com. https://www.kaggle. com/covid-19-contributions.
7. *B.M. Pavlyshenko*. Bitcoin Price Predictive Modeling Using Expert Correction. In 2019 XIth International Scientific and Practical Conference on Electronics and Infor- mation Technologies (ELIT),September 16 – 18, 2019 Lviv, Ukraine, pages 163–167, 2019.
8. *Bohdan M. Pavlyshenko*. Linear, machine learning and probabilistic approaches for time series analysis. In Data Stream Mining & Processing (DSMP), IEEE First International Conference, pages 377–381, 2016.
9. *Bohdan Pavlyshenko.* Machine learning, linear and Bayesian models for logistic re-gression in failure detection problems. In Big Data (Big Data), 2016 IEEE Interna- tional Conference on, IEEE, Washington D.C., pages 2046–2050, 2016.
10. *B. Pavlyshenko.* Using Bayesian Regression for Stacking Time Series Predictive Mod-els. In 2020 IEEE Third International Conference on Data Stream Mining & Pro- cessing (DSMP), pages 305–309, 2020.
11. *John Kruschke*. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press, 2014.
12. *Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin.* Bayesian data analysis. Chapman and Hall/CRC, 2013.
13. *Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell.* Stan: A probabilistic programming language. Journal of statistical software, 76(1), 2017.
14. Bayesian Model for COVID-19 Spread Prediction. GitHub.Com. https://github. com/pavlyshenko/covid19.

15. *Pascal Pons and Matthieu Latapy*. Computing communities in large networks using random walks. In International symposium on computer and information sciences, pages 284–293. Springer, 2005.
16. *Gabor Csardi, Tamas Nepusz, et al.* The igraph software package for complex network research. InterJournal, complex systems, 1695(5):1–9, 2006.
17. *Thomas M.J., Fruchterman and Edward M Reingold.* Graph drawing by force-directed placement. Software: Practice and experience, 21(11):1129–1164, 1991.
18. *Rakesh Agrawal, Ramakrishnan Srikant, et al*. Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB, volume 1215, pages 487– 499, 1994.
19. *Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, A Inkeri Verkamo, et al*. Fast discovery of association rules. Advances in knowledge discovery and data mining, 12(1):307–328, 1996.
20. *Chun-Kit Chui, Ben Kao, and Edward Hung*. Mining frequent itemsets from un- certain data. In Pacific-Asia Conference on knowledge discovery and data mining, pages 47–58. Springer, 2007.
21. *Karam Gouda and Mohammed Javeed Zaki*. Efficiently mining maximal frequent itemsets. In Proceedings 2001 IEEE International Conference on Data Mining, pages 163–170. IEEE, 2001.
22. *Ramakrishnan Srikant, Quoc Vu, and Rakesh Agrawal*. Mining association rules with item constraints. In Kdd, volume 97, pages 67–73, 1997.
23. *Mika Klemettinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, and A Inkeri Verkamo*. Finding interesting rules from large sets of discovered association rules. In Proceedings of the third international conference on Information and knowledge management, pages 401–407, 1994.
24. *Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal*. Discovering frequent closed itemsets for association rules. In International Conference on Database Theory, pages 398–416. Springer, 1999.
25. *Sergey Brin, Rajeev Motwani, and Craig Silverstein*. Beyond market baskets: Generalizing association rules to correlations. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data, pages 265–276, 1997.
26. *Bohdan Pavlyshenko*. Data Mining of the Concept" End of the World" in Twitter Microblogs. arXiv preprint arXiv:1302.2131, 2013.
27. *Bohdan M. Pavlyshenko*. Forecasting of Events by Tweets Data Mining. Electronics and information technologies, (10):71–85, 2018.
28. *Bohdan M. Pavlyshenko*. Can Twitter Predict Royal Baby's Name ? Electronics and information technologies, (11):52–60, 2019.

МОДЕЛЮВАННЯ ПОШИРЕННЯ COVID-19 ТА ЙОГО ВПЛИВУ НА ФОНДОВИЙ РИНОК ІЗ ВИКОРИСТАННЯМ ДАНИХ РІЗНИХ ТИПІВ

Б. М. Павлишенко

Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 Львів, Україна
b.pavlyshenko@gmail.com

У роботі розглянуто моделювання поширення COVID-19 на основі історичних даних. Для прогнозної аналітики поширення COVID-19 використано модель логістичної кривої. Для оцінки параметрів моделі використано байєсівську регресію. Цей підхід дозволяє отримати постеріорний розподіл ймовірності для параметрів моделі. У байєсівському виведенні можна використовувати задані експертом інформативні апріорні розподіли і результат можна розглядати як компроміс між історичними даними та думкою експерта. Це важливо у тих випадках, коли є невелика кількість історичних даних. Імовірнісний підхід дозволяє отримати функцію щільності ймовірності для цільової змінної. Показано, що байєсівську регресійну модель із використанням логістичної кривої можна використовувати для прогнозної аналітики поширення коронавірусу. Така модель може бути ефективною, коли є експоненційне зростання кількості підтверджених випадків коронавірусу. Важливо знайти максимум випадків зараження коронавірусом за добу. Цей максимум відображає половину часового періоду поширення коронавірусу. Очевидно, що нові дані та експертне формування апріорних розподілів параметрів моделі можуть суттєво коректувати попередні результати прогнозування. Пандемія коронавірусу має великий вплив на фінансовий ринок. Аналіз такого впливу є важливим, зокрема при формуванні стабільних інвестиційних портфелів. Досліджено вплив COVID-19 на динаміку індекса S&P-500 на фондовому ринку. Як регресійні ознаки, використано часові ряди кількості відвідувань сторінок Вікіпедії, які мають пряме чи опосередковане відношення до тематики коронавірусу. Як цільову змінну використано кількісні характеристики індексу S&P-500. Отримані результати свідчать, що різні ознаки мають різний вплив та різну невизначеність щодо цільової змінної. Найефективнішою та найменш варіативною серед розглянутих ознак була ознака на основі кількості відвідувань сторінки Вікіпедії про вакцину. Розглянуто вплив кризи, зумовленої пандемією коронавірусу на акції компаній на фондовому ринку та проведено порівняльний аналіз цього впливу із впливом кризи 2008 року та спадом ринку 2018 року. Для цього використано лінійну класичну та байєсівську регресії. Отримані результати показують, що різні кризи по-різному впливають на динаміку цін акцій внаслідок реалізації різних механізмів впливу. Підхід на основі байєсівського виведення дозволяє аналізувати невизначеність впливу різних фінансових криз. Результати показують, що невизначеність коронавірусної кризи більша порівняно з іншими кризами. Розрахунок невизначеності дозволяє робити оцінку ризиків для інвестиційних портфелів та різних фінансових та бізнес-процесів. COVID-19 активно обговорюється у соціальних мережах, тому характеристики повідомлень, зокрема у мережі Твітер можуть мати прогнозні власивості. Використовуючи теорію графів розглянуто зв’язки між користувачами мережі Твітер у масиві повідомлень, пов’язаних із COVID-19. Показано, що за допомогою алгоритмів теорії графів, можна виявляти спільноти користувачів та знаходити впливових користувачів за різними оцінками вершин. Показано, що використовуючи теорію частих множин та асоціативних правил, можна знайти семантичну структуру у масиві повідомлень в межах заданого тематичного поля.

Ключові слова: коронавірус, COVID-19, байєсівська регресія, фондовий ринок, прогнозна аналітика.

Стаття: надійшла до редакції 03.11.2020,
доопрацьована 07.11.2020,
прийнята до друку 08.11.2020