№ 13(27)
2023
НАУКА і ТЕХНІКА СЬОГОДНІ
серії: право, економіка, педагогіка, техніка, фізико-математичні науки

СЕРІЯ «Комп'ютерні науки»

UDC 316.77:004

https://doi.org/10.52058/2786-6025-2023-13(27)-897-908

Mints Oleksiy Yurijovych, Doctor of Science in Economy, Professor, Technical university Metinvest Polytechnic, Ukraine, Zaporizhzhia, 80 Southern Highway, 69008, tel.: (067) 7604954, http://orcid.org/0000-0002-8032-005X

INFORMATION MANIPULATION AND MANUFACTURED CONSENSUS IN SOCIAL MEDIA: DETECTION METHODS AND STRATEGIES

Abstract. This article addresses the issue of information manipulation and the deliberate creation of manufactured consensus within the dynamic landscape of social media. It examines the motivations and tactics employed by various actors, including government entities, political groups, commercial enterprises, and individuals, who engage in manipulative practices. The article explores the nature of artificial media activity, encompassing the use of bots, coordinated behavior patterns, manipulation of content, and trending topics.

The research investigates the specific characteristics of manufactured consensus formation across various social media platforms, including YouTube, Facebook, and Telegram. It proposes an approach for identifying and analyzing coordinated activity, drawing upon a combination of quantitative and qualitative methodologies.

A case study based on YouTube comments is presented to illustrate the practical application of detection methods for coordinated media activity. A scenario developed in Orange Data Mining allowed for the identification of such indicators of coordinated activity as bursts, posting of similar comments, changes in the emotional tone of comments, and time-based coordination. The analysis reveals that even neutral content with a relatively low audience can become a target for coordinated manipulative efforts, highlighting the scale of the problem.

The findings of this study underscore the critical need for continued research and development of sophisticated algorithms and methods for detecting and mitigating the impact of artificial and coordinated media activity. The spread of manipulation undermines trust in information in the digital space, making it difficult to distinguish

reliable data from manipulative content. Manufactured consensus contributes to the polarization of society, amplifying extreme viewpoints and suppressing moderate voices. Addressing this challenge requires a comprehensive approach, encompassing scientific, technological, educational, and legislative measures.

Keywords: social media, manipulation, propaganda, bot detection, coordinated activity, machine learning, natural language processing, manufactured consensus, information warfare, data mining.

Мінц Олексій Юрійович, доктор економічних наук, професор, Технічний університет Метінвест Політехніка, Україна, Запоріжжя, Південне шосе, 80, 69008, тел.: (067)7604954, http://orcid.org/0000-0002-8032-005X

МАНІПУЛЯЦІЯ ІНФОРМАЦІЄЮ ТА ШТУЧНИЙ КОНСЕНСУС В СОЦІАЛЬНИХ МЕДІА: МЕТОДИ ТА СТРАТЕГІЇ ВИЯВЛЕННЯ

Анотація. Ця стаття зосереджена на питанні маніпуляції інформацією та навмисного створення штучного консенсусу в динамічному ландшафті соціальних медіа. Вона досліджує мотиви та тактики, які застосовують різні актори, включаючи державні органи, політичні групи, комерційні підприємства та окремі особи, котрі займаються маніпулятивними практиками. Стаття розглядає характер штучної діяльності в медіа, що охоплює використання ботів, координовані моделі поведінки, маніпуляцію контентом та трендові теми.

Дослідження зосереджується на конкретних характеристиках формування штучної згоди на різних платформах соціальних медіа, таких як YouTube, Facebook та Telegram. Воно пропонує підхід для ідентифікації та аналізу координованої діяльності, використовуючи комбінацію кількісних та якісних методологій.

Проведено тематичне дослідження на основі коментарів у YouTube, яке демонструє практичне застосування запропонованих методів виявлення координованої медіа діяльності. Сценарій, що розроблений в Orange Data Mining, дозволив виявити такі показники координованої діяльності, як сплески активності, публікація схожих коментарів, зміни в емоційному тоні коментарів та часова координація. Аналіз результатів показав, що навіть нейтральний контент із відносно невеликою аудиторією може стати мішенню для координованих маніпулятивних зусиль, підкреслюючи масштаб проблеми.

Висновки підкреслюють критичну потребу в продовженні досліджень та розробці складних алгоритмів і методів для виявлення та пом'якшення впливу штучної та координованої медіа діяльності. Розповсюдження маніпуляцій підриває довіру до інформації у цифровому просторі, ускладнюючи відрізне-

ння надійних даних від маніпулятивного контенту. Штучний консенсус сприяє поляризації суспільства, посилюючи екстремальні погляди і придушуючи помірні голоси. Вирішення цього виклику вимагає комплексного підходу, що включає наукові, технологічні, освітні та законодавчі заходи.

Ключові слова: соціальні медіа, маніпуляція, пропаганда, виявлення ботів, координована діяльність, машинне навчання, обробка природної мови, штучний консенсус, інформаційна війна, дата майнінг.

Problem Statement: In the contemporary information society, digital media play a pivotal role in shaping public opinion. However, alongside positive aspects such as simplified, cheaper, and faster access to information, serious challenges emerge related to the manipulation of the information landscape and the formation of manufactured consensus [1]. This phenomenon presents an illusion of unanimous agreement on a specific issue, created through the deliberate use of technology and coordinated activities within digital media. Instead of the organic development of public opinion based on the free exchange of information and open discussion, manufactured consensus is imposed externally, creating a distorted representation of reality.

Manufactured consensus is formed through various mechanisms, including the use of bots and fake accounts to simulate widespread support, the artificial inflation of activity metrics (likes, comments, reposts), the dissemination of disinformation, and the targeted suppression of dissenting voices. The mechanism by which public consensus (or rather, its subjective perception) influences individual decision-making is grounded in the theory of reflexive control [2].

Recent global examples of coordinated media activity employed to create manufactured consensus include the 2016 US presidential election and the Anti-Vaccine Movement during the COVID-19 pandemic [1]. Similar tactics are utilized by the Russian Federation to justify its military actions [3]. The danger of these actions lies in their potential to distort the accurate perception of public opinion, erode trust in information within the digital realm, and contribute to the polarization of society [4].

Therefore, an actual task is to develop and investigate the phenomenon of information manipulation in social media, as well as to develop methods and tools for detecting actions aimed at forming manufactured consensus.

Analysis of Current Research and Publications: The topic of information manipulation and its detection is a subject of study for numerous scholars. Theoretical and practical aspects of this problem are examined by V. Lefebvre, S. Woolley, A. Vasara, D. Lazer, E. Hargittai and others. However, the constantly evolving media landscape, along with the emergence of new factors such as generative artificial intelligence, necessitates further research.

The aim of the study: This article aims to identify the characteristics of manufactured consensus formation in social media and propose methods for its detection.

Main results. The formation of coordinated and artificial media activity aimed at creating manufactured consensus in the contemporary information landscape is a complex process involving diverse actors pursuing various goals. Identifying these actors and understanding their motivations is crucial for effectively countering manipulation and disinformation.

The motivations behind actors initiating information manipulation in social media can be political, economic, or ideological. While this list is not exhaustive, it encompasses the primary drivers.

Political motivation for artificial media activity is characteristic of government entities seeking to cultivate a positive image, promote their political agenda, or discredit opposition. Intelligence agencies may also employ these methods within the framework of information warfare to destabilize situations in other countries. Conversely, artificial media activity, often based on the use of bots and algorithms, can be initiated by both state actors to enhance control over the information space and by private individuals or groups to achieve their own objectives. Overall, political actors, such as political parties and social movements, frequently resort to methods of manufacturing consensus to advance their ideas and mobilize supporters. The 2016 US presidential election serves as an example of artificial media activity driven by political motivation.

Artificial media activity stemming from *ideological motivation* aims to promote a particular belief system, worldview, or set of values, often with the goal of shaping social norms and influencing cultural discourse. Media platforms are utilized to disseminate ideological content, attract supporters, and build communities around shared beliefs. This can involve promoting specific narratives, demonizing opposing ideologies, and reinforcing group identity. For instance, a religious group might use social media to spread its beliefs and recruit new members, or an activist group might leverage online platforms to challenge dominant ideologies. The aforementioned Anti-Vaccine Movement exemplifies artificial media activity driven by ideological motivation.

Economic motivation underpins the media activities of commercial organizations, including corporations and PR agencies, which employ these methods to promote goods and services, cultivate a positive brand image, and engage in competitive struggles. In this context, artificial media activity can manifest in boosting popularity metrics and generating fake reviews and comments.

Individual actors, such as bloggers, influencers, and activists, should also be considered, as they can leverage their popularity and influence to promote specific ideas and foster manufactured consensus. Their motivations, in addition to those discussed above, might be rooted in the pursuit of personal fame or recognition.

Individual actors can engage in coordinated actions with their supporters or utilize bots and other tools of artificial media activity.

One aspect of analyzing and identifying artificial media activity is the number of accounts involved. This factor directly reflects the actor's influence and can serve as a supplementary indicator for identification.

Fig. 1 illustrates the connections between different types of information manipulation and their initiators.

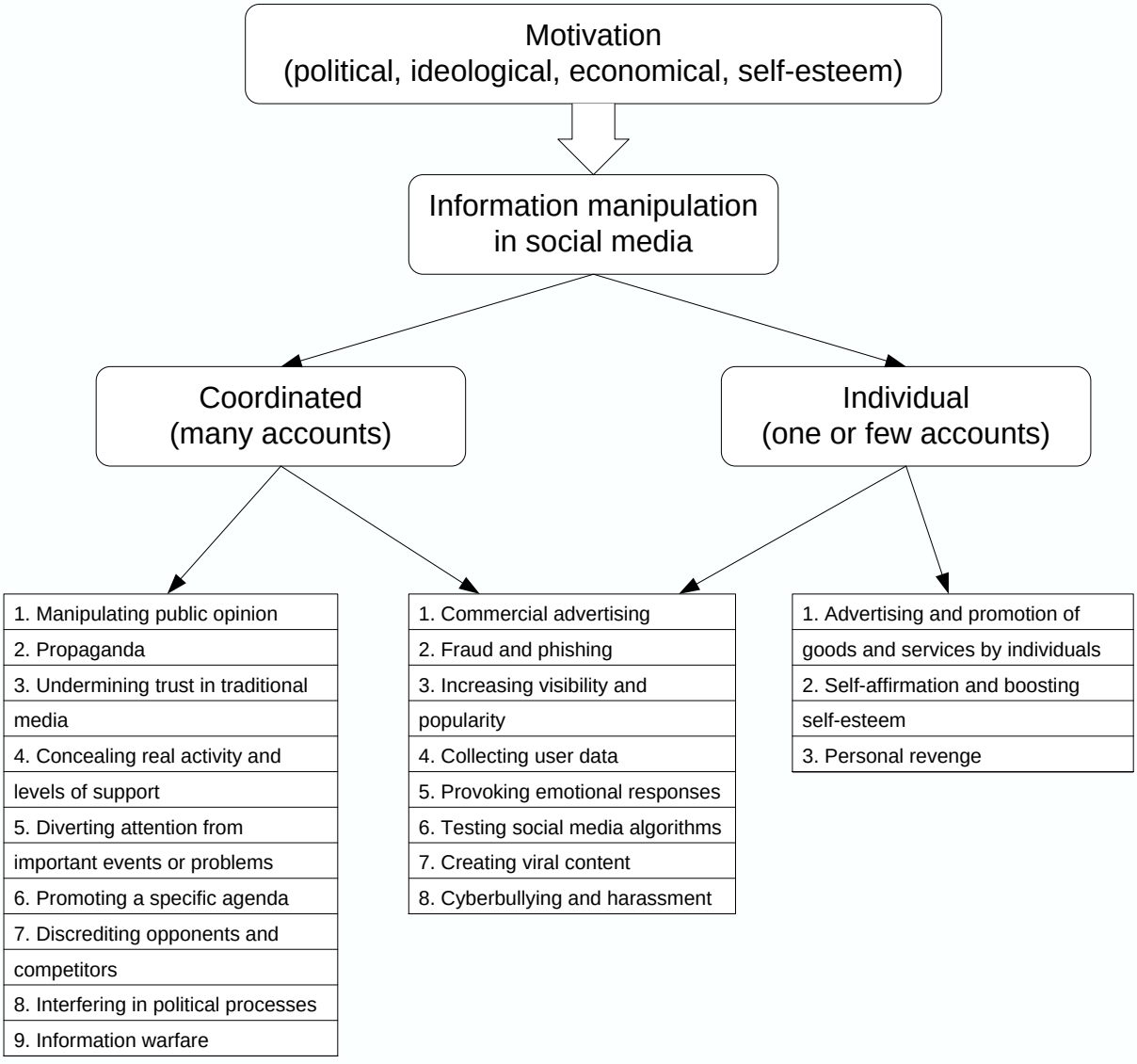

Fig. 1. Main types of information manipulation and their initiators.

It's important to understand that the boundaries between these categories can be blurred, and a single actor may pursue multiple goals simultaneously. However, Fig.1 clearly demonstrates that coordinated activity based on political or ideological motivations requires the involvement of the largest number of accounts.

Accounts that engage in coordinated media activity are often referred to as "bots". Initially, this word referred to a program designed to perform specific tasks in the digital space without direct human intervention. Software bots can be used to automate routine operations, gather information, and interact with users. However, they can also mimic the actions of a real user and be employed to manipulate the information environment and public opinion. Recently, the term "bot" has increasingly been used to describe a person who behaves like an automated program. This typically implies that the person acts mechanically, repeats the same actions, or follows instructions without critical thinking. Distinguishing between the actions of such a person and the operation of a program (especially one that utilizes generative artificial intelligence) in the media landscape is practically impossible, which has led to the widespread adoption of this terminology.

The increasing sophistication of "bots" (both human and AI-powered) has simultaneously made the task of identifying them more complex. The main indicators of artificial media activity are anomalies that can be observed in user behavior and information dissemination.

Among the anomalies in user behavior, the following can be highlighted:

- Coordinated bursts of activity, identical posting times, repetitive actions, and other unnatural activity patterns.
- Use of identical (or similar) content by different users, promoting the same narratives.
- Characteristic user profile appearance: minimal personal information, stock photos, recently created accounts, generated nicknames.
- Unusual connections between accounts involved in coordinated actions (mutual subscriptions, reposts, likes, comments).

Among the anomalies in information dissemination, the following can be highlighted:

- Artificial amplification of content through boosting likes, reposts, and comments.
- Irrelevant posting (e.g., spreading political narratives in comments to neutral posts, ignoring their topic and discussion history).
- Purposeful dissemination of disinformation content.
- Trend manipulation (supporting trends favorable to the actor and blocking or removing all others).

Each of these indicators alone does not guarantee that an account is a "bot," but their combination increases the likelihood. It's also important to note that different types of information manipulation correspond to different sets of anomalies and, consequently, different detection methods.

It should be considered that coordinated activity can be carried out from two sides – content creators and consumers (Fig.2).

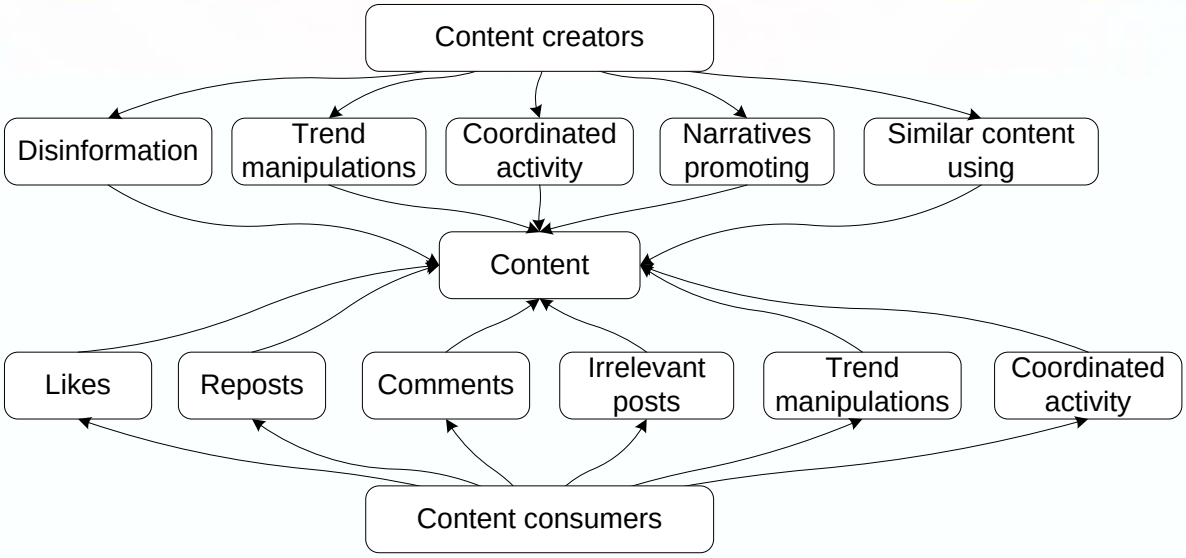

Fig. 2. Indicators of content manipulation from different sides.

Content creators are potentially more influential, as each post is viewed by a large number of followers. However, developing an account to achieve a significant level of influence is a complex and time-consuming task. On the other hand, individual content consumers (readers, followers) have limited influence on their own. However, a follower account doesn't require prior development, making it much easier to create and use for boosting likes, reposts, comments, and other engagement with content. In this case, the desired effect of coordinated activity is achieved by involving a large number of content consumers' accounts. The greatest impact is observed when coordinated activity combines efforts from both content creators and consumers.

It's worth noting that the number of influential accounts in media is limited due to the aforementioned difficulty in developing an account to a sufficient level of influence. This allows for the use of expert methods to analyze the activity of such accounts, which yield relatively reliable results. Therefore, a more challenging task, and consequently one of greater scientific interest, is identifying coordinated media activity originating from content consumers. This is particularly relevant because they are subconsciously perceived by other content consumers as "the people," making them ideally suited for manufacturing consensus when used effectively.

The complexity of identifying coordinated media activity from content consumers in social media stems from the vast volumes of generated data, the rapid speed of information dissemination, and the diverse formats of content. For example, the social network X (ex. Twitter) alone generates 80-100 gigabytes of textual data annually. This necessitates the application of various data analysis methods and technologies, including Big Data Analysis, Machine Learning, Natural Language Processing, and Network Analysis. Moreover, different types of media platforms have unique characteristics in terms of how coordinated activity manifests [5].

For video hosting platforms like YouTube and others primarily based on visual content, and where user interconnectivity is weak, the primary form of manipulation aimed at manufacturing consensus is coordinated activity in comments.

For social networks where user interconnectivity is central (Facebook, Instagram), reposts, likes, and other engagement indicators are more significant, although comments also play a role.

Finally, on platforms based on textual content (Reddit, Discord channels, Telegram), activity in comments is practically the only, yet highly effective, tool for manufacturing consensus.

As evident, comment activity is the primary form of coordinated media activity from content consumers. Let's examine potential approaches to identifying such activity using a real-world example. The example chosen is a set of comments on a YouTube video [6]. The source video contains neutral footage of the city of Mariupol just before the start of the military invasion, without any voiceover or other provocative insertions. The total number of views is approximately 50,000.

When working with real-world data, data analysis is typically preceded by data processing [7]. In this case, since most universal data analysis tools cannot directly work with media resources, data parsing is necessary, which involves creating a local copy of the data. In the example considered, the local data copy comprises 1,270 comments made from June 2022 to November 2023.

For further data analysis, the freely available software Orange Data Mining was used. The program implements a visual approach to constructing a data processing scenario. The constructed scenario is shown in Fig. 3.

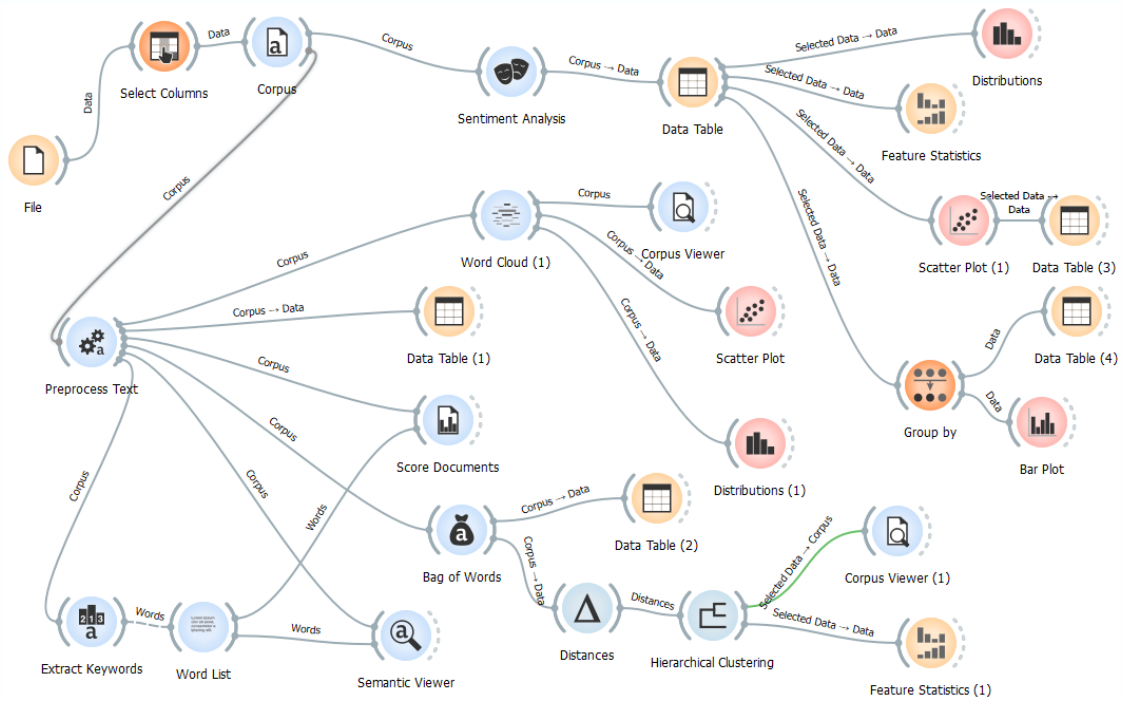

Fig. 3. Scenario for detecting coordinated media activity in YouTube comments.

This scenario allows for the identification of the following indicators of coordinated activity:

- Bursts of activity.
- Posting of similar comments.
- Emotional tone of comments and its dynamics.
- Time-coordinated commenting.

The dynamics of posting activity are presented in Fig. 4.

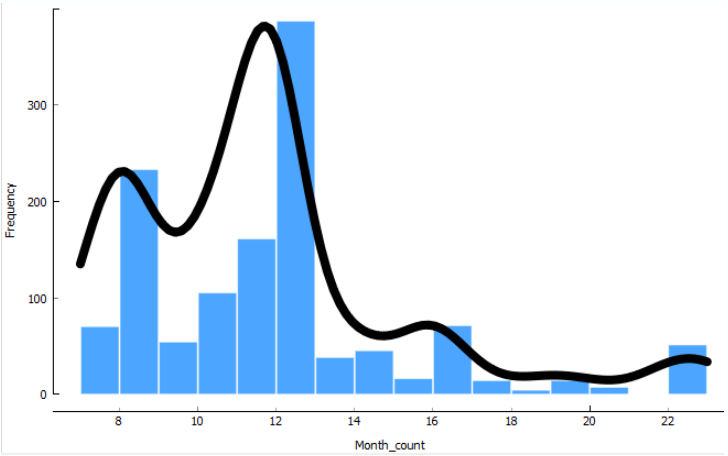

Fig. 4. The dynamics of posting activity by months

As evident from the analysis of Fig. 4, the number of comments is varying. Two distinct bursts are observed, followed by a significant decline. However, in this case, it's essential to consider YouTube's algorithms for internal video promotion, for which such a graph can be considered normal. An anomaly on this platform would be a uniform graph or a long-term graph with constant growth.

The posting of similar comments in the scenario under consideration is tracked using hierarchical clustering tools based on calculated Euclidean distances between

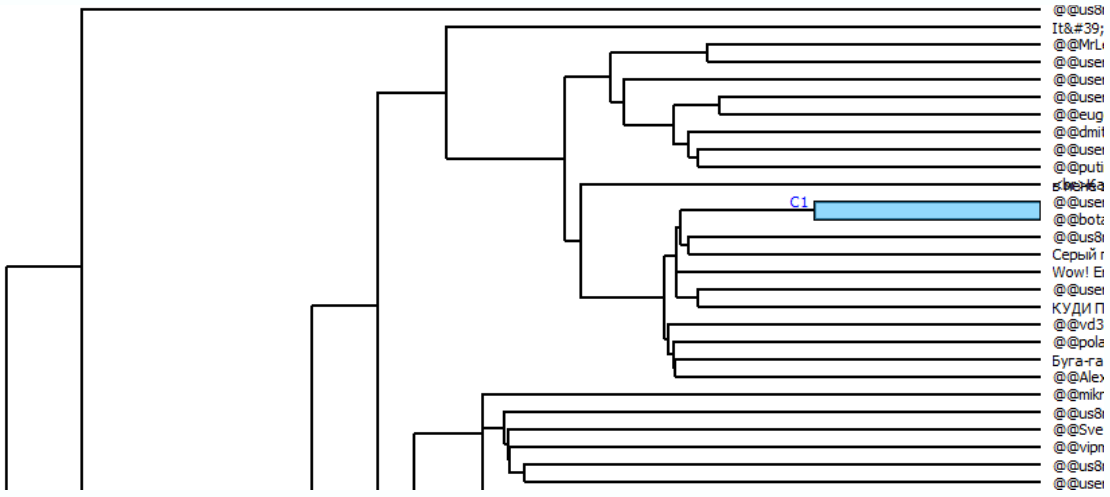

Fig. 5. The results of hierarchical clustering of comments.

The hierarchical clustering method is resource-intensive but allows for effectively finding texts containing similar or identical fragments. Fig. 6 shows the content of the cluster highlighted in Fig. 5. This can be identified as covert advertising for a certain YouTube channel and attributed to coordinated activity since it was done by different users using identical expressions at the same time. Such advertising also exemplifies the formation of manufactured consensus as it creates an illusion of unanimous recommendations for other users.

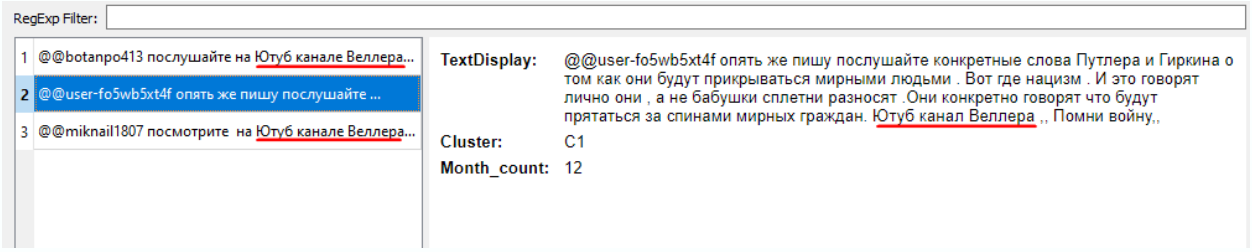

Fig. 6. Content of the cluster C1.

Analyzing the emotional tone of comments is also helpful for detecting and identifying coordinated activity. Obviously, in the absence of overt interference, the average level of emotions shouldn't fluctuate significantly over time. However, several pronounced anomalies are observed in the data under examination (Fig. 7).

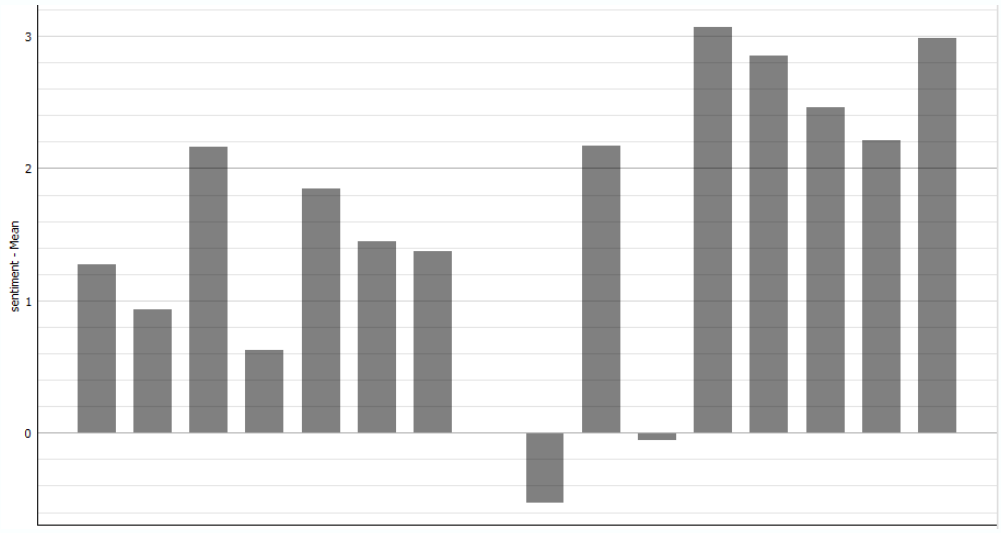

Fig. 7. Dynamics of change in emotional tone of comments by month

The anomalies observed in Fig. 7 include several drops in emotional tone to zero or even negative values. However, comparing this with the total number of comments (Fig. 4) reveals that the overall activity was low during those months. Therefore, the anomalous level may be due to random fluctuations. At the same time, the graph shows a steady increase in the average level of emotional tone over the

period under consideration, by approximately three times. This could indicate, if not coordinated activity itself, then at least its results. An analysis of the comments themselves shows that towards the end of the period, narratives about "how good things have become in the City" began to prevail.

Based on the analysis, it can be concluded that there is coordinated activity in the comments, but in limited quantities. This is more evident towards the end of the period under consideration when the video received a sufficient number of views to attract the attention of "bots."

Conclusion. The problem of manufacturing consensus poses a serious threat to democratic processes and the functioning of the information society. This technology can be used to manipulate elections and influence decision-making. The propagation of manipulations undermines trust in information within the digital space, making it difficult to distinguish reliable data from manipulative content. Manufactured consensus contributes to the polarization of society by amplifying extreme viewpoints and suppressing moderate voices.

Addressing this problem requires a comprehensive approach that encompasses scientific, technological, educational, and legislative measures. Scientific research in this area is crucial for understanding the mechanisms behind the formation of manufactured consensus and developing effective counterstrategies.

This article examines various types of information manipulation, identifies indicators of artificial media activity, and explores how these activities can manifest from both content creators and consumers. It proposes approaches to detect coordinated activity and develops and tests a model for its identification using the YouTube platform as an example. The conducted research, based on real-world data, demonstrates that even neutral content with a relatively low audience can become a target for coordinated activity.

Further development of this research topic should include refining algorithms for detecting bots and fake accounts, enhancing media literacy among users, and developing legislative norms to regulate activity in the digital space.

References:

1. Woolley, S. (2023). Manufacturing consensus: Understanding propaganda in the era of automation and anonymity. Yale University Press.

2. Lefebvre, V. (2015). *Conflicting structures*. Lulu. com.

3. Vasara, A. (2020). Theory of reflexive control: origins, evolution and application in the framework of contemporary Russian military strategy. National Defence University.

4. Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

5. Hargittai, E. (2015). Research methods in social media. Polity Press.

6. Mariupol. What it looked like a month before the war (2022). Retrieved from https://www.youtube.com/watch?v=i5K8MVq0MSM

7. Mints, A. (2017). Classification of tasks of data mining and data processing in the economy. *Baltic Journal of Economic Studies*, *3*(3), 47-52. https://doi.org/10.30525/2256-0742/2017-3-3-47-52

Література:

1. Woolley, S. (2023). Manufacturing consensus: Understanding propaganda in the era of automation and anonymity. Yale University Press.

2. Lefebvre, V. (2015). *Conflicting structures*. Lulu. com.

3. Vasara, A. (2020). Theory of reflexive control: origins, evolution and application in the framework of contemporary Russian military strategy. National Defence University.

4. Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

5. Hargittai, E. (2015). Research methods in social media. Polity Press.

6. Mariupol. What it looked like a month before the war (2022). Retrieved from https://www.youtube.com/watch?v=i5K8MVq0MSM

7. Mints, A. (2017). Classification of tasks of data mining and data processing in the economy. *Baltic Journal of Economic Studies*, *3*(3), 47-52. https://doi.org/10.30525/2256-0742/2017-3-3-47-52