UDC 004.93
DOI: 10.15587/2706-5448.2025.344630

Heorhii Chyzhmak,
Valeriy Sydorenko

DEVELOPMENT OF CLUSTERING MODELS FOR EXTENDED OPINION HOLDERS BASED ON AGGREGATED STYLOMETRIC AND SENTIMENT FEATURES OF CHAT MESSAGES

The subject of research is the methods and technologies for monitoring holder opinion groups in social media based on stylometric and sentiment features. One of the most important problems is the increasing complexity of text content, which makes user behavior analysis more difficult because of anonymity, informal language, slang, emojis, and non-standard writing styles. Stable, long-term behavioral patterns are not captured by methods based on single-message evaluation.

This research proposes a holder-level clustering method based on aggregated stylometric and sentiment features taken from several messages per user. The methodology includes agglomerative hierarchical clustering, which is enhanced by decision tree analysis for feature selection and cluster interpretability, quantile normalization, dimensionality reduction via PCA (LiveJournal provided six components explaining 81.7% of the variance, while Instagram provided four components explaining 83.5% of the variance), and data preprocessing (VarianceThreshold, removal of highly correlated features). Ultimately, the majority of users were covered by two clusters for LiveJournal and three clusters for Instagram. The result is a set of clustering models that efficiently group holders into logical, understandable clusters based on their overall communication style and emotional expression. The primary advantages of the proposed approach are as follows: holder-level aggregation ensures stability and consistency in profiling; two-stage clustering with intermediate feature selection enhances explainability; the method demonstrates cross-platform applicability, validated on both LiveJournal and Instagram. As a result, over time, more accurate and dynamic user profiles can be developed, enabling improved sentiment analysis, automated moderation, and customized user interaction. This approach offers significant benefits over conventional single-message analysis methods in terms of results transparency, behavioral insight depth, and profile stability. Customized social media recommendations, automated moderation, and social sentiment analysis can all benefit from the study's findings.

Keywords: *clustering models, natural language processing, semantic and sentiment analysis, explainable artificial intelligence.*

Received: 25.08.2025
Received in revised form: 06.11.2025
Accepted: 25.11.2025
Published: 29.12.2025

How to cite
Chyzhmak, H., Sydorenko, V. (2025). Development of clustering models for extended opinion holders based on aggregated stylometric and sentiment features of chat messages. Technology Audit and Production Reserves, 6 (2 (86)), 23–30. https://doi.org/10.15587/2706-5448.2025.344630

1. Introduction

The widespread use of social media and chat platforms in today's digital era has fundamentally changed the nature of communication, in particular for service providers within such industries as electricity and transportation. The complexity of textual content [1, 2] and anonymity of content creators pose a challenge to effective communication and timely feedback.

Stylometric features are in widespread use in text content analysis studies. The seminal work [3] showed that frequency-based linguistic characteristics, mainly the distribution of common function words, may reliably support authorship attribution. The subsequent surveys, for instance the review [4], confirm the effectiveness of stylometric methods for modern tasks of authorship analysis, including those related to online communication.

In fact, studies on social media further extend those ideas, such as by author profiling research. For instance, the PAN Author Profiling tasks [5] showed that the combination of stylometric and behavioral features supports the uncovering of stable user patterns in short and informal texts.

Recent research has pointed out the importance of text aggregation at the user level. In [6] demonstrated that a user-level representation shows substantially better performance compared to relying on standalone messages. Our approach further develops this line by adding together stylometric and sentiment features in a cross-platform setting, which enables to reveal more stable behavioral patterns.

While much progress has been made, there is a deficiency in stylometric analysis integrated with sentiment analysis for multi-platform data. Our work fills this gap by proposing a holder-level clustering approach that combines interpretability and practical applicability.

In research [7], the authors propose to evaluate the "holder pattern" [2] and "sentiment pattern" [8] as part of the evaluation of the "extended opinion" cortege [2] on the level of an opinion, i. e. base on one message. This article attempts to take a view of this issue from another perspective. Firstly, the authors expand the set of behavior features to add new characteristics. Secondly, clustering holders were made at the level of a holder based on aggregated stylometric and sentiment features extracted from a number of messages. This approach extends the single message evaluations, allowing the detection of more stable and consistent holder patterns that enable the identification of clusters of holders who share similar behavioral and emotional characteristics. Thirdly, the experiment was conducted on two distinct social media platforms: LiveJournal [9] and Instagram [10].

It also presents an overview of reflective and in-depth discussions that are typical of LiveJournal, while comments on Instagram tend to be emotional, brief, and highly interspersed with images. In this regard, the present research focuses on the respective patterns of communication deployed in both platforms through user behavior. This enhances understanding of how the basic characteristics of the platforms shape user behavior and expression of sentiment, showcasing the versatility and robustness of our clustering approach across varied social media landscapes.

Thus, *the object of research* is the process of monitoring holder opinion groups in social media based on stylometric and sentiment features. *The aim of research* is to create clustering models for extended opinion holders using aggregated stylometric and sentiment features of chat messages. This is to improve the existing target group finding techniques on social media.

To achieve the set aim, the following tasks must be completed:

– identify and extract sentiment-based and stylometric features from chat messages on various social media platforms using natural language processing (NLP) techniques;

– develop a method for integrating message-level attributes into trustworthy, representative user-level profiles;

– produce interpretable user clusters in line with XAI principles [11];

– confirm the clustering models' generalizability using datasets from multiple platforms (LiveJournal and Instagram).

2. Materials and Methods

2.1. Comparative analysis of holder's patterns in different social media chats

This paper sets out to investigate how stylometric and sentiment features are used in opinion group tracking on social media. A comparative study between Instagram and LiveJournal shows that users communicate and give their opinions using these sites in rather different ways. With a view to giving a full idea of how people communicate in their own special ways, this section investigates the various ways in which people behave and feel using these sites.

Let's consider the datasets of LiveJournal [9, 12] and Instagram [10, 13], which were obtained by web-scraping the comments on posts of popular writers within 2015–2019.

On LiveJournal platform, users are more active; send more messages per user, as shown in Table 1. Often, people come back to discussions after a while; thus, discussions last very long. Users are critically involved, in other words, they reflect seriously on posts and comments. Messages are longer and more complex; people often use ellipses, among other text patterns, as evidence that they broke off to think and continue their thoughts.

By contrast, on Instagram, there are only one or two short comments per post, and users post fewer messages. Positive reinforcement is the main means of engaging with others. Slang and emojis are used more in order to enhance visual expressiveness, and comments are more about answering promptly.

Sentiment and behavioral trends show that LiveJournal users are more critical and sarcastic, with roughly 34% of messages being negative. The platform allows for complex discussions because users are involved in multiple threads. In their more analytical writing, users employ more complex sentences and a larger vocabulary.

Positive messages make up about 76% of Instagram messages. The platform promotes quick, emotionally charged interactions with an emphasis on visual content. Emojis are used in comments by about 67% of users.

Behavior also is contingent upon typographic and device factors. The tendencies of LiveJournal users to use laptops or desktop computers result in relaxed writing styles and more detailed, well-structured comments. Instagram users are mostly accessing the site through mobile devices, so comments on that platform are much shorter and more direct. Auto-correction affects capitalization and name use, too.

Community interaction and overall activity are two very different things. The LiveJournal community engages in long discussions and regularly gives each other quite extensive critique. Instagram's activity, in contrast to LiveJournal, is fast and superficial; less in-depth communication and fewer threaded conversations take place.

These variations offer insight into the development of tailor-made communication strategies, and increase user engagement by showing how features of each platform drive user behavior and communication patterns.

Table 1

Comparative analysis between LiveJournal and Instagram

Characteristic	Characteristic	LiveJournal	Instagram
Num. (#) of Holders	Num. (#) of Holders	18,003	36,677
Total Messages (Msgs)	Total Messages (Msgs)	81,115	53,414
# of Words per Msgs	# of Words per Msgs	Median: 14. Mean: 29.04	Median: 7. Mean: 11.82
Msgs with Ellipsis, %	Msgs with Ellipsis, %	18.92%	0.00%
Avg. Discussion (D) Duration, weeks	Avg. Discussion (D) Duration, weeks	~33	~3.3
Avg. Msgs/Holder (H)	Avg. Msgs/Holder (H)	248 users with ≥ 30 messages	326 users with ≥ 10 messages
Sarcastic Msgs, %	Sarcastic Msgs, %	16.01%	9.23%
Cons. Pos. H, %	Cons. Pos. H, %	48.35%	76.02%
Cons. Neg. H, %	Cons. Neg. H, %	33.48%	11.79%
Txt-only H, %	Txt-only H, %	65.18%	33.44%
Txt + Emoji H, %	Txt + Emoji H, %	34.40%	66.56%
Mss start Upper %	Mss start Upper, %	61.62%	89.18%
Avg. Msgs/Thread	Avg. Msgs/Thread	2.66	N/A
Avg. Msg Like Count	Avg. Msg Like Count	N/A	0.49
D Depth (levels)	D Depth (levels)	Mean: 1.22. Max: 71	N/A
Avg. Time Between Msgs	Avg. Time Between Msgs	Min: 20 s. Median: 17 m	Min: 0 s. Median: 33 m

In light of these variations, consider classifying opinion holders according to aggregated features encompassing all accessible messages from a single user within a particular social media platform rather than assessing the characteristics of individual messages within a single discussion.

2.2. Holder clustering levels in social media chats

In [2, 7], the authors proposed a model for clustering holders at the opinion level, i. e., assessing the behavioral and sentimental features of text messages with their subsequent processing and clustering. Opinion-level analysis allows for a quick assessment of the discussion itself, or in other words, the role of the holder at the level of a particular discussion.

In contrast, holder-level clustering is designed to study holders based on their writing and emotional expressions in an array of messages from different conversations. In other words, opinion-level and holder-level evaluations solve different problems. In this regard, it is natural that at the level of different discussions, holders can from time to time represent different clusters. This difference can be visualized by an actor playing different roles. Thus, clustering at the level of opinions is a grouping of holders (actors) by roles within a particular discussion, while clustering at the level of holders gives an idea of the holder (actor) itself by the totality of its roles. Thus, as stated in Section 1, the purpose of this paper is to build and study models of clustering opinion holders by the totality of their roles in different discussions, i. e. at the holder level. The following sections detail the methodology used to achieve this aim.

The raw data needs to be processed to determine the behavioral and sentiment patterns of opinion holders based on text messages. The components of these vectors can be divided into two categories:

1) components that can be calculated based on the components obtained at the opinion level;

2) those that are impractical to calculate at the opinion level and therefore are unique at the holder level.

1. The first category includes components obtained by aggregating the relevant components at the opinion level

$$\overline{h}_i = F(\{\hat{h}_{i,j}\}_{j=1}^{n}). \tag{1}$$

Let $\hat{h}_{i,j}$ be the value of component i at the opinion level for the j-th message. The number n represents the total number of messages from a single author. The function F aggregates these n values into a single representative value. This results in the component $\overline{h}_i$ at the holder level. Thus, F transforms message-level features into a user-level feature.

Each feature $\hat{h}_{i,j}$ is calculated at the opinion level from a single message. It represents a behavioral or sentiment characteristic of that message. To obtain a holder-level representation, these values are aggregated across all messages from one holder. The number of messages per holder is denoted by n. An aggregation function F is applied to combine these values. This results in a single value $\overline{h}_i$ at the holder level.

Thus, $\overline{h}_i$ reflects stable patterns over multiple interactions.

As metrics for evaluating the holder's writing style and sentiment pattern, the features proposed in [7] were used. Also, authors introduce several new features:

- Feature h_{24} (int) counts the number of times the "..." symbol appears in a message.
- Feature h_{25} (float) measures the time from the first message in a discussion to the current one, in hours.
- Feature h_{26} (bool) indicates whether a message is sarcastic, based on the sentiment pattern op_2 [8].
- Feature h_{27} (int) records the number of likes received by the message.
- Features $h_{28} - h_{36}$ (int) count consecutive smiles and emojis, derived from decomposed sentiment components [8].

The next step was to select the aggregation function F or a series of functions $\{F_k\}$, which best fits the peculiarities of the data. In this case, (1) can be generalized as follows

$$\overline{h}_i^{(k)} = F_k(\{\hat{h}_{i,j}\}_{j=1}^{n}). \tag{2}$$

Two aggregation functions were tested: the arithmetic mean and the median. Due to outliers, the arithmetic mean was chosen for its comprehensive representation of holder characteristics. Unlike the median is relatively insensitive to outliers. Thus, (1) is formulated as follows

$$\overline{h}_i = \frac{1}{n}\sum_{j=1}^{n}\hat{h}_{i,j}. \tag{3}$$

2. The second category includes components meaningful in the context of the entire chat history.

Feature h_{37} (int) represents the number of messages by a holder in a discussion (which may be indicative of the level of activity).

Feature h_{38} (int) records maximum discussion depth (may be considered a measure of the level of nesting of a chat for a particular holder).

The next phase involves data preparation for analysis and dimensionality reduction.

2.3. Preprocessing and dimensionality reduction

The investigation was conducted on two datasets: one based on LiveJournal and the other on Instagram. The primary data utilized were the holder ID and a set of attributes determined on the basis of chat messages $(h_1,\ldots,h_{38})$.

In the initial stage of the process, low-variable features were removed using the VarianceThreshold transformer [14] with a threshold value of 0.1. Following the transformation, 18 features remained for LiveJournal and 14 for Instagram.

The second step was to remove highly correlated features with a threshold of 0.985. At this stage, 18 features remained for LiveJournal and 12 for Instagram. The features left after these transformations are shown in Tables 2 and 3, respectively.

Table 2

PCA component weights for LiveJournal (preliminary)

Holder attributes	*P*0	*P*1	*P*2	*P*3	*P*4	*P*5
h_3	−0.16	−0.4	−0.46	0.353	−0.30	−0.47
h_4	−0.2	0.05	0.262	0.085	−0.04	0.032
h_5	−0.197	−0.01	0.23	0.088	0.092	−0.09
h_6	−0.196	0.02	0.251	0.079	0.004	−0.025
h_7	−0.52	−0.18	−0.09	−0.79	−0.21	0.026
h_8	−0.265	0.086	0.289	0.14	−0.05	0.021
h_9	−0.13	−0.12	−2E−3	0.038	−0.03	−0.1
h_{10}	−0.23	0.031	0.175	0.11	−0.03	−0.11
h_{11}	−0.199	0.054	0.264	0.105	−0.02	−0.05
h_{12}	−0.31	−0.01	0.181	0.14	0.131	−0.15
h_{13}	−0.37	0.762	−0.50	0.12	0.038	−0.002
h_{15}	−0.23	−0.17	−2E−3	0.16	−0.14	−0.20
h_{25}	−0.04	0.017	−0.02	0.004	−0.01	0.077
h_{24}	−0.25	−0.27	−0.17	0.33	−0.19	0.816
h_{21}	0.11	0.226	0.147	0.02	−0.39	−0.07
h_{22}	−0.14	−0.22	−0.24	−0.03	0.713	−0.02
h_{38}	−0.13	−0.05	0.122	0.10	0.25	−0.03
h_{37}	−0.12	0.014	0.073	0.08	0.219	0.053

Table 3

PCA component weights for Instagram (preliminary)

Holder attributes	$P0$	$P1$	$P2$	$P3$
h_1	–0.003	–0.481	0.112	0.312
h_3	0.651	0.011	0.42	–0.177
h_4	0.098	0.036	–0.462	0.182
h_5	0.13	–0.027	–0.373	0.12
h_9	0.27	–0.002	–0.285	0.147
h_{10}	0.114	–0.122	–0.236	0.129
h_{11}	0.108	0.01	–0.418	0.111
h_{19}	–0.04	–0.475	0.261	0.611
h_{15}	0.667	0.042	–0.0	0.113
h_{25}	0.031	0.029	–0.16	0.044
h_{37}	–0.025	0.001	0.06	–0.022
h_{33}	0.06	–0.723	–0.22	–0.62

The third step is to normalize the feature distributions [15] using the quantile method [16]. This approach proved to be more efficient than others, such as logarithmization [17] and Box-Cox transformation [18], since the dataset included a significant number of outliers from both social media.

Subsequently, the PCA method [19] was employed to reduce the dimensionality of the data, with the principal components selected to explain at least 80% of the variation in the data. This resulted in the identification of 6 components for the LiveJournal dataset (Table 2) and 4 for Instagram (Table 3), which, respectively, explain 81.7% and 83.5% of the total variance.

The final stage of data preparation entailed the application of the Agglomerative Hierarchical Clustering algorithm [20] for the initial segmentation of opinion holders and the selection of significant features through the use of decision trees.

2.4. Preliminary clustering of holders in the space of aggregated stylometric and sentimental patterns

The initial clustering results are presented in Fig. 1 and 2, respectively.

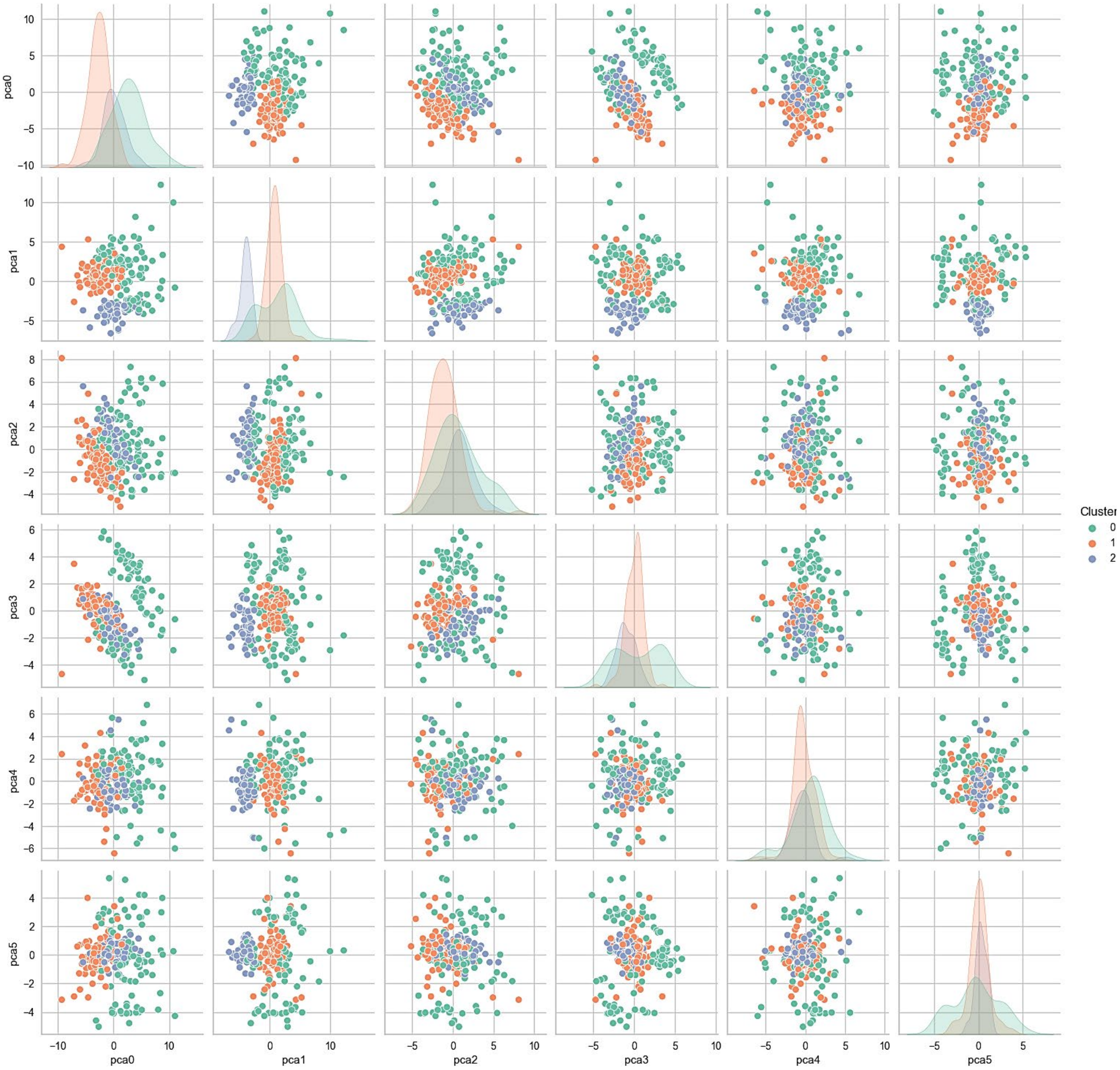

Fig. 1. Preliminary clusters for LiveJournal

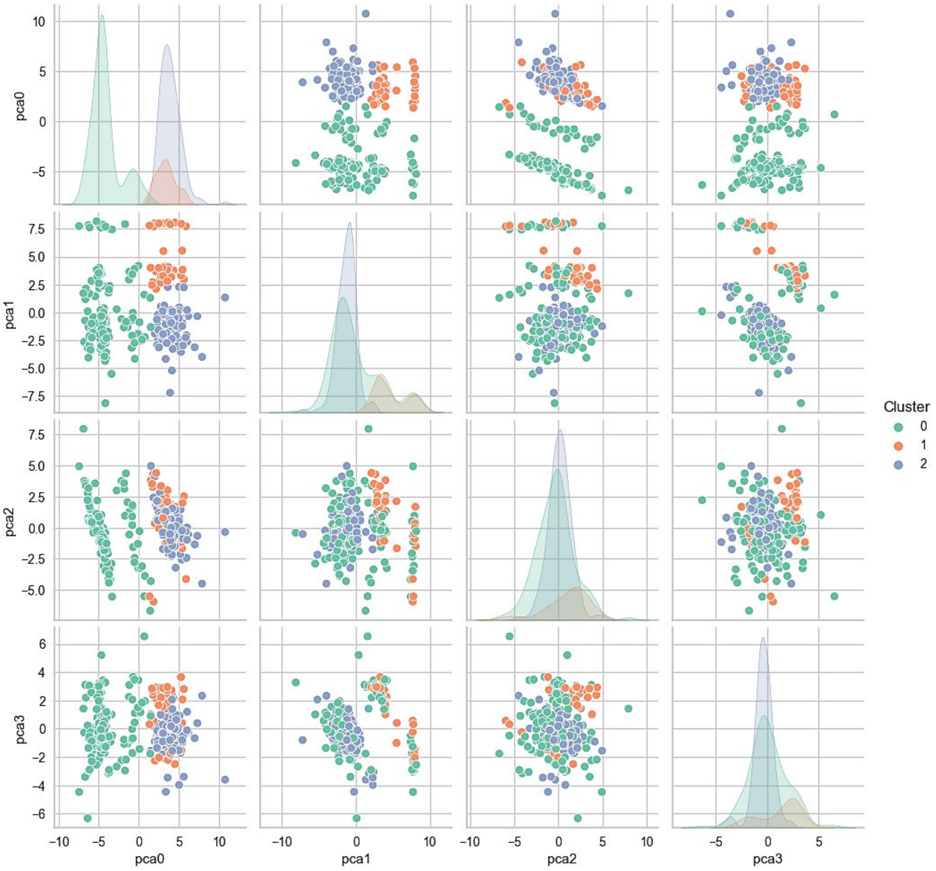

Fig. 2. Preliminary clusters for Instagram

Upon initial observation, it becomes evident that the holders do not form clearly delineated, homogeneous groups. However, subsequent application of the decision tree [13, 21] to the datasets based on the raw data and the obtained clusters yielded results that were relevant to those presented in the initial paper [7]. To facilitate the description of the obtained clusters, the following approach was proposed. Based on the obtained decision tree, the features that have an impact on predicting the holder class (threshold > 0.0) were selected. For the LiveJournal, the following features were identified: h_3, h_4, h_7, h_9, h_{13}, h_{24}. Similarly, Instagram has the following: h_2, h_{19}, h_{29}, h_{33}.

The selected features were used to repeat the entire pipeline, beginning with data pre-processing and concluding with clustering. For the LiveJournal dataset, three PCA components were generated, explaining 80% of the variance (Tables 4 and 5). For Instagram, two PCA components were identified, explaining 87% of the variance (Tables 6 and 7).

The analysis of loads (Table 5) permitted the formulation of a description of the PCA components for the LiveJournal.

The *P*0 component is mainly linked to the presence of references in the text. It is also associated with a distinctive use of space before commas. This may indicate the author's tendency to cite external sources. It could also reflect individual punctuation habits.

Component *P*1 characterizes the style of writing, delineating whether it is more or less formal, and also reflects the emotional coloring of the text. High values of this component may indicate texts with a positive mood, as evidenced by the presence of emoticons, whereas low values may indicate a negative mood.

The *P*2 component is indicative of the equilibrium between formal and informal elements within the text. High values of this component may be indicative of formal texts that exhibit a paucity of emoticons and a preponderance of punctuation marks.

A description of the PCA components for Instagram will be formed based on the data presented in Table 7.

Table 4

Cumulative explained variance ratio for LiveJournal (final)

Principal component index (Index)	Explained variance ratio (EVR)	Cumulative explained variance ratio
*P*0	0.3340	0.340
*P*1	0.296	0.636
*P*2	0.169	0.805

Table 5

PCA component weights for LiveJournal (final)

Holder attributes	*P*0	*P*1	*P*2
h_3	0.033	−0.486	−0.629
h_4	−0.11	−0.062	0.108
h_7	−0.4	−0.649	0.607
h_9	−0.024	−0.177	−0.034
h_{13}	−0.906	0.317	−0.272
h_{24}	−0.072	−0.454	−0.386

Table 6

Cumulative explained variance ratio for Instagram (final)

Index	EVR	Cumulative EVR
*P*0	0.475	0.475
*P*1	0.396	0.871

Table 7

PCA component weights for Instagram (final)

Holder attributes	*P*0	*P*1
h_3	0.975	0.204
h_{19}	0.023	−0.484
h_{33}	0.22	−0.851

The $P0$ component is indicative of the text's overall emotionality and is associated with the use of brief emoticons. It can be postulated that $P0$ reflects the text's overall emotionality or its "positivity", suggesting that the text is emotionally rich and may be more conversational in nature.

In turn, the $P1$ component is correlated with the style of emoji use. High values of $P1$ indicate a more formal and restrained style, wherein emojis are used less frequently and in shorter sequences.

In conclusion, opinion holders were clustered based on the final PCA components.

3. Results and Discussion

The application of the Agglomerative Hierarchical Clustering algorithm to the grouping of holders for each social media yielded the following results (Fig. 3–6). The analysis of the loadings (Table 9) enabled the formation of a description of the PCA components for the holders of the LiveJournal in the final experiment.

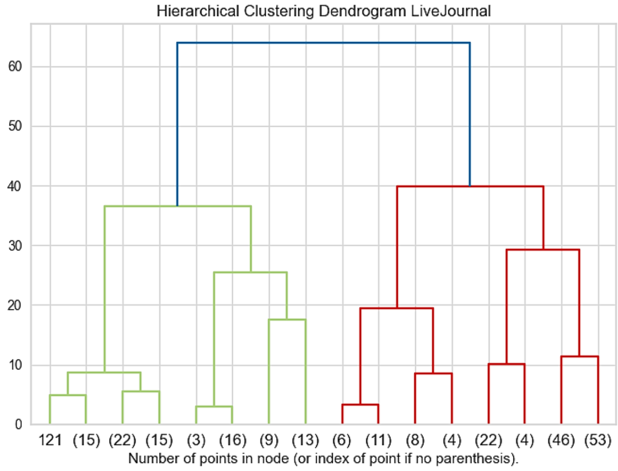

Fig. 3. Dendogram of final clusters for LiveJournal

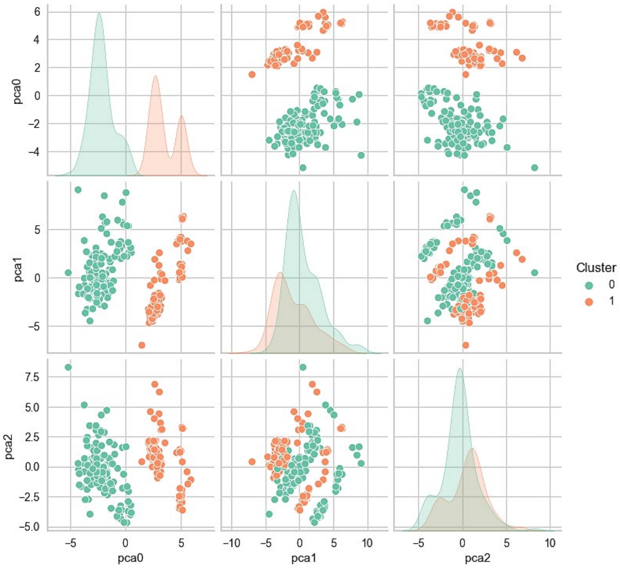

Fig. 4. Final clusters for LiveJournal

In cluster 1, the $P1$ values are higher, which is already indicative of a more formal style of writing. The $P0$ values can be low, which means a small number of references are present. The values of $P2$ may vary, as this component reflects the balance between formal and informal elements, which may differ even within the same cluster. One may, however, tentatively assume that the holders in cluster 1 may be authors of scientific articles, official documents, or holders of other types of texts characterized by a formal style and scarcity of emotive coloration. It was suggested that this cluster could take the name "Authors of scientific articles and official documents".

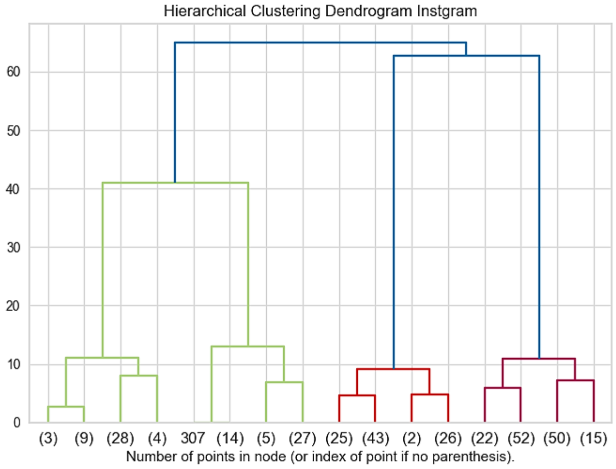

Fig. 5. Dendogram of final clusters for Instagram

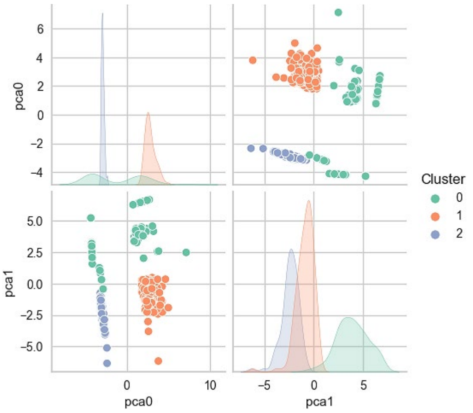

Fig. 6. Final clusters for Instagram

By contrast, the values of $P1$ for cluster 0 are relatively lower, suggesting a less formal style of writing. On the other hand, the values of $P0$ may be higher, indicating more reference variables. Further, the values of $P2$ may vary, though with more bias toward informal elements. It can be assumed that holders of cluster 0 are the authors of typical current social media posts, blogs, or other texts characterized by an informal style, use of emoticons, and other non-verbal elements. It was for this reason that this cluster was suggested to be labeled as "blog authors".

Similarly, an attempt was made to provide a verbal description of the holders on Instagram.

Cluster 0 is characterized by holders with high $P0$ values and relatively low $P1$ values. The texts of these authors are emotionally intense, with a large number of short emoticons, and are likely to be written in an informal style. In general, this cluster of holders can be described as "Authors of emotional texts".

Cluster 2 unites holders with low $P0$ values and high $P1$ values. The texts of these authors are more restrained, with fewer emojis and a more formal writing style. This cluster may be designated as "Authors of formal texts".

Cluster 1 unites holders with average values of both components. The texts of these authors exhibit a moderate level of emotionality and use emojis in a moderate amount. This can be designated as "Common users".

Accordingly, the proposed approach of double clustering with intermediate classification [21] enables the generation of a simplified view of social media holders that is both explanatory [22] and transparent [11]. The "simplified explanatory" division of holders into homogeneous groups allows to gain preliminary insight into the author based on the aggregated stylometric and sentimental features of their writing. Furthermore, there is scope for further research into a more granular division of holders by their characteristics, which may have a lower level of explainability but take into account a broader range of features [21].

The results of this research have significant practical implications for intelligent systems development in the field of social media analytics. The proposed approach to clustering opinion holders based on the aggregated features of stylometry and sentiment can be applied in order to enhance automated moderation systems, due to the identification of user behavior patterns and the detection of possibly problematic styles of communication over time. A good example is the "emotional users" cluster, characterized by heavy use of emojis, extreme sentiment polarity, and short, reactive messages. In other words, such users may be prioritized for fast human-in-the-loop moderation with the goal of preventing conflicts from escalating or capitalizing on positive virality. Conversely, the cluster entitled "common users" is featured by moderated emotionality and a balance of behavioral features; it represents the ideal segment for conducting A/B tests of novel services or recommendations of personalized content, since this group is engaged but less prone to extreme reactions.

It also enables real-time sentiment analysis improvement beyond a single message evaluation, allowing service providers to build more accurate user profiles for personalized interaction. In addition, the methodology has the potential to support the construction of adaptive recommender systems that consider not only the content preferences but also the emotional and behavioral characteristics of users. This is rather relevant for customer service platforms in various sectors, such as utilities, transportation, and e-commerce, where long-term analyses of user behavior contribute to higher user satisfaction.

While the approach proposed herein does have certain advantages, it has some limitations. First, datasets are from only two social media platforms, LiveJournal and Instagram, limiting the generalizability of results to other platforms with different interaction dynamics. Secondly, while the current feature set is rather extensive, it does not fully provide for multilingual or code-switching contexts, which could limit model performance in such linguistically varied environments. Thirdly, whereas interpretability of clusters is improved through PCA and decision trees, meaningful labeling requires domain expertise. Thus, future work can attempt to overcome these gaps.

4. Conclusions

1. The characteristics of various social media platforms impact both the expression of opinions and the behavioral patterns of users. The analysis shows that LiveJournal users (18,003 participants, 81,115 posts) compose texts that are, on average, 2.5 times longer (29.0 words per post) and utilize ellipses (18.9%) and sarcasm (16.0%) more frequently. In contrast, Instagram users (36,677 participants, 53,414 posts) favor brief, emotional comments (11.8 words), emojis (66.6% of users), and an encouraging attitude (76.0% of consistently positive holders).

2. The clustering of holders based on a single message (at the opinion level) and by the sum of aggregated characteristics (at the holder level) are aimed at solving different problems. At the opinion level, immediate reactions and emotional states in a single message are evaluated, while the holder-level approach allows for the identification of stable behavioral patterns and emotional trends that manifest themselves over a long period of time.

3. A model of social media data processing with multilevel feature selection and hierarchical clustering is proposed to form a simplified view of social media holders based on aggregated stylometric and sentimental features that support explanatory standards.

4. As a result were yielded two interpretable clusters for LiveJournal and three for Instagram. The six principal components for LiveJournal explain 81.7% of the variance, while the four components for Instagram explain 83.5%, confirming the effectiveness of dimensionality reduction. Furthermore, it can serve as a foundation for information technology [1], and it is designed to be universally applicable to any social media platform.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

The research was performed without financial support.

Data availability

Data will be made available on reasonable request.

Use of artificial intelligence

The authors have used artificial intelligence technologies within acceptable limits for grammar editing and assistance in identifying relevant literature sources. Specifically, GPT-5 (OpenAI) was consulted to explore additional literature relevant to the topic. All retrieved references were manually reviewed, verified, and incorporated by the authors.

Authors’ contributions

Heorhii Chyzhmak: Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Funding acquisition; ***Valeriy Sydorenko***: Conceptualization, Methodology, Resources, Writing – review and editing, Supervision, Project administration.

References

1. Sydorenko, V., Kravchenko, S., Rychok, Y., Zeman, K. (2020). Method of Classification of Tonal Estimations Time Series in Problems of Intellectual Analysis of Text Content. *Transportation Research Procedia, 44,* 102–109. https://doi.org/10.1016/j.trpro.2020.02.015
2. Sydorenko, V., Rychok, Y., Oladko, M. (2022). Method for Evaluation the Pattern of Internet Service Customers Based on Stylometric Analysis Oof their Text Content. *2022 IEEE 4th International Conference on Modern Electrical and Energy System (MEES),* 1–6. https://doi.org/10.1109/mees58014.2022.10005654
3. F. Mosteller and D.L. Wallace Inference and Disputed Authorship; The Federalist. Addison-Wesley Series in Behavioral Science; Quantitative Methods. Reading, Mass., Palo Alto, London, Addison-Wesley Publishing Company, Inc., 1964, XV p. 287 p., $ 12.50. (1965). *Recherches Économiques de Louvain, 31 (8),* 721–721. https://doi.org/10.1017/s0770451800020777
4. Stamatatos, E. (2008). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60 (3),* 538–556. https://doi.org/10.1002/asi.21001
5. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G. (2013). *Overview of the Author Profiling Task at PAN 2013. Working Notes of CLEF 2013 Conference.* Valencia: CEUR, 1179. https://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf
6. Giorgi, S., Preoţiuc-Pietro, D., Buffone, A., Rieman, D., Ungar, L., Schwartz, H.A. (2018). The Remarkable Benefit of User-Level Aggregation for Lexical-based Population-Level Predictions. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels: Association for Computational Linguistics, 1167–1172. https://doi.org/10.18653/v1/d18-1148

7. Chyzhmak, H., Sydorenko, V. (2023). Classification models of direct opinion holders in the space of stylometric and sentiment features of chat messages. *2023 IEEE 5th International Conference on Modern Electrical and Energy System (MEES),* 1–6. https://doi.org/10.1109/mees61502.2023.10402395
8. Rychok, Yu. S., Sydorenko, V. M. (2021). Model otsinky sentyment-komponent u zadachakh sentyment-analizu skladnoho tekstovoho kontenta. *Fizychni protsesy ta polia tekhnichnykh i biolohichnykh obiektiv*. Kremenchuk, 83–86.
9. *LiveJournal.* Available at: https://www.livejournal.com/
10. *Instagram.* Available at: https://www.instagram.com
11. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R. et al. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion, 99,* 101805. https://doi.org/10.1016/j.inffus.2023.101805
12. *GitHub – agentcooper/node-livejournal: LiveJournal API.* Available at: https://github.com/agentcooper/node-livejournal
13. *7 000 000 Russian comments from Instagram* (2025). Available at: https://t.me/danokhlopkov/395
14. VarianceThreshold. *Scikit-learn.* Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html
15. Kuhn, M., Johnson, K. (2013). *Applied Predictive Modeling.* New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3
16. Amaratunga, D., Cabrera, J. (2001). Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association, 96 (456),* 1161–1170. https://doi.org/10.1198/016214501753381814
17. Aitchison, J., Brown, J. A. C. (1958). The Lognormal Distribution. *The Incorporated Statistician, 8 (3), 145.* https://doi.org/10.2307/2986416
18. Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26 (2),* 211–252. Available at: http://www.econ.illinois.edu/~econ508/Papers/boxcox64.pdf
19. Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2 (11),* 559–572. https://doi.org/10.1080/14786440109462720
20. Nielsen, F. (2016). *Hierarchical Clustering. Introduction to HPC with MPI for Data Science.* Cham: Springer, 195–211. https://doi.org/10.1007/978-3-319-21903-5_8
21. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1 (5),* 206–215. https://doi.org/10.1038/s42256-019-0048-x
22. Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence.* National Institute of Standards and Technology. https://doi.org/10.6028/nist.ir.8312

✉***Heorhii Chyzhmak,*** *PhD Student, Assistant, Department of Computer Engineering and Electronics, Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine, e-mail: george.chizhmak@gmail.com, ORCID: https://orcid.org/0000-0001-9284-4195*

Valeriy Sydorenko, *PhD, Associate Professor, Department of Computer Engineering and Electronics, Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine, ORCID: https://orcid.org/0000-0002-4449-073X*

✉*Corresponding author*