DOI: 10.55643/fcaptp.3.56.2024.4402

LEVERAGING BIG DATA TECHNOLOGIES FOR ENHANCED PUBLIC PARTICIPATION IN PUBLIC FINANCIAL MANAGEMENT

Sergii Krynytsia
Candidate of Economy Sciences, Associate Professor, Doctoral Student of the Department of Public Finance, State Tax University, Irpin, Ukraine;
e-mail: serge.krinitsa@gmail.com
ORCID: 0000-0002-5569-4682
(Corresponding author)

Oksana Hordei
D.Sc. in Economics, Professor of the Department of Public Finance, State Tax University, Irpin, Ukraine;
ORCID: 0000-0001-6938-0548

Yuliia Kovalenko
D.Sc. in Economics, Professor of the Department of Financial Markets and Technologies, State Tax University, Irpin, Ukraine;
ORCID: 0000-0002-5678-3185

Alla Dankevych
Candidate of Economy Sciences, Associate Professor of the Department of Financial Markets and Technologies, State Tax University, Irpin, Ukraine;
ORCID: 0000-0001-5158-1018

Andrii Boldov
Vice-rector for European Integration and Digital Transformation, State Tax University, Irpin, Ukraine;
ORCID: 0000-0002-0411-3972

Received: 01/04/2024
Accepted: 10/06/2024
Published: 30/06/2024

ABSTRACT

The article is devoted to the topical issues regarding the implementation of Big Data technologies in public finance management. The application of Big Data has the potential to enhance transparency and accountability in the use of budgetary resources, increase trust in government, improve the efficiency of budget resource utilization, better understand citizens' needs, and engage the public in public finance management. The purpose of the study is to explore theoretical, methodological, and practical aspects, as well as to develop recommendations for the implementation of Big Data processing and analysis technologies to enhance public participation in public financial management. The article examines traditional methods of civil engagement in the budgetary process, identifies their disadvantages, and explores Big Data technology potential based on Computational Linguistics and Machine Learning to strengthen public participation. Developments in sentiment analysis and opinion mining have been adapted to the field of public finance. A generative model for analyzing public sentiment on social networks regarding public finance management has been constructed and tested. The approaches developed for using Big Data technologies can be implemented in the field of public finance to enhance public participation in their management as advisory tools for the realization of representative democracy and require further theoretical elaboration and practical application to improve the analysis of alternative sentiments, prevent manipulation of public opinion, and abuse within the network.

Keywords: Big Data, public finance, budget, public participation, sentiment analysis, opinion mining, Machine Learning, social media

JEL Classification: H30, H56, H72, C55, D70

INTRODUCTION

The world is changing rapidly under the influence of digitization and Big Data technologies. Big Data is already deeply integrated into the economy and societal life. Effective public finance management is a key factor in ensuring economic growth, prosperity, and national development. However, this sphere remains fairly conservative, largely due to established institutional frameworks (Kovalenko, 2014; Kovalenko, 2013). However, examples from the private sector clearly demonstrate that the implementation of Big Data and Machine Learning can provide significant advantages and substantially improve outcomes in the public sector (Pantielieieva, et al., 2018a). Big Data technologies have the potential to enhance transparency and accountability in the use of budgetary resources, increase trust in government, enable more efficient use of public funds, better understand the public's needs, and engage the citizens in public finance management.

The latter aspect is of particular importance since democratic principles on the one hand demand increasing citizen involvement in decision-making in the field of public administration, while on the other hand, budget transparency and accountability alone are insufficient to meet this requirement without adequate feedback channels (Krynytsia, 2023).

Thanks to the development of web technologies, the democratization of publications has led to a sharp increase in the expression of opinions and attitudes of citizens towards

various issues of public life. This information, despite massive manipulations and informational attacks, potentially serves as a powerful feedback tool with government bodies, particularly regarding issues of public finance management.

LITERATURE REVIEW

Theoretical studies in the field of public finance management were studied by various scientists. For example, J. Gruber (2010) examines the relationship between public finance and public policy, distinguishing the reasons, tools, and effectiveness of government intervention in the economy.

A. Khan, W.B. Hildreth, and J.R. Bartle (2004) offer a comprehensive approach to understanding and analyzing financial decision-making by governments in the context of the need to satisfy multiple and often conflicting interests.

The founders of the theory of fiscal federalism R. Musgrave (1971) and W. Oates (1968) point to the economic benefits of decentralization obtained by taking into account the interests of the local population in the supplying of public goods.

W. Congdon, J. Kling, and S. Mullainathan (2011) use ideas and methods from behavioral economics to analyze financial policy, specifically examining how behavioral factors influence governments and citizens to make public finance decisions.

Such studies became the basis for the justification and development of theoretical approaches regarding the need for active public involvement in public finance decision-making to increase their economic and social efficiency. For example, A. Halachmi and M. Holzer (2010) conclude that there is a need for greater citizen participation in the budget process by improving government accountability, resulting in increased trust in state institutions and the development of democracy.

C. Ebdon and A. Franklin (2006) consider public participation not only as transparency and accountability of the government in budgetary matters but also as deeper involvement of citizens in the budget process through participatory budgets, public hearings, etc.

Y. Zhang and Y. Liao (2011) investigate two-way local government-community communication on participatory budgeting.

I. Shybalkina (2021) explores various forms of public participation in the budget process, in particular, the involvement of marginalized and vulnerable population groups in budget planning issues to take into account their problems and experiences.

The term Big Data was first used by Professor of Computer Science at Rutgers University Sholom M. Weiss and Sydney University lecturer Nitin Indurkhya in 1998 in their work "Predictive Data Mining: A Practical Guide." Weiss and Indurkhya defined Big Data as very large collections of data, while also pointing out their significant potential for intelligent analysis (Weiss et al., 1998).

There is no single criterion that defines the minimum data volume required to classify it as big, but for a general simplified understanding, the following definition is used: "Big Data is any data that does not fit into an Excel spreadsheet" (Batty, 2013).

Alongside the sheer size, another necessary but insufficient characteristic of Big Data is its ability to grow rapidly. The first person to emphasize this aspect of Big Data was Silicon Graphics Chief Scientist John Mashey in his presentation "Big Data and the Next Wave of InfraStress: Problems, Solutions, Opportunities." Mashey noted that the amount of dynamic computer memory grows annually on average by 60% or fourfold in three years (Mashey, 1999). This 60% annual growth rate is now used as an unofficial criterion for classifying data as "Big" (Oleshchenko, 2021).

David Laney identified the third characteristic of Big Data - variety. Variety means Big Data comprises large collections of data, both structured and semi-structured or entirely unstructured (Laney, 2001). He also introduced the concept of 3Vs to denote the characteristics of Big Data: Volume, Velocity, and Variety.

The most common type of data format is structured data, which is readily processed and analyzed (using SQL, for example) and is represented in tables in accordance with a predetermined data model. Nonetheless, unstructured data is becoming increasingly significant in the overall data flow, including huge text fragments, audio and video recordings, and streaming media (Chen et al., 2022). These data can originate from a number of sources, including electronic sensors, video recordings made by cameras, news feeds, and user-shared content on blogs, vlogs, and social networking platforms.

Semi-structured data lacks a clearly organized structural model but can be relatively easy to analyze using tags or markers for semantic element separation (Chen et al., 2022). Examples include hash-tagged social media posts.

The term Big Data gained wide usage in 2008 with a special issue of Nature dedicated to the explosive growth of global information. As a result, the main credit for studying the Big Data phenomenon is often mistakenly attributed to the chief editor of the special issue, Professor Clifford Lynch of the University of California, Berkeley School of Information. Lynch included Value as the fourth of David Laney's three Vs. Lynch highlighted the capacity to uncover concepts and patterns from large datasets that result in innovations across a range of industries (Lynch, 2008).

The fifth characteristic (5th V) of Big Data was proposed by Dr. Arvind Sathi in the work "Big Data Analytics: Disruptive Technologies for Changing the Game" - Veracity. Sathi emphasizes that most Big Data comes from external sources lacking necessary management and homogeneity. Veracity refers to the credibility and correctness of data sources, as well as data suitability for use. Value extraction and data management have become important competitive advantages that every organization seeks (Sathi, 2013). In line with Sathi's ideas, futurist and business and technology strategy consultant Bernard Marr points out that although it can be more difficult to ensure data truthfulness and accuracy with such volumes, particularly for unstructured data like social media posts that contain hashtags, slang, and abbreviations, volume frequently makes up for shortcomings in quality and accuracy (Marr, 2014).

The Big Data phenomenon has become widely recognized, and it can be understood based on these 5Vs.

If initially, Big Data processing and analysis technologies were the subjects of interest in scientific research in Computer Science and Data Science, the development of this tool opened up prospects for business, particularly in the field of finance (improved decision-making, risk management, customer engagement in financial institutions, in financial markets, etc.), management, marketing, etc.

For instance, the research by Chen et al. (2012) shows the usefulness of big data analytics in risk management in financial institutions. Their study demonstrates how cutting-edge analytics techniques, such as predictive modeling and machine learning, enable more precise risk assessment and the early identification of possible market disruptions.

Cartea and Penalva (2011) investigate how algorithmic trading methods and market microstructure are affected by Big Data. The significance of high-frequency data and computational methods in creating intricate trading algorithms that take advantage of price dynamics and market inefficiencies is emphasized.

The financial sector has seen a transformation in fraud detection and compliance because of Big Data analytics. Wu, Wang et al. (2022) research highlights the efficacy of Big Data algorithms and Machine Learning in identifying fraudulent activities and guaranteeing adherence to regulatory mandates. Their conclusions emphasize how crucial anomaly identification and real-time data processing are to lowering financial risks.

Big Data analytics has great promise for managing portfolios and financial forecasts. According to studies by Aldridge and Avellaneda (2021), the use of alternative data sources and sophisticated data analytics techniques can improve asset price forecasts, portfolio efficiency, and investment opportunities. All of which help investors make wise decisions in volatile market conditions.

Scholars such as Verhoef et al. (2015) investigate how financial organizations might enhance their customer relationship management procedures by using Big Data analytics. Their research highlights how crucial data-driven insights are for personalized marketing, customer segmentation, and product recommendations, leading to increased customer satisfaction and loyalty. Thanks to Pang and Lee's (2008) pioneering work, research directions like sentiment analysis and opinion mining have emerged from the study of consumer behavior. Since the emergence of social media sites like Facebook and Twitter, researchers have looked into using big data techniques for user-generated content sentiment analysis and opinion mining.

The studies by Bollen et al. (2011) and Pak and Paroubek (2010) demonstrate the effectiveness of machine learning algorithms in sentiment classification and identification of confident content in large-scale social media datasets.

Big Data analytics allows companies to obtain valuable information from customer feedback and reviews. Scholars such as Liu (2012) have explored methods for product analysis and trend identification to provide insights for marketing strategies and product development initiatives. Businesses can evaluate customer satisfaction levels, spot emerging trends, and take preemptive measures to resolve possible problems by examining vast repositories of online reviews.

The rise of Big Data in finance poses ethical and confidentiality concerns notwithstanding its potential for transformation. According to Trinder's research (2019), financial institutions must ensure human involvement in algorithmic decision-making processes and practice self-regulation. Researchers like Mergel et al. (2016), who critically examine the ethical implications of data collection, storage, and use, are also concerned about these same issues. They emphasize the need

DOI: 10.55643/fcaptp.3.56.2024.4402

for strong regulatory frameworks and ethical principles to protect consumer interests and mitigate potential abuses of data-driven technologies.

To summarize, there is an increasing body of research on the subject of Big Data in finance and public management. These studies look at the diverse applications and implications of Big Data in a range of contexts, such as risk management, trading strategies, fraud detection, customer relationship management, financial forecasting, and ethical issues. Still unanswered are many of the issues surrounding Big Data's incorporation into the field of public finance management.

AIMS AND OBJECTIVES

The purpose of the study is to explore theoretical, methodological, and practical aspects, as well as to develop recommendations for the implementation of Big Data processing and analysis technologies to enhance public participation in public financial management.

To achieve this goal, the following tasks need to be realized:

- to define the potential of Big Data in public financial management;
- to investigate forms of public participation in public financial management and prospects for enhancing it through the use of Big Data processing and analysis technologies;
- to research and test the toolkit for sentiment analysis in social media based on Big Data technologies as a tool for civil engagement in public finance management.

METHODS

The study employed two main groups of methods: analysis and synthesis; in the process of analysis of scientific sources, the essential features of Big Data were identified; systematization - to systematize the possibilities of Big Data and the prospects of using Big Data technologies in public finance; analogies - to develop recommendations for applying the achievements of science and practice in the use of Big Data in various fields (computational linguistics, Data Science, private sector finance, etc.) in the sphere of public finance; graphical - for graphical representation of trends and relationships in data, facilitating the visualization of patterns and variations in indicators of social networks and open data portals of public finances in Ukraine; comparison - for comparing indicators over time, average indicators, etc.; semi-supervised Machine Learning - for building a training model for analyzing user sentiments on social networks; linguistic analysis - for binary classification of positive/negative emotional tone of text; mathematical modeling methods - for building a generative model for processing unstructured data to analyze public opinion on public finance management issues.

Statistical methods were also used, in particular, generative models based on Bayes' theorem are used to build a training model for processing unstructured data (Stuart and Ord, 1994):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1)$$

where A and B are events, P(A) and P(B) are the prior probabilities of A and B, P(A|B) is the conditional probability of event A occurring given that B is true, P(B|A) is the probability of event B occurring given that A is true.

Bayes' theorem forms the basis for naive Bayes classifiers:

$$\tilde{y} = \underset{k\in(1,\ldots,K)}{\operatorname{argmax}}\, p(C_k) \prod_{i=1}^{n} p(x_i|C_k), \quad (2)$$

where C_k - class, K - possible outcomes or classes, x = (x_1, …, x_n) – vector encoding some n features.

Since "Economic-mathematical modeling of the socio-economic system based on online Big Data algorithms makes it possible to predict consumer behavior based on the identification of business logic and to form a consumer profile in the decision-making system. This method is traditional, but the selection of characteristic functional features for forecasting efficiency and optimization of Slick-Through-Rate forecasting processes is special in view of machine learning as a tool for economic and mathematical modeling of the management decision-making system" (Kulyk et al., 2023), we can explore popular social networks.

We examined 1300 posts on popular social networks such as Facebook, Instagram, YouTube, and X (formerly Twitter), marked with the hashtag "#грошіназсу" (#moneyforarmy). Posts that are unrelated to this movement were filtered out (e.g., private volunteer fundraisers or commercial spam). Marketing analysis tools such as Keyhole and Awario, as well as custom scripts based on JavaScript and Python, were used for web scraping from the mentioned social networks.

RESULTS

One of the most important aspects of government administration is Public Financial Management (PFM). It includes all of the procedures used to gather, distribute, spend, and account for public resources. Therefore, tax collecting, government procurement, audits, and the full budgeting cycle are all included in Public Financial Management procedures. Responsible, transparent, and trustworthy management of public finances is a cornerstone of public administration reform. It is essential to provide the public with high-quality public services as well as creating and maintaining fair, sustainable economic and social conditions in the country. PFM involves a number of highly complex, technical processes and tasks, including macroeconomic forecasting, accounting, auditing, and budgeting allocation. Such procedures are complicated, which makes them opaque to the public and encourages corruption.

Allocative efficiency, operational efficiency, and general economic and fiscal stability have Traditionally been the main goals of PFM reforms. To this purpose, the majority of PFM reforms focus on technical methods to enhance the functionality of the Public Financial Management systems. Examples of these methods include revenue collection, public expenditure management, automation, modernization, and integration of PFM processes, as well as procurement systems.

PFM reforms are currently trending toward going beyond technical changes and emphasizing accountability, openness, and public engagement in PFM (Morgner & Chene, 2015).

According to the New Fiscal Transparency Code, the primary forms of such public participation include budget transparency: "publication of fiscal reports covering the entire public sector, preparation of full public sector balance sheets, and publication of audited annual financial statements" and "public participation in budget deliberations, in response to consultations with key stakeholder" (International Monetary Fund, 2014).

In what ways can municipal and federal governments improve budget transparency and public participation in Public Financial Management through the use of Big Data? First, let's discuss whether publicly available information from the Ministry of Finance and other government agencies qualifies as Big Data.

The maximum number of rows (records) in a Microsoft Excel spreadsheet is 1,048,576 (End, 2023). So based on this definition, the data collection on budget expenditures from the Unified Web Portal for Public Funds Usage of Ukraine already qualifies as "Big Data" in its first year of existence - the number of documents in the portal's database exceeded 10 million in 2016 (Figure 1).

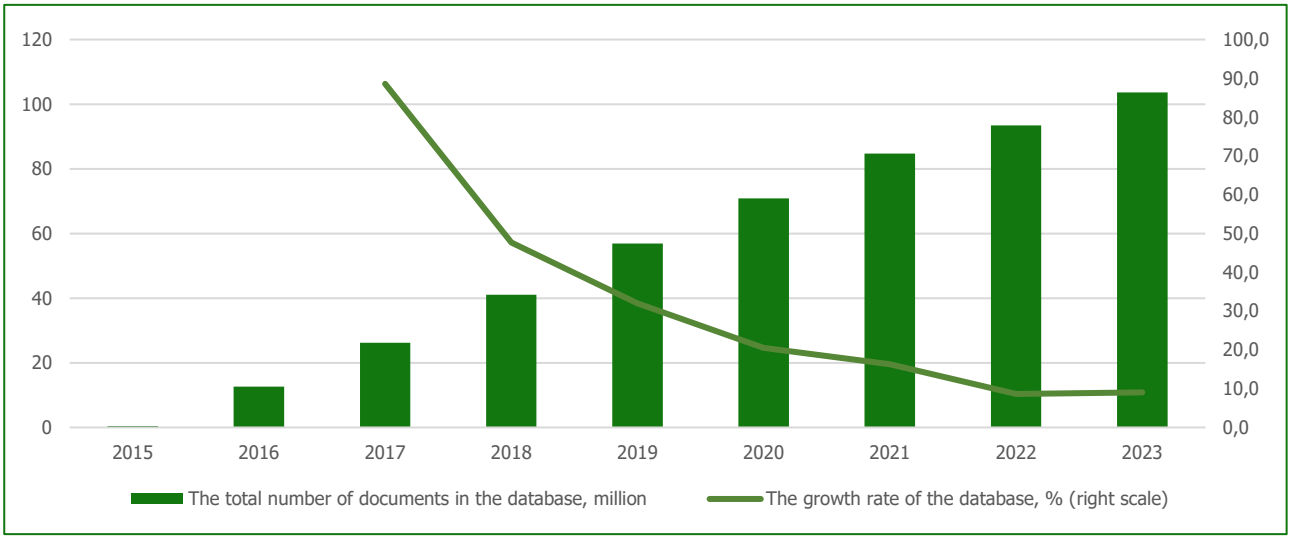

Figure 1. Volume and Growth Rate of the Unified Web Portal for Public Funds Usage Database in Ukraine in 2015-2023. *(Source: constructed by the authors based on data from Ministry of Finance of Ukraine, 2024)*

DOI: 10.55643/fcaptp.3.56.2024.4402

According to the Unified Web Portal for Public Funds Usage, as of January 1, 2024, 86.9 million documents and 228.8 million transactions were registered on the portal (Ministry of Finance of Ukraine, 2024).

As we can see from the data in Figure 1, the high growth rates of data on the Unified Portal for Public Funds Usage in Ukraine were observed only in the early years of its operation (until 2017), and now they have slowed to 10% per year, which is expected since with a fixed number of budgetary fund managers, data of this type cannot demonstrate constant high growth rates.

Another state resource for public finance, the Public Electronic Procurement System Prozorro, demonstrated high database growth dynamics. In 2020, the number of announced procurements on the platform increased by 171%, and the number of proposals increased by 160%. However, by 2021, the growth rate of the database had dropped to only 44% (Smart-Tender, 2022). Due to the Russian invasion of Ukraine in 2022 and the permission for direct contracting, the growth of the database significantly decreased. In times of peace, with the increase in the number and competitiveness of trades and the number of participants, such data growth rates may accelerate. However, it should be understood that the limited number of budgetary fund managers will still impose certain restrictions on the growth of the database at high rates. Thus, the data collection of electronic public finance systems in Ukraine is not Big Data in the exact sense of the term.

Furthermore, scientists in their recent research focus more on other characteristics of Big Data, such as their velocity growth and variety (Goswami et al., 2019).

What data can be useful for implementing Big Data processing and analysis technology in public financial management? Let's concentrate on some of the other important aspects of Big Data, such as their variety (structured, unstructured, and semi-structured data).

Approximately 80% of relevant data is unstructured, according to recent studies, and is typically not exploited for decision-making (Benz and Müller, 2023). Their application has a lot of promise, though.

Take the private sector as an illustration. Unstructured data is used in banking and FinTech to investigate consumer behavior, reduce risk, and enhance banking products and interactions (Pantielieieva et al., 2018b). These same directions can be useful in the public sector as well. Drawing an analogy with the private sector, the public sector also offers goods, but they are social goods, the quality of which can be improved by studying the behavior, thoughts, and preferences of consumers, in this case, the public. Such study will provide feedback from society on the quality of public services, thus ensuring public participation in public administration, as one of the main requirements for democratizing society.

Therefore, public participation in public financial management is important for ensuring the efficiency of budgetary expenditure management. Unlike corporate finance, where management efficiency is primarily determined by economic efficiency criteria, for public finances, such criteria are not the main ones. Instead, the optimization of budgetary expenditures should be based on social efficiency criteria.

Social efficiency cannot be unambiguously expressed solely by quantitative indicators calculated, for example, based on financial reporting data. Addressing this problem long before the onset of digital transformations in finance relied on the distribution of state power (and hence public finances) according to the principle expressed by Milton Friedman in his work "Capitalism and Freedom" in 1962: "The second broad principle is that government power must be dispersed. If government is to exercise power, better in the county than in the state, better in the state than in Washington. If I do not like what my local community does, be it in sewage disposal, or zoning, or schools, I can move to another local community, and though few may take this step, the mere possibility acts as a check. If I do not like what Washington imposes, I have few alternatives in this world of jealous nations" (Friedman, 1962). The principle of decentralizing state power to the specific consumer of public goods developed into the concept of fiscal federalism, formulated by Wallace E. Oates: "the provision of public services should be located at the lowest level of government encompassing, in a spatial sense, the relevant benefits and costs" (Oates, 1999). Thus, the efficiency and effectiveness of public finances should be ensured by providing feedback on the efficiency of public services and proposing alternative approaches to service delivery.

However, the requirements of economic efficiency entail a certain distance between the specific consumer of public goods and the decision-making center in public finances. In Ukraine, as a result of the administrative-territorial reform of 2020, the lowest level of local self-government and fiscal federalism is the Hromada (Territorial Community). Most hromadas were formed by combining individual cities, towns, and villages, sometimes to the size of former districts, which previously served as an intermediate, more aggregated link between two other levels of administrative-territorial units (local councils and regions). The population of the largest rural hromada reaches 43,000 people, the area of the largest hromada exceeds 2,000 km^2, and the largest number of settlements included in the hromada is 81 (Territorial Communities, 2024).

Under such conditions, traditional methods of involving the public in addressing local issues, including implementing budget policies locally, such as public hearings, become ineffective, particularly due to limited time frames, locations, and difficulty in accessing the venues. Other traditional methods also have their well-known drawbacks. The electoral process is discrete, and citizens can express their attitude toward local or government policies once every 4-5 years (in peaceful conditions), significantly reducing the effectiveness of such feedback. Referendums on every local or nationwide issue, following Switzerland's example, are too costly. The same can be said for sociological surveys, especially if they are conducted at the level of each hromada or expand surveys on any issues related to government or local authority decisions. And "foot voting", whose theoretical basis was developed by Charles Tiebout (1956) and mentioned in the quote from Milton Friedman above, mostly carries a probabilistic character. While this method could represent the community's overall sentiment, it has limitations when it comes to impartially assessing the efficacy of particular municipal government initiatives. First of all, citizens will need to make major efforts to relocate and modify their surroundings. In actuality, a sizable portion of the populace is unable to perform such acts because of diverse monetary, familial, and other responsibilities. Second, since a variety of factors, like employment possibilities and the health of the real estate market, can affect people's decisions to relocate, "foot voting" may not always accurately reflect the motivations behind and effects of local government policies.

Thus, digital channels of communication between communities and local authorities are gaining increasing importance. By the way, these channels, due to their prevalence and the reduction of the digital divide between large and small population centers, are becoming powerful tools for citizen communication with authorities at any level, not just the lowest local level.

Among such digital communication channels, electronic petitions and voting within participatory budgets have already gained popularity (Krynytsia, 2023). These data themselves serve as alternative sources of information for analyzing the effectiveness of budgetary policies and decision-making and thus can be used in the complex management of public finances based on Big Data analysis. However, these data are clearly insufficient for quality and comprehensive analysis and determination of effectiveness. The submission of applications and voting within the participatory budget before the Russian invasion was limited by periodicity once a year and was held only in certain large cities of Ukraine; during the state of war, it is not carried out at all. Electronic petitions are becoming increasingly popular, but their consideration is limited by a sufficiently high threshold of verification by public signatures (although even those petitions that have not passed such verification can be a source of un- and semi-structured data), restrictions on the number of petitions (due to spam concerns, blur of votes by signatories for petitions on similar topics, etc.), their non-systematic and sporadic nature.

Moreover, as research in the field of public relations shows, one of the significant problems of modern methods of analysis and sociological science is the large time gap between the problem and the conclusions and proposals of managerial decision-making (Isett, 2016).

At the same time, large volumes of unstructured data, which can practically analyze the social efficiency of budgetary policy and manage public finances in real-time, remain overlooked. Among them: are posts on social networks, citizen interviews, data from sociological surveys, expert reviews and reports, news feeds, and so on.

Leinweber (2009) distinguishes 4 groups of such data:

- news created by authoritative news agencies. These sources a priori have the highest reputation;
- pre-news - derived from primary sources of information such as reports, databases, survey data, etc. The reputation level depends on data interpretation;
- rumors - Leinweber includes blogs and websites that transmit thematic news, reviews, and analytics here. Their reputation depends heavily on the author's reputation, from the highest as in news from authoritative news agencies to unverified rumors and conspiracy theories, which have the lowest reputation;
- social media - Leinweber ranks them at the lowest level of the reputation scale, as the entry barriers are very low, and the ability to produce information is high. Thus, social networks are potentially a source of "dangerously inaccurate information".

Although social media are positioned as a potential source of inaccurate and misleading information, Leinweber, like other researchers, acknowledges that it is a valuable source of information that, when carefully applied, can be useful in analyzing big data and machine learning. Such data enables working with unique data sets, which are the most promising (Lopez de Prado, 2018).

In this context, we can draw parallels with Voice of the Customer (VOC) methodologies, which are already actively used in research for the benefit of commercial companies. They utilize alternative sources of Big Data, primarily social media data, to study consumer behavior and relationships with customers. Methods of studying such behavior using Big Data have been termed "sentiment analysis" (Hurwitz, 2015) or "opinion mining" (Kashyap, 2017).

DOI: 10.55643/fcaptp.3.56.2024.4402

Alternative data (whether text, images, audio, or video recordings or streams) traditionally fall into the qualitative category in economic analysis. Unlike quantitative data, qualitative data requires expert processing and evaluation when using traditional analysis methods. Qualitative data in the analysis have been used before, mostly limited to formalized sources such as sociological surveys, expert analytics, or news from top-rated agencies. With the significant increase in information flows, primarily due to the development of social networks, blogging, and other public information activities, the importance of these data is growing. According to the portal Statista "Estimates suggest the amount of data uploaded to social media globally doubles every two years" (Social Media & User-Generated Content, 2023).

On one hand, this is a powerful tool for analyzing consumer sentiment regarding public services and, more broadly, engaging the public in addressing socio-economic issues and, in particular, managing public finances. On the other hand, the use of social media data and other User-Generated Content (UGC) faces several practical challenges:

- users of social media platforms generally bear no responsibility for generating false, manipulative, or simply erroneous data. Consequently, a significant portion of alternative unstructured data may simply be fake, and without proper cleansing, the results of analysis would be erroneous;
- there are certain barriers to collecting social media data from service providers;
- collecting such data runs into moral and increasingly legal aspects of abuse of personal data;
- insufficient competence of public finance management officials, from the government level to the level of local self-government bodies, in the field of Big Data analysis.

The first problem is associated with the fact that social networks often fall victim to mass abuses due to the relative ease of creating anonymous accounts and the low level of identity verification. Compared to resources where user identification requires a Digital Signature or BankID (for example, platforms for submitting petitions), social networks do not impose such strict restrictions. This opens up wide opportunities for the spread of fake information, manipulation, as well as mass operations with vote stuffing and creating artificial trends. The absence of a clear mechanism for verifying identity makes social networks vulnerable to fraud and abuse. This is quite a serious problem, as the number of organized manipulations on social networks has increased by 150% since 2017, and currently, at least 70 countries are engaged in computer propaganda aimed at influencing public opinion (Sjouwerman, 2020).

To address this problem, technologies such as Big Data, Machine Learning, and Artificial Intelligence based on them are helpful once again. Big Data and Machine Learning play an important role in detecting fraud and manipulation on social networks. They enable the analysis of large volumes of data to identify patterns and anomalies that may indicate fraud. Machine Learning algorithms can be trained to recognize various forms of manipulation, such as fake accounts, botnets, and coordinated inauthentic behavior.

Furthermore, through advanced Machine Learning models based on Artificial Intelligence, fraud such as deepfake technology, data manipulation, and phishing, can be detected. The irony is that the same technologies are used to evade fraud detection, such as bypassing anomaly detection systems, which may indicate suspicious activity through so-called data poisoning (Constantin, 2021). Therefore, combating abuses and fraud on social networks requires constant improvement of algorithms and methods, reminiscent of the ongoing battle against the spread of computer viruses.

The next problem is related to access to gathering such data. In principle, any public posts on social media are accessible to any user, but organizing the collection of large volumes of such data requires the involvement of software tools, known as web scraping. Through web scraping, one can obtain publications, dates, comments, hashtags, user reactions, author locations, and so forth. This pertains to the collection of non-personal data. However, personal data (collecting information by login or name) is protected by the General Data Protection Regulation (GDPR), and scraping them without the individual's consent is prohibited (Milenkoski, 2023).

Although the legality of web scraping non-personal data has been separately affirmed by a US appeals court (Whittaker, 2022), whose jurisdiction covers most social media company owners, these companies do not welcome web scraping practices without their personal consent.

Meta, the owner of Facebook and Instagram, sues both software development companies for web scraping tools and individuals who collected data from these two social networks without Meta's consent (Wadhwani, 2022).

The owner of the social network X (formerly Twitter), Elon Musk, announced in the summer of 2023 limitations on the daily viewing of posts for each user (and consequently a complete ban on viewing content by unauthorized users), citing extreme levels of web scraping (Mask, 2023). Musk's Twitter also files lawsuits against individuals resorting to such data collection methods (Anurag, 2023).

The position of social media owners is understandable. In the context of the fourth industrial revolution, where information becomes the primary commodity, such assets have very high value. Therefore, it is quite likely that in the future, mass collection of Big Data from social networks will ultimately become officially paid (currently, this can be done, for example, unofficially through marketing analytics systems using paid accounts). Thus, governments and local authorities need to weigh the price of such information and the effectiveness of its use for analysis and decision-making in the realm of public finances.

The third problem partly touches upon the legislative prohibition of collecting personal data. However, arrays of non-personalized data still contain information about, for example, authors of posts and comments, so further analysis can be personalized, even to the extent of forming a specific profile of the individual, which, of course, violates their privacy (Mergel, 2016). As demonstrated by the practice of researching misinformation in social networks, even within the Meta company itself, they do not know how to ensure the confidentiality of their users (Donovan, 2020).

Regulation needs to address questions such as: Who can use such data and for what purposes? What volume of such data should be collected? How will identification information be stored and protected? Therefore, governments, researchers, and civil society partners need to work together to develop specific rules and mechanisms to ensure fairness, accountability, and transparency in the collection, processing, and storage of big data.

Finally, the fourth problem is related to the challenges faced by public finance officials in effectively processing large volumes of information, structured or unstructured, analyzing such data correctly, interpreting the results, and making appropriate managerial decisions. Meeting these requirements will require resources different from those typically used in government bureaucratic structures, especially in smaller jurisdictions.

Due to the complexity of sentiment analysis, there is no single standardized approach to its implementation. However, general approaches can be represented by an iterative process (Figure 2).

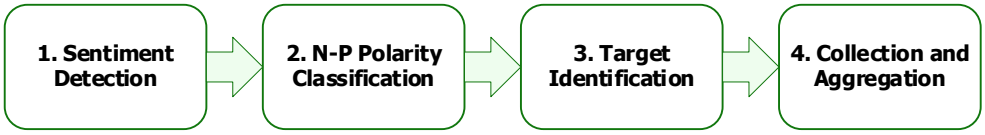

Figure 2. Sentiment Analysis Procedure. *(Source: constructed by the authors based on Delen, 2020)*

After gathering the necessary social media data, the first step in sentiment analysis is to identify objectivity through the calculation of objectivity-subjectivity (O-S) polarity. In other words, at this stage, distinctions are made between factual statements and opinions, filtering out simple factual statements. Typically, determining opinion relies on checking adjectives in the text (Delen, 2020).

The second stage involves classifying polarity using binary classification of positive or negative sentiment toward a particular issue highlighted in the publication. The main challenge here is that, unlike the formal language of news agencies, social media publications may contain rather informal text, rich in slang and, most problematically, sarcastic expressions, where binary classification of positive/negative can be misinterpreted. However, developments in this direction are underway using heuristic approaches, although identifying sarcasm in the text remains a non-trivial task in Natural Language Processing (NLP) (Buyya et al., 2016).

The third stage involves determining the purpose of the expressed sentiment. Often, a single publication may have several purposes, so the task of analyzing them includes distinguishing between them, including from a comparative standpoint if applicable.

Finally, in the fourth stage, the identified sentiments of individual textual data points and goals are aggregated and transformed into a single indicator of sentiment.

These methods and strategies are in the field of Computational Linguistics, thus we won't go into detail about particular algorithms here. Those who are interested can refer to Srinivasa-Desikan (2018), Vajjala et al. (2020), Kurdi (2017), Jurafsky (2023), and other relevant sources.

It is important to remember that data from social media and other unstructured sources is typically classified as unstructured data. This means that traditional data analysis methods relying on quantitative metrics are not suitable for such data types. Therefore, such data require fundamentally new approaches to processing, content extraction, and analysis.

DOI: 10.55643/fcaptp.3.56.2024.4402

The field of Data Science contains a variety of approaches to solving this problem, largely based on Machine Learning. Machine Learning on this type of data can provide more concise, semantically rich, descriptive patterns in the data that better reflect their internal properties. Machine learning technologies, such as semantic networks, include Natural Language Processing (NLP), created to facilitate access to such information and extract necessary data for analysis (Kurdi, 2017).

Some of this data is semi-structured because users tag them with hashtags, and such tagged data is more amenable to Supervised Learning (SL). Supervised learning uses labeled data to teach a computer to classify objects or achieve the expected level of correctness of training data (Kashyap, 2017). Supervised learning methods seek a data model that most accurately reflects the relationships between the target variable and predictors - independent variables (for example, using the same correlation-regression model). Of course, not all social media users tag their posts with hashtags, and even if they do, it does not guarantee that they are relevant to the information of interest to researchers of this data. However, algorithms based on Machine Learning and Artificial Intelligence can also be applied to determine relevant markers of unstructured information (Mitra, 2012).

However, the use of supervised learning has several drawbacks and obstacles:

1. The important aspect is not only the quantitative determination of the frequency of using certain markers but also the informational content (positive or negative), i.e., the author's attitude toward a particular object or phenomenon. The main difficulty lies in determining the context of the story from the standpoint of the feelings expressed in terms of emotional content by the author. In other words, we risk losing the context of the entire story and misinterpreting its content by relying solely on hashtags.
2. From a cost and time-saving perspective, it is necessary to minimize the costs of processing and cleansing such data, reducing them to an acceptable level.
3. Supervised learning methods suffer from well-known drawbacks of traditional modeling, including incorrect model specification, limited inferences, incorrect hypotheses, and so on.

To overcome these drawbacks, unsupervised (UL) or semi-supervised learning (SSL) is called upon.

Unsupervised learning (UL) does not require the prior formulation of hypotheses, there are no target or final variables, and the analysis itself aims to study the structure of the data and identify relationships between them (Kashyap, 2017). The model is built by determining structures and patterns using a systematic reduction of redundancy or organizing data by resemblance.

Semi-supervised learning (SSL) combines the advantages of SL and UL. The model is trained on a portion of the data defined by the researcher based on a pre-formulated hypothesis and defined structure. Subsequently, the model is tested on the rest of the unstructured data using generative models (semi-supervised generative adversarial network - SGAN) (Klaas, 2019). It is excellent for transforming raw text or images into tables with quantitative data representations, ready for computerized assessment. SSL gives us a set of fast computer methods designed to extract meaning from voluminous data, such as text (Aldridge and Avellaneda, 2021).

Machine learning algorithms are applied, in particular, to identify hidden patterns and other aspects in social media, capturing the mood of a document. Sentiment analysis or opinion mining was primarily modeled as a supervised learning problem (Buyya, 2016), meaning that the data for modeling must be labeled. The laboriousness of this process can be addressed by semi-supervised learning, where only a portion of the sample (labeled data) is used for training (Figure 3).

The naive Bayes classifier determines the probability of observation x_i belonging to one of the classes C_k assuming independence of variables (to prevent the so-called "curse of dimensionality" with exponential growth of sample size).

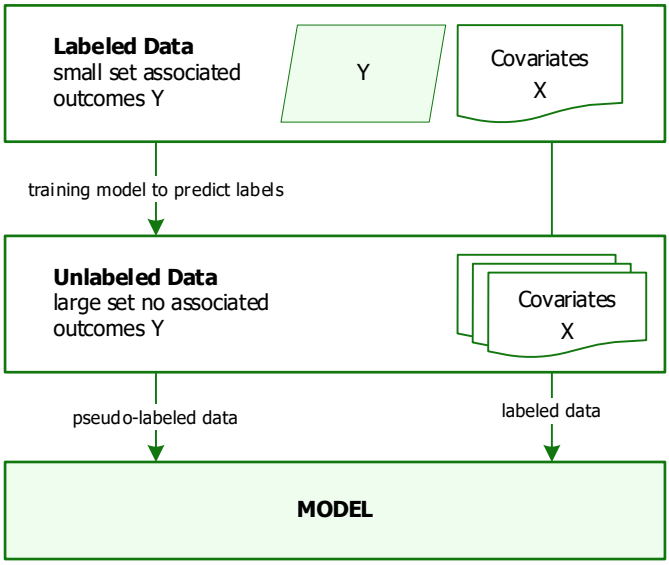

Figure 3. Semi-supervised learning. *(Source: Adapted from Aldridge and Avellaneda, 2021; Potrimba, 2022)*

If we adopt the naive Bayes classifier for classifying social media posts to assess authors' attitudes toward public finance issues, the number of classes in our case will be a finite K, like categories classifying various aspects of public finance management or related events. The dictionary containing a typical set of words about public finance has a size of $/X/$. Each new post i can be assigned to class C_k, where $k \in K$, and has $/x_i/$ words if it contains the word w_j from the dictionary. In this case, $x_{ij} \geq 1$, otherwise $x_{ij} = 0$. Each new post is generated according to the probability distribution described by the multinomial parameter F, thus the probability of post classification is $P(x_i/C_k, F)$. Based on Bayes' theorem:

$$P(x_i|F) = \sum_{k\in K} P(C_k|F)\, P(x_i|C_k, F), \tag{3}$$

The parameter that describes the multinomial distribution for the set of word probabilities:

$$F_{w_j|C_k} \equiv P\left(w_j\middle|C_k, F\right), \tag{4}$$

$$\sum_j F_{w_j|C_k} = \sum_j P\left(w_j\middle|C_k, F\right) = 1, \tag{5}$$

According to the naive Bayes classifier, the words in each post are conditionally independent of each other, given the class labels. Then, the probability of classifying the post as class C_k given a word from the dictionary w_j is described by the proportionality equation:

$$P(x_i|C_k, F) \propto P(|x_i|) \prod_{w_j\in X} P\left(w_j\middle|C_k, F\right)^{x_{ij}}, \tag{6}$$

And the complete generative model of post classification will take the form:

$$P(x_i|C_k, F) \propto P(|x_i|) \sum_{k\in K} P(C_k|F) \prod_{w_j\in X} P\left(w_j\middle|C_k, F\right)^{x_{ij}} \tag{7}$$

Let's demonstrate the application of the generative model (7) for analyzing public sentiments regarding the use of local budget funds in Ukraine amidst the war.

The reform of the administrative-territorial structure and budget decentralization in Ukraine from 2015 to 2020 significantly strengthened the financial base of local budgets, primarily at the lowest level of the budget system - the budgets of local communities (hromadas). The Russian invasion in 2022 and the sharp escalation of hostilities led to widespread mobilization of the population to the Armed Forces of Ukraine. This had the side effect of a sharp increase in personal income tax receipts (from the income of servicemen) in local communities where the rear units of the Armed Forces of Ukraine are based. Additionally, there were relocations of enterprises and businesses from the occupied territories and areas affected by hostilities. According to the Budget Code of Ukraine, 60% of personal income tax receipts are credited to the budget of the respective local community (64% from 2023, 40% to the budget of the city of Kyiv) (Budget Code of Ukraine, 2023).

DOI: 10.55643/fcaptp.3.56.2024.4402

Thus, in 2023, the revenues of the Kyiv budget (excluding interbudgetary transfers) increased by UAH 10.5 billion, or 17.2% compared to 2021, and personal income tax revenues increased by UAH 7 billion, or 23.7% (Reports on the implementation of the budget of the city of Kyiv, 2021-2023). In certain smaller territorial communities, there were noticeably greater gains in revenues. For instance, the municipal of Cherkasy's budget revenues climbed by 49.4% in 2023 compared to 2021, and the receipts from personal income taxes increased by 88.6%. (Reports on the implementation of the budget of the city of Cherkasy, 2021-2023).

Since the Budget Code of Ukraine clearly delineates the expenditure powers of budgets at different levels, and funding for national defense falls within the competence of the government and accordingly the state budget of Ukraine, local authorities continued to increase funding for expenses within their competence, even in conditions of significant budget deficits, primarily due to increased revenues. These expenses include road repairs, landscaping, investment projects in the field of municipal economy, etc.

Such inefficient spending of public funds during wartime led to the emergence of the public movement "Money for the Armed Forces" (Klymkovetsky, 2023), aimed at encouraging local authorities to review their expenditure policies funded by local budgets in favor of the defense. Since the Budget Code of Ukraine was amended by the Verkhovna Rada of Ukraine in November 2023 and, temporarily until the end of the martial law, personal income tax received from the financial support of military servicemen was centralized in the state budget, the protests did not cease, as the public's perception of inefficient use of local budget funds remained unchanged.

Although the Budget Code prohibits local budgets from directly financing national defense (for example, purchasing weapons or ammunition), local authorities can still finance such needs without violating existing legislation. This can be done, for instance, through financing territorial defense units, targeted local programs to support the Armed Forces, or simply by providing a subvention to the state budget of Ukraine (Grechka, 2023).

A vocabulary was compiled, including terms like "army," "local budget," "money for the army," "Kyiv City Administration" (the venue for actions in Kyiv), "local council," "drones" (as a generalization of funding for arms in the public consciousness), "pavement" (a similar generalization of inefficient spending of local budget funds), and so on. "Negative" terms were also identified, such as pay requisites in posts indicating private fundraising rather than budget funding, which helped filter out irrelevant content.

The application of the training model on labeled data and its implementation for pseudo-labeling of unstructured data allowed us to conclude the relevance of the proposed generative model (7) (Figures 4 and 5).

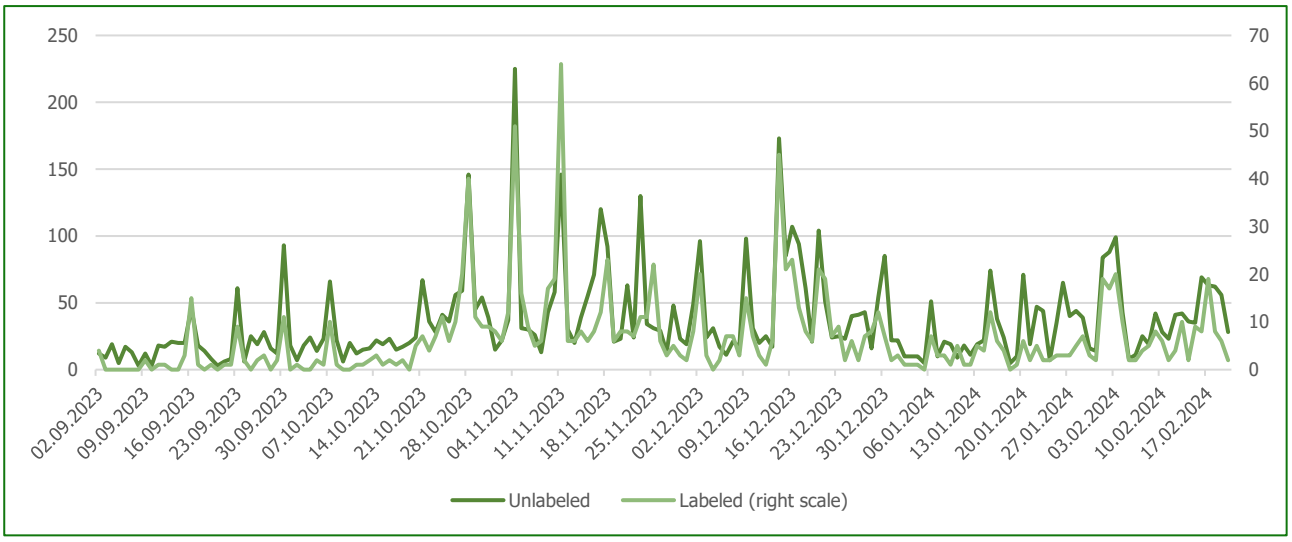

Figure 4. Number of posts in popular social networks about the "Money for the Armed Forces" campaign.

It is clearly visible that the dynamics of unstructured data, both in terms of posts and engagements, mirror the dynamics of labeled data. The weekly cycle (the "Money for the Armed Forces" protests are held every Saturday since autumn, so most posts come at the end of the week) demonstrates the dynamics of both structured and unstructured data.

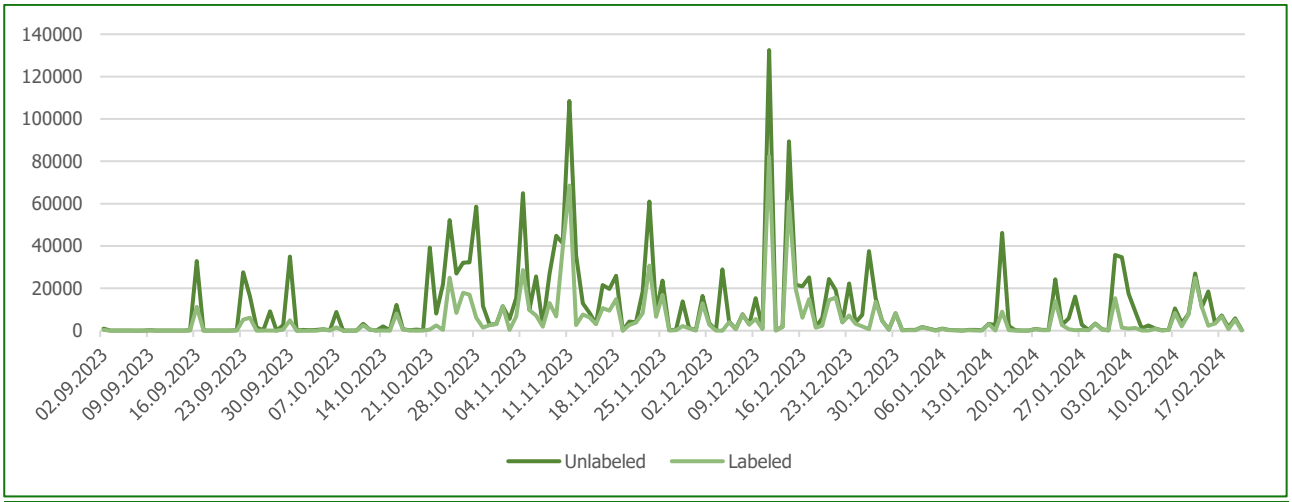

Figure 5. Number of engagements (reactions, comments, and shares) in posts on popular social networks about the "Money for the Armed Forces" campaign.

Unstructured data processed using the generative model (7) also mirrors the dynamics of labeled data, which correspond to significant events that impacted the subject of the protests, such as the session of the Kyiv City Council on December 14, 2023, where the draft city budget for 2024 was discussed, or February 1, 2024, when the Cherkasy City Council considered a controversial decision to reallocate more than half of the funds previously allocated in the city budget for 2024 for the financing of the Armed Forces in favor of the development budget (including the purchase of municipal equipment) and the mayor's bonuses (18000, 2024).

The number of posts expressing a negative attitude towards the public initiative "Money for the Armed Forces" detected by the mentioned methods was minimal. It generally corresponds to the highest societal support for the Armed Forces among all government institutions (Matiash, 2023). This is in reference to the polarity classification using a binary classification of positive or negative attitudes of the post author towards a specific issue, concerning the studied topic.

What about assessing the positive or negative attitudes towards the issue expressed in social media posts among engagements, it can effectively be evaluated only through the analysis of comments or one's opinion when sharing, while reactions are mostly interpreted as positive (there is a discussion about interpreting the meaning of the angry smile on Facebook, but it cannot be definitively interpreted, as it can represent either a negative attitude towards the author's position or support for the author in a negative context regarding the issues expressed in the post). The results of the conducted analysis of binary classification of positive or negative attitudes in comments to posts supporting the "Money for the Armed Forces" campaign on the 4 most popular social networks are presented in Figure 6.

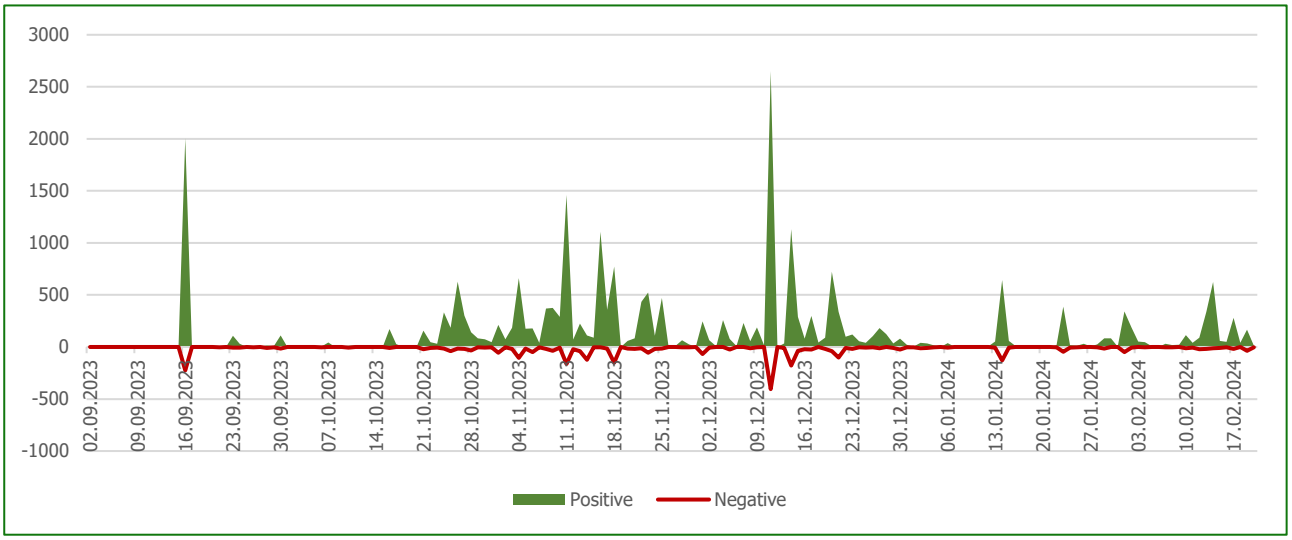

Figure 6. Binary classification of polarity attitudes towards the "Money for the Armed Forces" campaign in comments on posts from popular social networks.

DOI: 10.55643/fcaptp.3.56.2024.4402

As we can see, overall positive attitudes towards the public initiative significantly outweigh the negative ones. Negative comments primarily focus on societal fears that such a campaign aims to roll back the achievements of previous democratic reforms, particularly decentralization efforts, or is aimed at removing inconvenient representatives of local self-government for the central authorities (especially in Kyiv).

Overall, the following formula can be used to generate an integral support coefficient based on the examination of public sentiment regarding support (or lack thereof) for a given process in the country's or a particular region's socioeconomic life:

$$\beta_{C_k} = \frac{\sum_i E_i^+}{\sum_i E_i^-} \quad (8)$$

where E_i^+ and E_i^- represent positive and negative reactions regarding the issue classified by the feature C_k.

In our case, the integral support coefficient for the "Money for the Armed Forces" campaign is 8.5, indicating the predominant support for this initiative by the public and serving as an indicator for local authorities regarding societal preferences regarding the effectiveness of distributing funds from local budgets for their designated purposes.

DISCUSSION

Certainly, such a quantitative approach to assessing societal sentiments is rather simplified, but it demonstrates a further direction of research that, in our opinion, could move away from the binary classification of positive/negative attitudes. Some developments in this direction already exist, for example, Bouazizi (2019). However, the proposed approaches are mainly aimed at classifying the emotions of social network users (for example, fun, happiness, anger, sadness, hate, etc.). At the same time, there are certain difficulties with the interpretation of negations (which is not always a positive/negative attitude polarity switch), the above-mentioned sarcasm, as well as biases. However, these aspects are precisely important in the analysis of public sentiment regarding financial policy issues, and not the exact classification of the emotional coloring of users' messages in social networks.

CONCLUSIONS

The social and economic life are actively incorporating Big Data technologies. However, in the public finance sphere, as well as in the public sector as a whole, which is inherently conservative, these technologies have not yet gained sufficient traction. Nonetheless, as demonstrated in the private sector, their utilization can yield significant benefits and dividends. Big Data technologies in public finance can contribute to preventing abuses in government procurement, assist in budget planning, enhance transparency and accountability of government and local self-government in the use of budgetary funds, improve the efficiency of public administration and budget policy optimization, and so forth.

One of the promising directions for the development of Big Data technologies and machine learning based on them is to enhance the efficiency of management decisions in public finance through public engagement.

In the practice of public finance management, there are various forms of public participation. In particular, the public can traditionally exercise its attitude to state policy through elected representative democracy, but this process is discrete and has known drawbacks. Other forms of public participation are budget transparency and information, public hearings, participatory budgets, electronic petition mechanisms, etc.

A promising tool for ensuring public participation in public finance management is sentiment analysis or opinion mining, particularly through social media analysis.

We conducted a study of sentiment analysis tools using Big Data technology, developed a generative model of social media post classification, and tested it through social network analysis using the example of the public initiative "Money for the Armed Forces".

Studies have shown that developed approaches to sentiment analysis in computational linguistics can be implemented in the sphere of public finance to take into account public attitudes toward various issues related to budgetary expenditure distribution, taxation, government procurement, prioritization of funding for projects of national and local importance, and so forth. However, considering the imperfection of these technologies and the proliferation of manipulations and abuses, such approaches in the near future require further theoretical elaboration and practical testing and can only serve as

advisory tools for decision-making by traditional democratic institutions such as parliament or local self-government bodies, and in no way should they replace representative democracy.

Subsequent research could concentrate on examining alternative viewpoints, and their relevance to the matter at hand, as well as on averting manipulation of public opinion, fending off informational attacks, and other related topics.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

All authors have contributed equally.

FUNDING

The Authors received no funding for this research.

CONFLICT OF INTEREST

The Authors declare that there is no conflict of interest.

REFERENCES

1. 18000. (2024, February 1). Cherkasy deputies gave the military less than half of the promised funds earlier. *18000*. https://18000.com.ua/strichka-novin/cherkaski-deputati-dali-vijskovim-menshe-polovini-obicyanix-ranishe-koshtiv/
2. Aldridge, I., & Avellaneda, M. (2021). *Big Data Science in Finance*. John Wiley & Sons.
3. Anurag. (2023, July 14). Elon Musk's Twitter sues four individuals for illegal data scrapping. *Gizmochina*. https://www.gizmochina.com/2023/07/14/twitter-sues-four-individuals-illegal-data-scrapping/
4. Batty, M. (2013). Big data, smart cities, and city planning. Dialogues in Human Geography. *SageJournals*, 3(3), 274-279. https://doi.org/10.1177/2043820613513390
5. Benz, M., & Müller, M. (2023, November 14). 80% of Data Is Generally Considered Unstructured Data and Is Left Unused for Decision Making. *Squirro*. https://squirro.com/squirro-blog/4-valuable-insights-banks-can-gain-unstructured-data 2023
6. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007
7. Bouazizi, M., & Ohtsuki, T. (2019). Multi-class sentiment analysis on Twitter: Classification performance and challenges. *Big Data Mining and Analytics*, *2*(3), 181-194. https://doi.org/10.26599/BDMA.2019.9020002
8. Budget Code of Ukraine. (2023). https://zakon.rada.gov.ua/laws/show/2456-17#Text
9. Buyya, R., Calheiros, R., & Dastjerdi, A. (2016). Big Data: Principles and Paradigms. Morgan Kaufmann. https://dhoto.lecturer.pens.ac.id/lecture_notes/internet_of_things/Big%20Data%20Principles%20and%20Paradigms.pdf
10. Cartea, A., & Penalva, J. (2011, May 30). Where is the Value in High Frequency Trading? Banco de Espana Working Paper, 1111. http://dx.doi.org/10.2139/ssrn.4554933
11. Chen, C., Murphy, N. R., Parisa, K., Sculley, D., & Underwood, T. (2022). *Reliable Machine Learning*. O'Reilly Media, Inc.
12. Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165-1188. https://doi.org/10.2307/41703503
13. Congdon, W.J., Kling, J.R., & Mullainathan, S. (2011). *Policy and Choice: Public Finance through the Lens of Behavioral Economics*. Washington, DC: Brookings Institution Press.
14. Constantin, L. (2021, April 12). How data poisoning attacks corrupt machine learning models. *CSO*. https://www.csoonline.com/article/570555/how-data-poisoning-attacks-corrupt-machine-learning-models.html
15. Delen, D. (2020). *Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners*, 2nd Edition. FT Press.
16. Donovan, J. (2020, January 14). Redesigning consent: Big data, bigger risks. *Misinformation Review*. https://misinforeview.hks.harvard.edu/article/big-data-bigger-risks/
17. Ebdon, C., & Franklin, A. (2006). Citizen Participation in Budgeting Theory. *Public Administration Review, 66.* 437-447. https://doi.org/10.1111/j.1540-6210.2006.00600.x
18. End, N. (2023). The Excel Row Limit is 1,048,576 Rows. *Row Zero*. https://rowzero.io/blog/excel-row-limit
19. Friedman, Milton. (1962). *Capitalism and Freedom*. University of Chicago Press.
20. Goswami, S., Kumar, A., & Mukherjee, S. (2019). *Big Data Simplified.* Pearson Education India.

DOI: 10.55643/fcaptp.3.56.2024.4402

21. Grechka. (2023, November 6). "Money for the Armed Forces"? Can communities direct funding to the military? https://gre4ka.info/suspilstvo/76220-hroshi-dlia-zsu-chy-mozhut-hromady-napravliaty-finansuvannia-armii/

22. Gruber, J. (2010). *Public Finance and Public Policy* (Third Edition). Worth Publishers.

23. Halachmi, A., & Holzer, M. (2010). Citizen Participation and Performance Measurement: Operationalizing Democracy Through Better Accountability. *Public Administration Quarterly, 34,* 378-399. https://doi.org/10.2307/41288353

24. Hurwitz, J., Kaufman, M., & Bowles, A. (2015). *Cognitive Computing and Big Data Analytics*. John Wiley & Sons.

25. International Monetary Fund (2014). IMF Survey: New Fiscal Transparency Code to Improve Policies and Accountability. https://www.imf.org/en/News/Articles/2015/09/28/04/53/sopol080714a

26. Isett, Kim, R., Brian W., & VanLandingham, G. (2016). Caveat Emptor: What Do We Know about Public Administration Evidence and How Do We Know It? *Public Administration Review, 76*(1), 20–23. https://doi.org/10.1111/puar.12467

27. Jurafsky, D., & Martin, J. (2023). *Speech and Language Processing*. Third Edition draft. Stanford. https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

28. Kashyap, P. (2017). *Machine Learning for Decision Makers*. Apress Berkeley, CA. https://doi.org/10.1007/978-1-4842-2988-0

29. Khan, A., Hildreth, W., & Bartle, J. (2004). *Financial Management Theory in the Public Sector*. Praeger.

30. Klaas, J. (2019). *Machine Learning for finance*. Packt Publishing. https://proquest.safaribooksonline.com/9781789136364

31. Klymkovetsky, M. (2023, September 16). A rally gathered under the walls of the KMDA: people demanded to direct money "to the army, not to paving stones". *Hromadske*. https://hromadske.ua/posts/pid-stinami-kmda-zibravsya-miting-lyudi-vimagayut-spryamuvati-groshi-na-armiyu-a-ne-brukivku

32. Kovalenko, Yu. (2013). Standards within the Code of Good Practice for financial activities. *Actual Problems of Economics, 148*(10), 8–14. https://www.researchgate.net/publication/291850207_Standards_within_the_Code_of_Good_Practice_for_financial_activities

33. Kovalenko, Yu. (2014). Research toolkit for transformations in financial activities. *Actual Problems of Economics*, *154* (4), 51–58. https://www.researchgate.net/publication/288301861_Research_toolkit_for_transformations_in_financial_activities

34. Krynytsia, S. (2023). Modern trends in the development of digital technologies and their impact on public finances. *Collection of scientific papers of the State Tax University, 2*(2023), 82-120. https://doi.org/10.33244/2617-5940.2.2023.82-120

35. Kulyk, P., Hurochkina, V., Patsai, B., Voronkova, O., & Hordei, O. (2023). Maximizing customer satisfaction and business profits through Big Data technology in Society 5.0: a crisis-responsive approach for emerging markets. *CEUR Workshop Proceedings,* 3465, 82–94. https://ceur-ws.org/Vol-3465/paper09.pdf

36. Kurdi, M. (2017). *Natural Language Processing and Computational Linguistics 2: Semantics, Discourse and Applications*. ISTE Ltd. https://doi.org/10.1002/9781119419686

37. Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Velocity and Variety**.** META Group Research Note. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

38. Leinweber, D. (2009) *Nerds on Wall Street.* John Wiley & Sons.

39. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

40. Lopez de Prado, M. (2018). *Advances in financial machine learning*. John Wiley & Sons.

41. Lynch, C. (2008). Big data: Science in the petabyte era. *Nature*, 455, 1-50. https://www.nature.com/nature/volumes/455/issues/7209

42. Marr, B. (2014, March 6). Big data: The 5 Vs everyone must know. https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know

43. Mashey, J. (1999). Big Data and the Next Wave of InfraStress Problems, Solutions, Opportunities. https://www.usenix.org/conference/1999-usenix-annual-technical-conference/big-data-and-next-wave-infrastress-problems

44. Mask, E. (2023, July 1). To address extreme levels of data scraping & system manipulation. [X post]. X. https://twitter.com/elonmusk/status/1675187969420828672

45. Matiash, T. (2023, July 26). Most Ukrainians trust the Armed Forces, volunteers, and the State Emergency Service, according to a survey. *Livyy Bereh*. https://lb.ua/society/2023/07/26/567094_bilshist_ukraintsiv_doviryayut_zsu.html

46. Mergel, I., Rethemeyer, R., & Isett, K. (2016). Big Data in Public Affairs. *Public Administration Review*, *76* (6), 928-937. https://doi.org/10.1111/puar.12625

47. Milenkoski, M. (2023). Legal and Privacy Challenges of Data Scraping in the Digital Age. *GDPR Local*. https://gdprlocal.com/legal-and-privacy-challenges-of-data-scraping-in-the-digital-age/

48. Ministry of Finance of Ukraine. (2024). Spending. Unified Web Portal for Public Funds Usage of Ukraine. https://spending.gov.ua/new/statistics/documents

49. Mitra, G., & Mitra, L. (2012). *The Handbook of News Analytics in Finance*. https://doi.org/10.1002/9781118467411

50. Morgner, M., & Chene, M. (2015). Public Financial Management. *Transparency International*. https://knowledgehub.transparency.org/topics/public-financial-management-parent-label

51. Musgrave, R. (1971). Economics of Fiscal Federalism. *Nebraska Journal of Economics and Business, 10*(4). https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwigt7Lg79CGAxUmEBAIHca9CjsQFnoECBwQAQ&url=https%3A%2F%2Fcooperative-individualism.org%2Fmusgrave-richard_economics-of-fiscal-federalism-1971-autumn.pdf&usg=AOvVaw1-m8sovNI-Yl1xo6NXp4Oj&opi=89978449

52. Oates, W. (1999). An Essay on Fiscal Federalism. *Journal of Economic Literature, 37*(3), 1120-1149. https://www.jstor.org/stable/2564874

53. Oates, Wallace E. (1968). The Theory of Public Finance in a Federal System. *The Canadian Journal of Economics / Revue Canadienne D'Economique, 1*(1), 37-54. https://doi.org/10.2307/133460

54. Oleshchenko, L. (2021). *Technologies for processing Big Data.* Igor Sikorsky KPI. https://ela.kpi.ua/server/api/core/bitstreams/dedcb0bb-b3b2-46d7-98b4-6977fd4f8628/content

55. Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10),* 1320-1326. https://doi.org/10.17148/IJARCCE.2016.51274

56. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval, 2*, 1-135. https://doi.org/10.1561/1500000011

57. Pantielieieva, N., Krynytsia, S., Zhezherun, Y., Rebryk, M., & Potapenko, L. (2018a). Digitization of the economy of Ukraine: Strategic challenges and implementation technologies. *Proceedings of the 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT 2018),* 508-515. https://doi.org/10.1109/DESSERT.2018.8409186

58. Pantielieieva, N., Krynytsia, S., Khutorna, M., & Potapenko, L. (2018b). FinTech, Transformation of Financial Intermediation and Financial Stability. *International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T 2018 - Proceedings*, 553–559. https://doi.org/10.1109/INFOCOMMST.2018.8632068

59. Potrimba, P. (2022, December 16). What is Semi-Supervised Learning? *Roboflow.* https://blog.roboflow.com/what-is-semi-supervised-learning

60. Reports on the implementation of the budget of the city of Cherkasy (2021-2023). https://chmr.gov.ua/ua/text.php?s=33&s1=368&s2=437

61. Reports on the implementation of the budget of the city of Kyiv (2021-2023). https://kyivcity.gov.ua/publichna_informatsiia_Tag_166122/

62. Sathi, A. (2013). *Big Data Analytics, Disruptive Technologies for Changing the Game*. 2nd Edition, MC Press Online, 73.

63. Shybalkina, I. (2021). Toward a Positive Theory of Public Participation in Government: Variations in New York City's Participatory Budgeting. *Public Administration, 100.* https://doi.org/10.1111/padm.12754

64. Sjouwerman, S. (2020, October 1). How Social Media Manipulation Threatens Your Business — And What You Can Do About It. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2020/10/01/how-social-media-manipulation-threatens-your-business--and-what-you-can-do-about-it

65. Smart Tender. (2022). Prozorro summary and main system changes for 2021. https://smarttender.biz/blog/view/pidsumki-roboti-prozorro-ta-golovni-zmini-u-sistemi-za-2021-rik/

66. Social Media & User-Generated Content. (2023). Statista. https://www.statista.com/markets/424/topic/540/social-media-user-generated-content/#overview

67. Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics*. Packt Publishing Ltd.

68. Stuart, A., & Ord, K. (1994). *Kendall's Advanced Theory of Statistics*. Edward Arnold.

69. Territorial Communities. (2024). https://decentralization.ua/newgromada

70. Tiebout, Ch. (1956). A pure theory of local expenditures. *Journal of Political Economy, 64*(5), 416–424. http://www.jstor.org/stable/1826343?origin=JSTOR-pdf

71. Trinder, B. (2019). Big Data and Financial Ethics: The Significant Capabilities of Artificial Intelligence Necessitate Human Guidance and Input. *Seven Pillars Institute Moral Cents, 8*(1), 25-30. https://sevenpillarsinstitute.org/wp-content/uploads/2019/05/Big-Data-Finance-Ethics-ED.pdf

72. Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020*). Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media.

73. Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2015). From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing. *Journal of Retailing, 91*(2), 174-181. https://doi.org/10.1016/j.jretai.2015.02.005

74. Wadhwani, S. (2022, July 6). Meta Files Two Lawsuits Over Illicit Data Scraping from Facebook and Instagram. *Spiceworks*. https://www.spiceworks.com/tech/tech-general/news/meta-sues-for-data-scraping/

75. Weiss, S. M., & Indurkhya, N. (1998). *Predictive data mining: A practical guide*. Morgan Kaufmann Publishers.

76. Whittaker, Z. (2022, April 18). Web scraping is legal, US appeals court reaffirms. *TechCrunch*. https://techcrunch.com/2022/04/18/web-scraping-legal-court/

DOI: 10.55643/fcaptp.3.56.2024.4402

77. Wu, Sh., Wang, N., & Wang, K. (2022). Internet Financial Risk Management in the Context of Big Data and Artificial Intelligence. *Mathematical Problems in Engineering*, 1024. https://doi.org/10.1155/2022/6219489

78. Zhang, Yahong, & Liao, Yuguo. (2011). Participatory Budgeting in Local Government. *Public Performance & Management Review, 35*, 281-302. https://doi.org/10.2753/PMR1530-9576350203

Криниця С., Гордей О., Коваленко Ю., Данькевич А., Болдов А.

ВИКОРИСТАННЯ ТЕХНОЛОГІЙ BIG DATA ДЛЯ ПОСИЛЕННЯ УЧАСТІ ГРОМАДСЬКОСТІ В УПРАВЛІННІ ПУБЛІЧНИМИ ФІНАНСАМИ

Стаття присвячена актуальним питанням упровадження технологій Big Data в управління публічними фінансами. Застосування Big Data має потенціал підвищення прозорості та відповідальності у використанні бюджетних коштів, зростання довіри до влади, удосконалення ефективності використання ресурсів бюджету, кращого розуміння потреб громадян і залучення громадськості до управління публічними фінансами. Метою дослідження є опрацювання теоретико-методичних і практичних аспектів, а також розроблення рекомендацій з упровадження технологій обробки та аналізу Big Data з метою посилення участі громадськості в управлінні публічними фінансами. Досліджено традиційні методи залучення громадськості до бюджетного процесу, виявлено їхні недоліки та потенціал технологій Big Data, базованих на прийомах комп'ютерної лінгвістики й машинного навчання, до посилення громадської участі. Розробки в царині аналізу настроїв та аналізу думок адаптовано до сфери публічних фінансів. Побудовано й апробовано генеративну модель аналізу настроїв громадськості в соціальних мережах стосовно управління публічними фінансами. Вироблені підходи щодо використання технологій Big Data можуть бути імплементовані в царину публічних фінансів із метою посилення громадської участі в управлінні ними як дорадчі інструменти реалізації представницької демократії та потребують подальшого теоретичного опрацювання й практичного застосування з метою поліпшення аналізу альтернативних думок, запобігання маніпуляціям суспільною думкою та зловживанням у мережі.

Ключові слова: Big Data, великі дані, публічні фінанси, бюджет, громадська участь, аналіз настроїв, аналіз думок, машинне навчання, соціальні мережі

JEL Класифікація: H30, H56, H72, C55, D70