BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **25**, No 1
Sofia • 2025 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2025-0003

To Share or Not to Share: Investigating Drivers for Sharing Online News Using Automated Machine Learning and Probabilistic Modeling

Anton A. Gerunov

Sofiq University "St. Kliment Ohridski", Faculty of Economics and Business Administration, 125 Tsarigradsko Shosse Blvd, 1113 Sofia, Bulgaria
E-mail: a.gerunov@feb.uni-sofia.bg

Abstract: *The article leverages advanced machine learning to investigate what drives sharing behavior for online news content. To this end, it investigates a dataset of 39,797 pieces of individual news and uses 59 different features to outline the key influences on the number of shares. Initially, an automated machine learning framework is applied to choose the optimal model among 1,000 contenders and then this model is used to study the likelihood of sharing further. Causality links are investigated in more detail by recourse to a Bayesian Belief Network, which elucidates the transmission mechanisms and the direct quantitative effect of relevant predictors. Overall results show that the volume and uniqueness of content, appropriate keywords, and the article's position in the information network are all important predictors for sharing. Conversely, negative polarity is shown to be the most important blocker.*

Keywords: *News sharing, Viral content, Machine learning, AutoML, Bayesian network.*

1. Introduction

The increased digitization of many economic and business activities has led to blurring the boundaries between real and virtual customer experiences. On a practical note, this underlines the increased importance of modeling and predicting online user behavior. One of the most important emerging online behaviors is the actions that lead to a sudden and explosive rise in popularity of a given piece of content. It internet-lingo is known as becoming "viral", comparing the rising popularity of the content to the explosive growth of an infectious disease.

The major difference is that while viruses have a set of tractable properties that can explain their transmission, this is not the case for online content. Thus, the prediction of virality (the number of shares) becomes a key problem for both businesses and academics. This question needs to be addressed by investigating large quantities of unstructured or semi-structured data. Thus it is a better strategy to move beyond traditional econometric tools and embrace machine learning instead. This article attempts to study what drives the sharing behavior of online news by leveraging both a quasi-Bayesian approach and a set of automated machine-learning algorithms. Those enable the analyst to better understand both the inner workings of online sharing behavior, as well as to forecast how many times a piece of content is shared.

The article is structured in the following way. Section two presents a short review of the extant literature, outlining what we currently know about drivers of online sharing. The third section presents the data and shows several stylized facts about the dynamics of online sharing. Section four moves on to implement an automated search for the best-fitting machine learning algorithm. The following section presents an exploratory analysis of possible causal links, using conditional probability to construct a Bayesian network of relevant features. The last two sections discuss the results and present a few concluding remarks.

2. Literature review

The study of online content sharing is important across all digital outlets but it becomes ever more necessary with the advent and increased maturity of social networks and other online media. Overall research on news sharing has tended to focus on a few main strands of enquiry. Those include investigating the impact of the characteristics of users of organizations, the content, networks, and their interactions, affect online news sharing [1].

In a seminal article [2] bluntly asks the question of what makes a piece of content viral and tries to answer it with a dataset of New York Times articles. The analytic tool of choice is a logistic regression. The authors take a psychological approach and investigate how particular emotions and their valence shape online content diffusion. It seems that strong emotions evoke more extensive sharing, with positive content being more likely to be shared than negative one. Arousal is a key predictor – both in terms of amplifying positive (awe) and negative (anger, anxiety) emotions. Conversely deactivating emotions such as sadness tends to decrease the probability of sharing.

It is not only the emotional content of the shared news that matters – the emotional state of the sharing agent is also crucial. It seems that emotions from the negative polarity tend to induce greater sharing likelihood. Hoewe and Parrott [3] have investigated sharing behavior following the 2016 US presidential election. They find that those users who experienced the greatest amount of anger are most likely to seek and share information about the election results – both online and through interpersonal communication. Anxiety and enthusiasm turned out to be weaker predictors, while hopefulness had practically no effect on sharing behavior. In a similar vein, Cantwell and Kushlev [4] show that people who felt more anxious were also more likely to share news during the global COVID-19 pandemic.

The cognitive processes around sharing can also be divided into the stages of sharing – before, during, and after. Wang and Fussel [5] point out that before sharing content assessment and evaluation the social value of sharing takes place.

During sharing the users tend to decide the audience of the post and adjust their behavior accordingly. Finally, after sharing two critical cognitive processes take place – the process of expecting feedback (and deriving utility from it) as well as the process of revisiting previous sharing.

Some of the results obtained in [2] also seem to be replicable across different samples, particularly regarding the importance of emotions. Wadbring and Ödmark [6] find that more interactions and shares tend to go to news with positive content. Thompson, Wang and Daya [7] investigate individual motivations of Facebook users when sharing online news. They leverage a Structural Equation Modeling (SEM) approach to look into what factors are important predictors of the intention to share news. Two key motivators emerge – the status-seeking motive (credibility and reputation), as well as the information-sharing motive. Their model fails to find sufficient proof that socializing, entertainment, or pastimes are motives for sharing [7].

Additionally, there seems to be a difference in sharing attitudes between regular consumers and opinion leaders. In two experiments Bobkowski [8] shows that regular users are more likely to share news that has informational utility, while opinion leaders are mostly agnostic between useful and non-useful news sharing. A final nuance here is given by Ihm and Kim [9]. They use a survey of 400 users to show that those who are more motivated by self-presentation, including creating relationships and building a reputation, are more likely to engage in sharing behavior. Furthermore, Beam, Hutchens and Hmielowski [10] leverage a multilevel modeling on a sample of 403 online users to check whether more informed users tend to share more. While they do find that increased exposure to online news leads to increased sharing, there is no evidence that increased sharing is in any way connected to knowledge. It is not the more knowledgeable internet sleuths that shape the landscape of online opinions and communication.

Shi, Rui and Whinston [11] investigate in more detail the importance of closeness between different agents for providing virality. Looking at data for sharing behaviors on Twitter (so-called retweets), they find out that closeness is not particularly important but rather weak ties (uni-directional followers) have a greater likelihood of resharing. This shows that Granovetter’s weak ties theory neatly translates into the online environment as well.

Sharma and Cosley [12] show that individual preferences for sharing certain items are crucially important, but agents also try to customize their shared content based on the expected recipient. Building upon this individualized psychological line of investigation, Scholz et al. [13] hypothesize that content sharing is driven by a value-based decision process. They assert that it is neural value signals that respond to both the attributes of the media content and to social influence. There is a further neural process that integrates content features and social influences that impinge on final decisions. The features of the media under question seem to be an important predictor of whether it is going to be shared. Results by Hsiao [14] show that this may be a positive feedback system, showing that sharing eventually leads to a higher level of both internet satisfaction and life satisfaction. Interestingly, the life satisfaction effects are greater for males than for females.

Finally, content that is interesting, surprising, or useful also tends to be shared more widely. The topic of the news also matters – while front-page news tends to be dominated by crime, the shared content is more likely to be political [6].

It is worth noting that content sharing online is driven by user perceptions, and not necessarily the objective value or characteristics of what is reposted. Factors such as issue framing believability, bias, perceptions of importance, and influence impinge on user sharing intentions [15]. Furthermore, those seem to be framed by the espoused political ideology (ibid). Similar results are found in other recent research. Bhagat and Kim [16] leverage survey data of 513 active internet users to show that the most important drivers behind sharing behavior are news quality, source credibility, perception of civic engagement, and influence on others, as well as social influences.

While there seems to be a growing understanding of the drivers behind online news sharing, the prediction of an actual number of shares remains a challenge. A strand of research has taken up machine learning models and algorithms to tackle it. Studies have been particularly fruitful and abundant in the field of studying fake news, but those approaches can naturally and seamlessly extend to other types of news.

Leveraging approaches from natural language processing and regression modeling, Trilling et al. [17] have studied relatively large datasets of around 870,000 links that have been shared around 100 Million times on the social network Facebook. They find that the drivers of viewing and sharing are different. For example, political news tends to be viewed less but shared more. Prado-Romero et al. [18] test actual news coming from Yahoo News and investigate the utility of various classifiers that prove useful to the tasks of predicting topic popularity.

Meghawat et al. [19] propose the analyst go beyond simply textual features of the shared content and instead move to a multi-modal feature dataset. More concretely, they use a set of machine learning algorithms (including random forest and neural network) on an enriched dataset including text, social and contextual information, as well as metadata. Their major conclusion is that a Convolutional Neural Network (CNN) has excellent performance on multimodal datasets.

The proliferation and growing popularity of machine learning methods also means that a large variety of such algorithms has been applied to the problem of understanding news sharing and detecting fake news. For a review, the reader is referred to [20] and the references therein. Still, it is worth noting that the most popular algorithms used are the Naïve Bayes classified, decision trees, Support Vector Machines (SVM), neural networks, random forests, and XG Boost (ibid.). The current state of the art seems to be opportunistic fitting of advanced models to existing datasets and measuring their performance. The usual set of well-known predictors performs well on this task as well. Gerunov [21], for example, shows how 109 different machine-learning algorithms can be applied to five different forecasting tasks, with variants of random forest and neural networks registering consistently good performance.

The research challenge then is not to fit another model but rather to understand the causal links between sharing drivers and to automate the prediction process so that it functions seamlessly at scale. The current article focuses on both tasks by consecutively applying automated machine learning to find the optimal prediction for the drivers under investigation and leveraging a Bayesian network to shed more light on their causality.

3. Data and stylized facts

The twin tasks of prediction and causality modeling of such a complex phenomenon call for a sufficiently large and detailed data set. The article thus uses data provided in [22]. The dataset contains structured data on 39,797 news items published on the popular online portal Mashable. The authors (ibid.) apply Natural Language Processing (NLP) methods to key quantitative characteristics and features of each of the texts. They include variables such as word count, hyperlinks to different sources, emotional charge, sentiment polarization, publication category, day of the week, etc. The total number of features provided is 60. The variable under investigation is naturally the number of shares. In [22] this data is used to demonstrate the performance of a decision support system with an automated component. They also test five different forecasting algorithms, finding that a random forest model produces the best results.

A more detailed overview of the dataset with its full set of descriptive statistics is available in [23], and the list of explanatory features is added as an appendix to the article. Here we will outline a few stylized facts about the data at hand that inform both modeling and further conclusions. On average, an article contains 546.5 words in its content, and its title consists of 10 words. While we observe significant differences for the volume of content (σ = 471), which is not the case for titles (σ = 2.11). Most articles consist of repeated words, with an average ratio of 1 unique to two non-unique words (μ = 0.55), but here again, the variance is quite large.

The number of hyperlinks these articles contain is considerable. On average, the materials have 11 links, of which 3 are to the same website. Both statistics vary widely, with the article with the most links having 304. Articles tend to be richly illustrated, with an average of nearly 5 images (μ = 4.54, σ = 8.31) and at least one video (μ = 1.25, σ = 4.11). In terms of topics covered, most articles were published in the World section (21%), followed by Technology (19%) and Entertainment (18%). In addition, the database provides data on the number of shares of the best, worst, and average keywords, as well as sharing of links to the same site. There is also data on the days of publication, with the most frequent publications being on Tuesday or Wednesday (19%).

Fernandes, Vinagre and Cortez [22] also perform a corpus analysis, inferring main themes based on a latent Dirichlet distribution and reporting the proximity of the given article to any of the five main themes. Those themes are also included in the dataset as variables with the prefix LDA. For more details on this model, we refer the reader to the original publication by Blei, Ng and Jordan [24]. This dataset also contains information about the level of subjectivity of the article, as well as about the level of emotional polarization. It is noteworthy that both are relatively high in their various levels of measurement. The overall subjectivity of the article averaged 34%, and its overall polarization – 16%. These measures are arrived at by determining what proportion of words are subjective and what proportion are polarizing, relative to certain fit-for-purpose corpora. For more information on sentiment analysis in online communication, we refer the reader to the review article in [25].

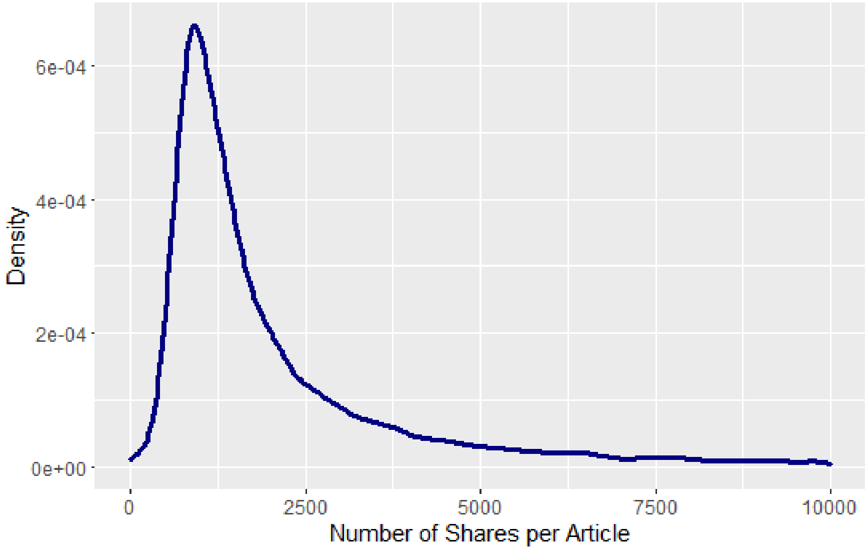

Fig. 1. Density of the target variable: number of shares

The target variable in the dataset is the number of shares of an article, where we note that on average an article is shared 3.395 times, but this varies widely. The significant differences are highlighted both by the high values of the standard deviation (σ = 11.627) and by the very wide range in which the dependent variable varies. The least popular article was shared only once, and the most popular one - as many as 843,300 times. Judging by the skewness and kurtosis values, we can say that the distribution of the target variable is not Gaussian. This is an expected result given that many phenomena in the digital world follow exponential distributions [26].

This is readily apparent in Fig. 1. The distribution is far from normal and follows the approximate shape of the lognormal distribution. Thus, there are many articles with a relatively low number of shares peaking at around 1000, as well as an extremely small number of articles with a huge number of shares. For clarity, the graph presents articles with up to 10000 shares, but the distribution tail is very long, and the most popular article reaches over 843,000 shares. Such disparities in popularity are typical of many phenomena in the digital environment and are sometimes described as “winner-takes-all” situations, i.e., a disproportionately large reward is obtained by the top performers.

The correlation matrix of all the features shows that the average absolute correlation of the number of shares with the remaining independent variables is only r_μ = 0.0087. The highest positive correlation of shares is with the number of shares of the average-popular keyword of the given article (r = 0.11), and the highest negative correlation is with the topic LDA02, as here the correlation coefficient is $r = -0.06$.

In short, real-world data for online news sharing features a few characteristics that pose significant challenges to traditional statistical methods. These are namely:

- Non-normal distribution of most variables, thus compromising the normality assumption inherent in many traditional statistical tests.
- Non-linearity of data, thus making the modeling through traditional linear models challenging barring a careful transformation of data.
- The large size of the dataset both in terms of features and in terms of observations, makes some traditional econometric methods obsolete and calls for a careful interpretation of *p*-values obtained.

Those stylized facts underline the need for a set of novel machine-learning algorithms for modeling such data. Due to the scale of the task, model evaluation will also need to be different., Evaluation needs to move away from *p*-values and coefficient sizes and into forecast accuracy metrics (e.g., Root Mean Squared Error, RMSE) and relative variable contributions (e.g., Gini impurity).

4. Model selection and interpretation

Once presented with such a challenging forecasting task such as predicting news virality, the researcher may opt for opportunistic model fitting. It is well-known that several popular machine learning algorithms tend to produce excellent fit in a large number of potentially interesting tasks. Some researchers even question the need for additional ones [27]. Following this logic, the researcher may choose to fit those standard models and select the best one among those, based on a relevant error metric. Among those algorithms that constitute the best practice in forecasting one can include the traditional multiple regression and k-Nearest Neighbors (kNN), as well as more novel approaches such as the neural network, random forest, and the support vector machine.

We split the total sample into a training set that consists of 80% of all observations (total of $N = 31{,}716$ observations, and a test set that contains the other 20% (thus reaching $N = 7928$ observations). To avoid overfitting and estimate correct out-of-sample error metrics, we train all the models on the training set and then generate predictions with the test set, i.e., with data the models have not seen before. Based on those predictions many error metrics are calculated – the mean error, the root mean squared error, the mean absolute error, the mean percentage error, and the mean absolute percentage error. All the error metrics are calculated by generating predictions for out-of-sample data, thus minimizing the risk that results are driven by model overfitting. For more details on the train/test procedure and a more involved review of the error metrics, the reader may consult [21]. The results from this exercise are shown in Table 1.

The error metrics give similar qualitative conclusions. For concreteness, we will focus on the most theoretically sound one – the RMSE – to perform model selection. It seems that across the traditional high performers, the random forest algorithm (a tree-based ensemble algorithm) has the lowest RMSE, standing at 12,698.11. The regression, neural network, kNN, and SVM all have RMSEs above 13,000 but their performance is still satisfactory.

Table 1. Performance of alternative forecasting algorithms

Algorithm	Mean Error, ME	Root Mean Squared Error, RMSE	Mean Absolute Error, MAE	Mean Percentage Error, MPE	Mean Absolute Percentage Error, MAPE
Multiple linear regression	176.21	13,120.37	3185.35	–240.80	262.42
Neural network	3506.38	13,338.84	3506.38	99.90	99.90
k-Nearest Neighbors	319.83	13,213.30	3309.44	–175.19	203.24
Random forest	–50.97	12,698.11	3219.62	–267.90	285.32
Support vector machine	1348.31	13,101.14	2633.16	–103.29	133.17
AutoML: gradient boosting machine	107.81	11,505.41	3223.93	–244.81	266.09

At this point, one may find it inexplicable that given the increases in computational capacity and algorithm efficiency, experiments are confined to such a limited number of algorithms. We thus propose to move away from opportunistic model fitting and into a more formalized and rigorous procedure of finding the best-performing prediction. This can be achieved by fully leveraging the capability of the so-called AutoML (automatic machine learning). AutoML is defined as automatically constructing the machine learning pipeline on a limited computational budget [28], the place where machine learning meets automation [29]. Essentially, AutoML starts with a pre-selected group of algorithms and then trains a large number of models based on those algorithms, tuning their parameters as it goes. Finally, an information criterion (such as the Akaike or the Bayesian one) is used to select the best-performing model. Conceptually, this is close to the idea of performing a grid search for optimal classifiers and predictors (see also [30]). For a more detailed overview of AutoML, the reader is directed to the review articles in [28].

The article leverages one of the most popular AutoML frameworks – H2O (see [31]) and uses it for the prediction problem at hand. H2O's automatic machine learning functions contain the following state-of-the-art models:

- Distributed Random Forest;
- Extremely Randomized Trees;
- Generalized Linear Model (GLM) with regularization;
- XGBoost Gradient Boosting Machine;
- H2O Gradient Boosting Machine;
- Multi-layer Artificial Neural Network;
- 2 Stacked Models, of which one contains all the models trained and the other – the best-in-class model.

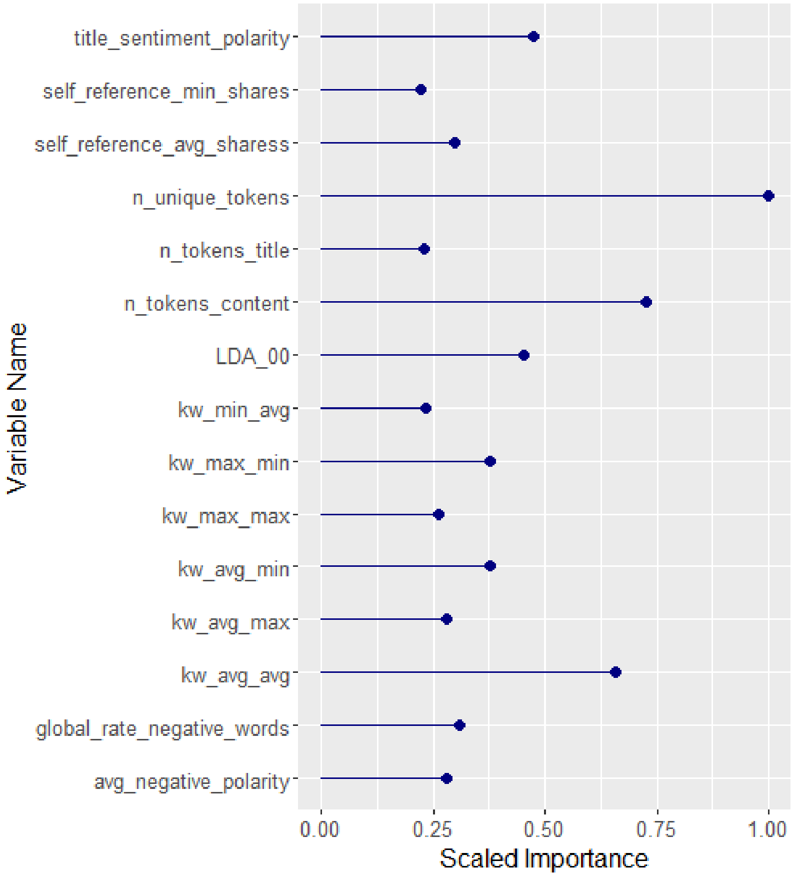

Fig. 2. Density of the target variable: number of shares

For a more detailed overview of the framework and its algorithms, the reader may consult [31]. We set automated model training, configuring AutoML to train 1000 different models and using a cross-validation procedure to select the best one (with the lowest RMSE). The model returned is a Gradient Boosting Machine with 50 internal trees, and its RMSE on the same test data set stands at 11,505.41. This number stands markedly lower than the error metrics of other contenders and clearly shows that automated model fitting and selection can provide superior forecasting results. Moreover, AutoML presents a standardized, automated, and rigorous framework for performing machine learning tasks and moves beyond expert intuition and into the realm of exhaustive inquiry.

From the perspective of model quality, the AutoML procedure also seems to be a viable approach. An analysis of the residuals shows that there is a very limited discernible trend between fitted values and residuals, pointing to the lack of systematic errors (see Fig. 1 in Appendix A).

The model selection procedure provides another key benefit – the possibility to elicit model drivers, thus understanding the underlying features that make for news virality. With a relatively large number of predictors and tens of thousands of observations, it is useful to move beyond statistical significance to better understand model drivers. More particularly, one can say that an important variable decreases the prediction error generated by the model. We investigate the increase in the model standard error when each of the variables is removed. From this, we conclude that the most important variables are the ones responsible for the largest improvement in forecast accuracy. Their influence is scaled so that the most important one has a score of 1, and the least important – a score of 0. The fifteen most important variables are presented visually in Fig. 2.

News sharing seems to be largely driven by four main groups of variables:

- Numbers of words – the single most important predictor is the number of unique words in the article itself (n_unique_tokens). It is then followed by the volume of the content – the total words in it (n_tokens_content), and also the length of the title (n_tokens_title). Longer pieces of content are more likely to be shared.
- The keywords that are featured in the article – the average keywords in an article with minimum shares (kw_min_avg), the worst, average, and best keywords in articles with average shares (kw_avg_min, kw_avg_avg, kw_avg_max), and the worst and best keywords in articles with maximum shares (kw_max_min, kw_max_max).
- Number of links – the number of links has a robust and positive effect. Network effects are visible here – the more the referenced articles are being shared, the more the referencing article is. This holds for articles with minimum and average shares (self_reference_min_shares, self_reference_avg_shares).
- Sentiment and polarity of the article – the rate of negative words in the content (global_rate_negative words), the average polarity of negative words (avg_negative_polarity), and the title polarity (title_sentiment_polarity) affect whether the article is going to be shared. However, negative polarity seems to be a relatively less important driver of sharing than factors such as the number of unique tokens or keyword quality.

This may very well be considered a model mining exercise. We tested sequentially alternative algorithms and estimated a large number of potential models (with their specific parameters) The variable importance may thus be an artifact of the machine learning process (i.e., what makes prediction better) and not necessarily reflective of underlying dynamics. This is why the AutoML approach can benefit from being supplemented with a more formal analysis of causality. The next section presents such an approach, leveraging Bayesian Belief Networks (BBNs).

5. Bayesian causality analysis

A better understanding of news-sharing behavior depends on rigorously modeling the causal links between different features of the news content (e.g. volume, keywords, polarity). A possible venue to do this is to take a probabilistic route and estimate the links between different variables, thus creating a network of effects. To do this we leverage a BBN. The network algorithm constructs a visualization of variables under study (features, events, variables, etc.) and shows their causal links in the form of a Direct Acyclic Graph (DAG). Such models enable the researcher to conduct a deeper investigation of effects, as the network algorithm constructs a local conditional probability distribution for every node. We will briefly outline the specifics of a Bayesian Belief Network, but a fuller treatment can be found in [32-35].

Let us denote the probability of a given event *A* occurring with *P*(*A*), and that of an event *B* – with *P*(*B*). The Bayesian theorem relates the probability of *A* occurring, given that *B* has occurred, or its conditional probability *P*(*A*|*B*) as follows:

(1) $$P(A \mid B) = P(A \cap B)/(P(B)).$$

Thus, one can estimate an unknown probability given a set of known ones. Building upon that, the Bayesian network creates a whole set of interconnected network nodes with each of them having a specific probability distribution. The underlying construction of the network assumes that any given node V_i depends on its parent nodes alone (pa(V_i)) and no other variable in the system. This leads to the following:

(2) $$P(V_i) = P\big(V_i | \mathrm{pa}(V_i)\big).$$

The sum of probabilities at every node is one, and thus all the events in the network are completely characterized:

(3) $$\sum_{V^*} P\big(V_i = V^* | \mathrm{pa}(V_i)\big) = 1.$$

The network computes not only the individual conditional probabilities but also an aggregated probability distribution *P*(*V*) defined as follows:

(4) $$P(V) = \prod_{V_i \in V} P\big(V_i | \mathrm{pa}(V_i)\big).$$

The Bayesian network is not a deterministic model – the resulting structure and precise parameter values may vary slightly with the chosen computation algorithm or criteria for the optimization procedure. Those effects usually diminish with increasing the sample size. It is then useful to calculate BBNs over several iterations to ensure consistency and robustness of the results obtained. This also extends to the specific network computation algorithm that is utilized.

The Bayesian network is not learned all at once. Instead, it goes through three typical phases. The standard approach begins with learning the structure of the network (i.e., the nodes and the edges present). In the second phase, the algorithm determines the direction of those edges. The direction of the edges can also be interpreted as causal links – i.e., the presence of a directed edge implies that the first node’s existence influences the probability of the second node’s existence. This process of situating edges and nodes and estimating the direction of the edges repeats over several iterations, creating alternative versions of the network. In the final phase, an information criterion (e.g., the Bayesian or Akaike one) is used to determine the optimal network structure.

We estimate a Bayesian Belief Network following this standard approach on the news-sharing data set. While there are some alternative network estimation algorithms (see [36]), there seems to be no consensus on the best ones. The BBN is thus calculated using one of the most popular ones – the hill climbing algorithm (see [37]). We iterate over rounds of calculating the network, finding that the structure remains mostly invariant, with edge directions showing robustness and stability. The final Bayesian Belief Network is graphically presented in Fig. 3.

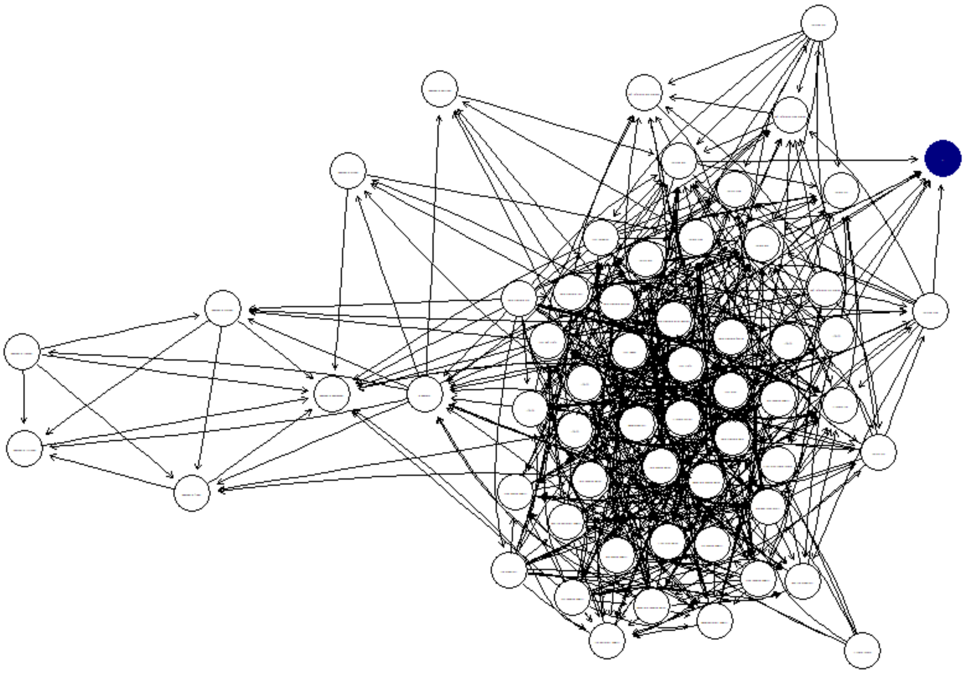

Fig. 3. Schematic structure of bayesian belief network for news sharing

It consists of one strongly connected central component with a rich pattern of influences between different article features. On further inspection, the few outstanding nodes on the left side of the network turn out to be the days of the week. The outer nodes of the network that influence the strongly connected component in the middle are usually connected with the emotional polarization or the number and type of keywords. These influence other features but are not influenced by them as keywords and emotional polarity are somewhat exogenous to the overall structure. In a similar vein, the target variable – the number of article shares is on the edge of the network. It is directly influenced by many features, but it does not influence any feature. This is expected as sharing behavior is the result of other nodes and takes place after they are formed. Such a structure of the network seems to capture the expected causality links. It also gives further credence to the idea that the current network is well-identified.

The analyst may not only be interested in the overall network structure but may want to focus on direct drivers of sharing behavior in more detail. The nodes that directly affect sharing are the following:

- Number of links (num_hrefs);
- Entertainment channel (data_channel_is_entertainment);
- Best keywords for maximum shares (kw_max_max);
- Average keywords for minimum shares (kw_min_avg);
- Best keywords for average shares (kw_max_avg);
- Average keywords for average shares (kw_avg_avg);
- Minimum shares of referenced articles on the platform (self_reference_min_shares);

- Closeness to LDA topic 3 (LDA_03);
- Average negative polarity (avg_negative_polarity).

The Bayesian network allows one to follow the influence of each of those factors on the number of shares. Furthermore, each of those factors is a node in the network that is in turn influenced by its parent nodes. In a way, the network enables the reconstruction of causality links flowing from a given node to the result.

We further model the influence of each parent node on the target one (number of shares). To this end, we fit the parameters of the Bayesian network using a maximum likelihood estimator for the regression coefficients of each node's effect. Results are presented in Table 2.

Table 2. Key factors influencing news-sharing behavior

Feature	Name of variable	Contribution to conditional density
Closeness to LDA topic 3	LDA_03	875.8545
Number of links	num_hrefs	27.173
Average keywords for average shares	kw_avg_avg	1.795558
Minimum shares of referenced articles on the platform	self_reference_min_shares	0.026155
Best keywords for maximum shares	kw_max_max	–0.00129
Best keywords for average shares	kw_max_avg	–0.20048
Average keywords for minimum shares	kw_min_avg	–0.47214
Is channel "Entertainment"	data_channel_is_entertainment	–697.467
Average negative polarity	avg_negative_polarity	–1649.28

The highest direct influences on the number of shares are the closeness to LDA Topic 3 (broadly corresponding to more involved content) and the number of links that the article features. News items that load strongly on LDA3 get on average 876 more shares, while each link in the article increases the number of shares by 27. Both effects are strong and practically significant. On the other hand, the number of average (medium quality) keywords is notable, but with limited impact – one more keyword increases shares by two, on average. The number of referenced articles and other keywords also have some impact, but it is of relatively small size.

The two large blockers to further sharing seem to be whether the article comes from the entertainment channel. If yes, the news item gets almost 700 shares less, on average. It seems that people are reluctant to share lighter stories as they may impinge negatively on the user's self-constructed image. This insight is close to the results obtained in [8]. Finally, negative polarity leads to fewer shares. It is the strongest predictor in the sample. Each point on the negative polarity scale leads to an average decrease in the number of shares of about 1650. This result is clear in the current sample, and it also replicates previous research showing it (e.g., [2]) but it seems far from conclusive. Previous research has not reached a consensus on the effect of negative polarity on sharing. While some authors find a clear correlation with more negative news being more likely to become viral [40], others [41] show that audiences are not using negative as a cue for sharing. Matthews, Bélair-Gagnon and Lewis [42] explain this pattern of behavior as individuals may choose not to share negative news so that they do not ruin their social connections and reputations. In sum, the results obtained give credence to the idea that negative news is not necessarily more likely to become viral.

6. Discussion

The results obtained paint a deeper and more nuanced picture of online sharing behavior. Recent advances have allowed the research community to obtain a deeper understanding of the fundamental drivers behind online content virality and be able to predict it with increased accuracy. This article proposes the division of the two tasks so that insights from the forecasting and modeling exercise can feed into causality investigation, and vice versa. The precise determination of causality links goes beyond just statistical evaluation. It also has to be theoretically justified, and the research design has to ensure that the time sequences of events are met and that no intervening variables obscure the relationship. Despite this, we can outline a few preliminary conclusions based on previous results in the extant literature and the machine learning and probabilistic modeling presented in this article.

First, it seems that the transition from traditional statistical methods to a set of more advanced machine learning algorithms for forecasting news virality is a reasonable strategy. The error rates for traditional methods are notably higher compared to even those of out-of-the-box machine learning approaches. The RMSE in the new sharing forecasting tasks of the multiple regression stands at 13,120.37, and that of the kNN stands at 13,213.30. In contrast, the SVM registers lower RMSE, and the random forest model RMSE stands at 12,698.11. The large samples and the rich data set structure (especially in terms of features) call for more advanced forecasting approaches.

Second, opportunistic modeling that leverages popular machine learning algorithms improves upon classical methods, but this process can be formalized and expanded. We have leveraged an AutoML approach searching for the best model that is optimally parameterized for the prediction problem at hand. Iterating over 1,000 alternatives, the algorithm finds that a Gradient Boosting Machine (GBM) with regularization that has 50 internal trees is an optimal fit. The model RMSE is much below all the other contending alternatives – RMSE = 11,505.41. This result shows that an automated machine-learning framework can outperform human experts doing theoretically informed model fitting. This article has leveraged the H2O framework in particular. Still, the proliferation of AutoML and the ensuing frameworks and approaches provide choices for the researcher.

Third, we can expand the causality investigation of a pertinent process beyond an automatically fitted model. Instead, we construct a more detailed probabilistic network of influences using a Bayesian Belief Network estimated with a hill-climbing algorithm. The edges of the network can be interpreted as pointers for causality and their direction is consistent with both results from extant literature as well as with insights obtained from the GBM variable importance. The network provides not only direct causal links but also the transmission mechanisms from one variable (or feature) through the next, all the way into the terminal target variable (number of shares). We have demonstrated that the qualitative insights from the network overlap with those from the GBM, thus triangulating and further validating our results.

Fourth, this two-pronged approach allows one to glean substantive insight from the data at hand. It seems that sharing behavior is largely driven by four main factors. The first one is the number of unique and total words in the article content and title – they matter a lot. The more unique words the article has, and the bigger it is, the more likely users are to share it. The second group consists of the number of keywords (best, average, worst) per type of article (popular or not). The larger the number of keywords, the more likely sharing is. This effect is of somewhat limited practical significance as it is not quantitatively large. The third group reflects the position of the news article in the overall network of information. This is proxied by the number of links it has to other articles and the number of times those articles are shared. A larger number of links leads to more virality, with an average of 27 more shares per link. The effect is practically significant and of average size. The final group of predictors has to do with emotions and polarity. Results show that negative news items are less likely to be shared than positive ones – the largest direct effect of all the predictors under study. While people avoid sharing negative news, they are also reluctant to share entertainment pieces, thus pointing to a propensity to share positive "serious" content. The Bayesian network allows for a further study of those drivers, showing that the number of keywords, links, and emotional content have a direct influence on the number of shares. Volume metrics work indirectly by influencing intermediate nodes.

However, there are some limitations to the current study. While it has leveraged an extensive database with quantitative indicators, it does not consider the full set of relevant online interactions. Particularly, we do not study the behavioral biases and cognitive constraints that decision-makers experience in an online environment. Previous experimental research has shown those to be important influences over actual behavior [38]. The data used here also does not include information on important features of the online environment such as the presence of marketing campaigns, malicious actors spreading fake news, and the effect of online influencers and opinion leaders. Further research is needed to better understand the effects those have on the virality of online content.

Overall, those results may be used both for forecasting the sharing behavior for existing news as well as designing online content for virality. Digital marketers may choose to create positive news with many links and pertinent keywords and let the wisdom of the Internet make it viral. While this will increase the chance of many shares, we should note that many factors relevant to predicting the long tail may not be included in the current research, or indeed – in any easily accessible dataset. The personal characteristics and psychological wiring of individual users spring to mind as prime examples. Despite those limitations, current results do provide some insight into sharing behavior and further research will likely further elucidate it.

7. Conclusion

This article investigates what makes pieces of news viral. In essence, this is a forecasting problem that tries to identify and model the drivers behind user-sharing behavior online. Stepping upon a rich foundation of previous research, we propose a twin-peak approach to tackling this problem. Initially, we use automated machine learning to search for the best forecasting model among a very large sample of possible candidates. Leveraging H2O, we investigate 1000 automatically fitted models. The winner of this simulated competition – a variant of GBM– outperforms standard econometric and machine learning approaches. Investigating variable importance shows that uniqueness and volume of content, as well as keywords, place in the information network, and emotional sentiment, are important predictors.

Such models leave a bit to be desired – while the forecasted number itself is remarkably accurate, one is left to wonder about the exact causality links. To further study those, we propose leveraging probabilistic modeling by constructing a Bayesian Belief Network. It shows how different causal effects propagate and what features get influenced until reaching the target variable – the number of shares. Results point out that keywords, links, and sharing of linked content, topics, and emotional content directly affect sharing. Content volume and word uniqueness do so only indirectly. The BBN also allows one to calculate the strength of effects and we thus find that closeness to a certain topic cluster and number of links have the strongest positive direct influence. In contrast, the type of content (entertainment or not) and the negative polarity have the strongest negative direct influence.

Concluding, such a two-pronged approach can adequately answer the needs of researchers and practitioners to understand and forecast user behavior online. The main limitation of this approach is the scarcity of psychological features of the individual users themselves – an undoubtedly crucial piece of the sharing puzzle. This is a naturally intriguing venue for further research, especially if the psychological and cultural characteristics of users can be reconstructed using readily available behavioral and engagement data. Additionally, more sophisticated models such as Long Short-Term Memory (LSTM) neural networks or advanced language models such as BERT or transformer-based models can be used to tackle this research problem. Still, the initial insights presented in this paper may serve as a useful vantage point for both creating viral content as well as better understanding the subtle nuances of consumer behavior online.

Acknowledgments**:** This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, Project SUMMIT BG-RRP-2.004-0008-C01.

References

1. Kümpel, A. S., V. Karnowski, T. Keyling. News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks. – Social Media + Society, Vol. **1**, 2015, No 2, 2056305115610141.
2. Berger, J., K. L. Milkman. What Makes Online Content Viral? – Journal of Marketing Research, Vol. **49**, 2012, No 2, pp. 192-205.

3. Hoewe, J., S. Parrott. The Power of Anger: How Emotions Predict Information Seeking and Sharing After a Presidential Election. – Atlantic Journal of Communication, Vol. **27**, 2019, No 4, pp. 272-283.
4. Cantwell, O., K. Kushlev. Anxiety Talking: Does Anxiety Predict Sharing Information about COVID-19? 2021.
5. Wang, L., S. R. Fussell. More Than a Click: Exploring College Students' Decision-Making Processes in Online News Sharing. – Proceedings of the ACM on Human-Computer Interaction (GROUP), Vol. **4**, 2020, pp. 1-20.
6. Wadbring, I., S. Ödmark. Going Viral: News Sharing and Shared News in Social Media. – OBS-Observatorio, Vol. **10**, 2016, No 4, pp. 132-149.
7. Thompson, N., X. Wang, P. Daya. Determinants of News Sharing Behavior on Social Media. – Journal of Computer Information Systems, 2019.
8. Bobkowski, P. S. Sharing the News: Effects of Informational Utility and Opinion Leadership on Online News Sharing. – Journalism & Mass Communication Quarterly, Vol. **92**, 2015, No 2, pp. 320-345.
9. Ihm, J., E. M. Kim. The Hidden Side of News Diffusion: Understanding Online News Sharing as an Interpersonal Behavior. – New Media & Society, Vol. **20**, 2018, No 11, pp. 4346-4365.
10. Beam, M. A., M. J. Hutchens, J. D. Hmielowski. Clicking vs Sharing: The Relationship between Online News Behaviors and Political Knowledge. – Computers in Human Behavior, Vol. **59**, 2016, pp. 215-220.
11. Shi, Z., H. Rui, A. B. Whinston. Content Sharing in a Social Broadcasting Environment: Evidence from Twitter. – MIS Quarterly, Vol. **38**, 2014, No 1 pp. 123-142.
12. Sharma, A., D. Cosley. Studying and Modeling the Connection Between People's Preferences and Content Sharing. – In: Proc. of 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2015, pp. 1246-1257.
13. Scholz, C., M. Jovanova, E. C. Baek, E. B. Falk. Media Content Sharing as a Value-Based Decision. – Current Opinion in Psychology, Vol. **31**, 2020, pp. 83-88.
14. Hsiao, C. C. Understanding Content Sharing on the Internet: A Test of a Cognitive-Affective-Conative Model. – Online Information Review, 2020.
15. Su, M. H., J. Liu, D. M. McLeod. Pathways to News Sharing: Issue Frame Perceptions and the Likelihood of Sharing. – Computers in Human Behavior, Vol. **91**, 2019, pp. 201-210.
16. Bhagat, S., D. J. Kim. Examining Users' News Sharing Behavior on Social Media: Role of Perception of Online Civic Engagement and Dual Social Influences. – Behavior & Information Technology, 2022, pp. 1-22.
17. Trilling, D., J. Kulshrestha, C. de Vreese, D. Halagiera, J. Jakubowski, J. Moeller, C. Vaccari. Is Sharing Just a Function of Viewing? – Predictors of Sharing Political and Non-Political News on Facebook. 2022.
18. Prado-Romero, M. A., A. Coto-Santiesteban, A. Celi, G. Stilo. A Time-Sensitive Model to Predict Topic Popularity in News Providers. – Intelligent Data Analysis, Vol. **24**, 2020, No S1, pp. 123-140.
19. Meghawat, M., S. Yadav, D. Mahata, Y. Yin, R. R. Shah, R. Zimmermann. A Multimodal Approach to Predict Social Media Popularity. – In: IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'18), IEEE, 2018, pp. 190-195.
20. Manzoor, S. I., J. Singla. Fake News Detection Using Machine Learning Approaches: A Systematic Review. – In: 3rd IEEE International Conference on Trends in Electronics and Informatics (ICOEI'19), 2019, pp. 230-234.
21. Gerunov, A. Performance of 109 Machine Learning Algorithms across Five Forecasting Tasks: Employee Behavior Modeling, Online Communication, House Pricing, IT Support, and Demand Planning. – Economic Studies, Vol. **31**, 2022, No 2, pp. 15-43.
22. Fernandes, K., P. Vinagre, P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. – In: Portuguese Conference on Artificial Intelligence. Cham., Springer, 2015, pp. 535-546.
23. Gerunov, A. Automated Approaches for Operational Risk Management. Sofia University "St. Kliment Ohridski", 2020 (in Bulgarian).
24. Blei, D. M., A. Y. Ng, M. I. Jordan. Latent Dirichlet Allocation. – Journal of Machine Learning Research, Vol. **3**, January 2003, pp. 993-1022.

25. Yue, L., W. Chen, X. Li, W. Zuo, M. Yin. A Survey of Sentiment Analysis in Social Media. – Knowledge and Information Systems, Vol. **60**, 2019, pp. 617-663.
26. Ratkiewicz, J., S. Fortunato, A. Flammini, F. Menczer, A. Vespignani. Characterizing and Modeling the Dynamics of Online Popularity. – Physical Review Letters, Vol. **105**, 2010, No 15, 158701.
27. Fernández-Delgado, M., E. Cernadas, S. Barro, D. Amorim. Do We Need Hundreds of Classifiers to Solve Real-World Classification Problems? – The Journal of Machine Learning Research, Vol. **15**, 2014, No 1, pp. 3133-3181.
28. He, X., K. Zhao, X. Chu. AutoML: A Survey of the State-of-the-Art. – Knowledge-Based Systems, Vol. **212**, 2021, 106622.
29. Yao, Q., M. Wang, Y. Chen, W. Dai, Y. F. Li, W. W. Tu,. Y. Yu. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. – arXiv preprint arXiv:1810.13306, 2018.
30. Hutter, F., L. Kotthoff, J. Vanschoren. Automated Machine Learning: Methods, Systems, Challenges. – In: Springer Nature Link, 2019, p. 219.
31. LeDell, E., S. Poirier. H2O Autonomy: Scalable Automatic Machine Learning. – In: Proc. of AutoML Workshop at ICML, Vol. **2020**, July 2020.
32. R. G. Cowell, Ed. Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks. Springer Science & Business Media, 2006.
33. Korb, K., A. E. Nicholson. Bayesian Artificial Intelligence. 2nd Edition, Chapman & Hall/CRC, 2010.
34. Højsgaard, S. Graphical Independence Networks with the gRain Package for R. – Journal of Statistical Software, Vol. **46**, 2012, No 10, pp. 1-26.
35. Wang, C. H., H. Y. Cheng, Y. T. Deng. Using Bayesian Belief Network and Time-Series Model to Conduct Prescriptive and Predictive Analytics for Computer Industries. – Computers & Industrial Engineering, Vol. **115**, 2018, pp. 486-494.
36. Scutari, M., C. E. Graafland, J. M. Gutiérrez. Who Learns Better Bayesian Network Structures: Constraint-Based, Score-Based, or Hybrid Algorithms? – In: Proc. of International Conference on Probabilistic Graphical Models PMLR, August 2018, pp. 416-427.
37. Gámez, J. A., J. L. Mateo, J. M. Puerta. Learning Bayesian Networks by Hill Climbing: Efficient Methods Based on Progressive Restriction of the Neighborhood. – Data Mining and Knowledge Discovery, Vol. **22**, 2011, No 1, pp. 106-148.
38. Gerunov, A. Networks of Risk. – In: Risk Analysis for the Digital Age. Cham, Springer International Publishing, 2022, pp. 115-156.
39. LeDell, E., S. Poirier. H2O Autonomy: Scalable Automatic Machine Learning. – In: Proc. of AutoML Workshop at ICML, Vol. **2020**, July 2020, San Diego, CA, USA, (ICML'20).
40. Watson, J., S. van der Linden, M. Watson, D. Stillwell. Negative Online News Articles Are Shared More on Social Media. – Scientific Reports, Vol. **14**, 2024, No 1, 21592.
41. van der Meer, T. G., A. Brosius. Credibility and Shareworthiness of Negative News. – Journalism, Vol. **25**, 2024, No 1, pp. 61-80.
42. Mathews, N., V. Bélair-Gagnon, S. C. Lewis. News is "Toxic": Exploring the Non-Sharing of News Online. – New Media & Society, Vol. **26**, 2024, No 8, pp. 4629-4646.

Appendix A. Residual analysis

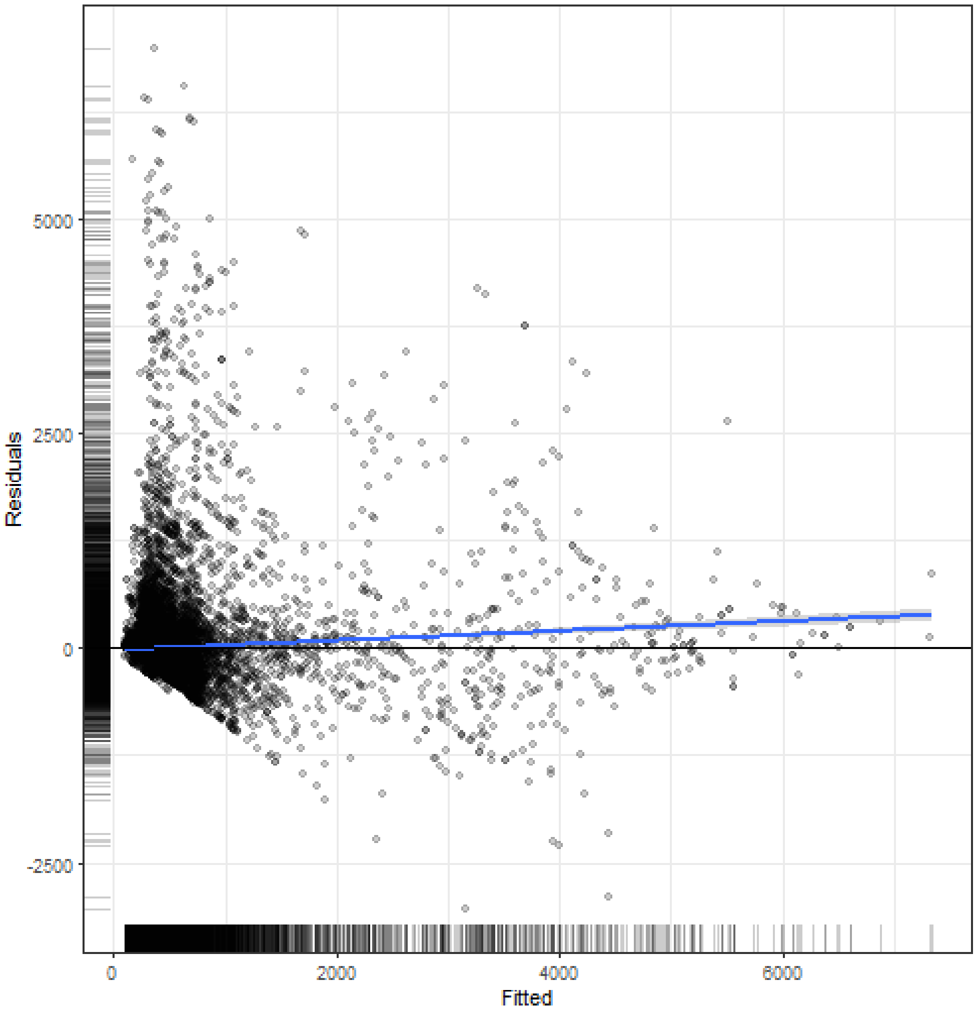

Fig. 1A. Residual analysis plot for best performing AutoML gradient boosting model

Received: 05.12.2024, First Revision: 30.01.2025, Second Revision: 19.02.2025, Accepted: 21.02.2025