Modeling Cascade Growth: Predicting Content Diffusion on VKontakte

Anna Moroz[1][0000−0001−9442−0580], Sergei Pashakhin[1][0000−0003−0361−2064] and Sergei Koltsov[1][0000−0002−2932−2746]

Laboratory for Social & Cognitive Informatics, National Research University
Higher School of Economics, Saint Petersburg, Russia
{asmoroz;spashahin;skoltsov}@hse.ru

Abstract. Online social networks have become an essential communication channel for the broad and rapid sharing of information. Currently, the mechanics of such information-sharing is captured by the notion of cascades, which are tree-like networks comprised of (re)sharing actions. However, it is still unclear what factors drive cascade growth. Moreover, there is a lack of studies outside Western countries and platforms such as Facebook and Twitter. In this work, we aim to investigate what factors contribute to the scope of information cascading and how to predict this variation accurately. We examine six machine learning algorithms for their predictive and interpretative capabilities concerning cascades' structural metrics (width, mass, and depth). To do so, we use data from a leading Russian-language online social network VKontakte capturing cascades of 4,424 messages posted by 14 news outlets during a year. The results show that the best models in terms of predictive power are Gradient Boosting algorithm for width and depth, and Lasso Regression algorithm for the mass of a cascade, while depth is the least predictable. We find that the most potent factor associated with cascade size is the number of reposts on its origin level. We examine its role along with other factors such as content features and characteristics of sources and their audiences.

Keywords: News diffusion · Machine learning · Information cascades · Online social networks · Cascade size prediction.

1 Introduction

Today's world is overloaded with an enormous amount of information circulating through various environments. Some environments – emerged from the development of technologies, and their penetration in individuals' daily life – became widely and skillfully exploited by media on the global level [1]. Social networking platforms and services as one of such channels provide individuals with facilities for rapid communication and sharing of texts, photos, videos, links to external resources, or any other digital pieces of information with other users [2].

The mechanism of such information spreading, in general, involves a source, that posts information in the first place, and a circle of users exposed to this content. Users could be tied to a source by friendship/followership relations, comprising an audience, or accidentally being exposed to the post. Then, some of this audience may repost the message and become a source for their audience. This chain of sharing actions generates a hierarchical tree of information resharing, usually referred to in the literature as a cascade [3,4]. Cascades capture information spread better than the 'small world' model [5,6], and allow studying social influence in networks [4,7]. Although information cascades are capable of reaching an enormous number of users, they vary in size and rarely become large [8]. It is still unclear what factors contribute to the scope of information cascading and how to predict this variation accurately [9].

Currently, there are two approaches to this task: generative (deductive) and feature-based (inductive) [10]. The generative approach involves characterizing and modeling the process of content becoming popular in a social network. Although it provides excellent interpretability, it predicts poorly the variation observed in real-world cascades and may miss possibly valuable predictors. The feature-based approach formulates this task as a regression/classification problem that could be solved using a learning algorithm and a set of features with varying contributions to the explained variance, providing a framework for both prediction and explanation.

Previous works related to the prediction of cascade growth had been investigating factors connected to it. Although some progress in successful prediction has been achieved, a consensus on what features are the most essential to it is not established. Specifically, Cheng and others in [9], aiming to predict whether the size of information cascade will exceed the predetermined number, discovered that content features of an original (root) post (i.e., attached images and captions of the post), although being weak predictors on their own, affect the influence of structural (friendship/followership networks' properties) and temporal (the speed of reshares) features. Simultaneously, they report that the average connectivity of the first reposters contributes to the increasing accuracy of prediction. An alternative finding concerning the influence of content features was reported in [11], in which features of an author of a tweet along with tweet features appeared to be the most essential for prediction. Another study by Hong and colleagues [12] indicated that the best model performance is achieved when contextual features are used along with temporal ones and that user activity aspects enhance marginal predicting performance. One more work [13] experimented on features sets for prediction by applying a hybrid methodology for feature selection. The results were that channel (information source) features, i.e., the author's followers, and content features (whether a post contains URL and/or image), were repeatedly given the best rank by several feature selection algorithms. In the interim, a study by Tsur and others [14] proved that a hybrid model incorporating several feature types (i.e., contextual, structural, and temporal ones) predicts information dissemination much better than partial models, evidencing that no predicting feature type's influence is concealed by the presence of other predicting feature types. There was also an attempt to propose a model able to integrate all notable findings of the previous research in [15], that was successfully empirically tested on the Twitter resharing cascades afterward. However, the authors' highest result of prediction hardly accounted even for a half of the variation in the cascade size.

Additionally, studies on the topic had reported that complex models outperform simpler models applied for prediction. Precisely, the finding is valid for the work by Cao and co-authors [10] that involved a deep learning extension of the Hawkes processes model in comparison to regularized linear regression, both utilized to predict the influence of a retweet path. The already mentioned work by Cheng and colleagues [9] also proved that non-linear algorithms perform a little better than linear ones when predicting the variation in a cascade's size.

In this work, we seek to perform an accurate prediction of cascade growth using six types of regression machine learning algorithms with a wide set of feature categories. We opt for the regression task for two reasons. The size of a propagation cascade, with no additional transformations implemented, is a numeric value. The classification formulation requires transforming this value either to a binary or multinomial variable, which involves dividing the range of values into categories based on some threshold. As cascade's size is commonly power-law distributed [9,16,17], defining such threshold is problematic and reduces the amount of likely valuable information. Hence, this study aims to determine which algorithms are the best for predicting cascade's growth, and to explore which features are the most strongly associated with the eventual scope of a cascade. Although we consider factors associated with information propagation, we aim for a methodological contribution: to investigate the applicability of several learning algorithms to the prediction of cascades observed in a real online network.

Although the most popular social networking service for studying information propagation is Twitter, we choose VKontakte (VK) to address the lack of studies on Russian-speaking platforms and possible discrepancies connected to local audiences and media. Moreover, VK is the most popular Russian-speaking service, similar to Facebook, with an audience of 73.4 million users per month [18].

We divide a notion of cascade's size into three differentiated metrics: its depth, width, and mass. We apply the notion of a cascade level, considering that there is a hierarchy of reposting from node to node. Thus, we define these metrics as follows:

- **width** is the biggest amount of nodes at one of the levels of a resharing cascade;
- **mass** is the total number of nodes at all levels of a cascade;
- **depth** is the maximum number of levels in a propagation cascade.

The rest of this paper is organized in the following way. The next section discusses the data used in the study. Then we elaborate on the methodological pipeline, after which the results of modeling are summarized. Finally, section 5 provides conclusions and considers the meaning and value of the findings.

2 Data

2.1 Dataset Description

As it was already mentioned, the data needed for modeling and further analysis were retrieved from VK service. The scope of the data was narrowed to the official public pages of leading state-owned Russian television channels (see Table 2).

The data included (1) reposting data – the chains of reposts for top posts from the channels, (2) reposters data – publicly available profile information of users who at least once have reshared one of the top posts at any level of a cascade, (3) channel's summary statistics – total numbers of posts' comments, likes to posts, likes to posts' comments for each channel computed for the top posts, (4) post metadata – the popularity figures for each post (i.e., post's number of comments, likes, and reposts, the latter accounting for the number of nodes on the first level of a resharing cascade), and, finally, (5) topic modeling data, a matrix containing probability distributions of 86 labeled topics over news texts posted on behalf of the channels, so that each studied information cascade has some probability of belonging to all of the topics. The topic modeling procedure was reported in [19].

The full set of features used for prediction can be found in Table 1. The final dataset used for fitting and assessing the models consisted of 4,424 observations. Its detailed description by sampled channels is provided in Table 2.

Table 1. List of features within each feature category

Feature Category	**Feature**	**Data Type**
Root post features	N of comments to each root post	Numeric
	N of likes to each root post	Numeric
	N of reposts of each root post on the first level of a cascade	Numeric
Channel features	The total N of comments to channel's top posts	Numeric
	The total N of likes to channel's top posts	Numeric
	The total N of likes to comments to channel's top posts	Numeric
	N of channel's followers	Numeric
Channel's audience features	The average age of those subscribed to a channel	Numeric
	The cumulative N of followers of those subscribed to a channel	Numeric
	The prevalent sex ('1' for female and '2' for male) of those subscribed to a channel	Binomial
Content features	Distributions of probabilities of shared information, text of a root post, belonging to each of the established by topic modeling procedure [19] topics. 86 separated variables in total.	Numeric

Table 2. Dataset description.

News Community	N of Posts	N of Reposts	N of Unique Reposters	N of 1st-level Reposts	Max Cascade's Width	Max Cascade's Depth	Max Cascade's Mass
RIA News	953	44,703	27,871	40,325	1,048	10	1,151
Russia Today	944	17,661	11,725	15,772	591	6	749
RBC	706	28,030	14,313	24,536	665	8	1,169
Dozhd	593	9,233	5,892	7,593	278	8	405
NTV	441	10,262	7,480	8,448	1,279	12	1,413
Russia 24	252	8,072	6,120	6,733	784	6	907
Russia-Culture	123	3,530	2,050	2,704	109	9	273
MIR24	113	5,412	1,264	4,984	153	8	176
Channel-5	101	2,960	2,312	2,208	126	7	217
Channel 1	72	23,755	18,261	20,656	1,170	6	1,627
TVC (News)	64	392	311	270	10	4	18
Russia-1	51	5,505	3,246	3,281	332	9	614
Monson	8	1,684	1,435	1,500	234	4	255
InoTV	3	15	15	15	6	1	6

2.2 Training and Testing Datasets

The compiled dataset was split into two samples: the training set for fitting models, and the test set one for assessing their performance, as the standards of predictive machine learning modeling entail [20].

There are several ways of splitting the initial dataset, and the most common strategy is to randomly assign 80% of observations to the training sample and 20% to the testing sample. In the case of our data, we can split the dataset by date of posts' publication. Thus, models fitted on the older data would be extrapolating to more recent cascades with more reliable predictions in terms of external validity. Hence, the data was split as follows: the training set included all posts up to October 2017 (including), and the test set covered the period from November 2017 to February 2018. Ultimately, the convention of 80% by 20% data split was met.

2.3 Normalization

In our case, the target variables – width, depth, and mass of a cascade – were found to be power-law distributed. To comply with the assumptions of the penalized linear regression models – requiring a normally distributed outcome variable – the resulting data was duplicated, and normalization procedure was applied so that two sets of data were obtained, non-normalized and normalized. For normalization, the scaling of matrix-like objects algorithm [21] was applied.

3 Methods

3.1 Algorithms and Hyperparameters

For modeling cascade's metrics, a set of algorithms of increasing complexity with 5-fold cross-validation was chosen. The set included three kinds of penalized linear regressions, i.e., Lasso regression, Ridge regression, and Elastic Net regression, Decision Tree algorithm, and more complex ensemble methods, such as the Random Forest and Gradient Boosting Machine algorithms.

The hyperparameters were set to default for Decision Tree, as it was found that changing hyperparameters does not affect performance significantly on this data after a series of preliminary tests. In the case of Random Forest, the picking-up procedure revealed the most optimal hyperparameters for the algorithm. As for Gradient Boosting, the algorithm automatically chose the hyperparameters out of a specified list of values based on the RMSE metric. The same was done for all three types of penalized linear regressions. The detailed settings of hyperparameters are present in Table 3.

3.2 Fitted Models

Models were built in a three-stage procedure aimed at defining the best-fit for predicting cascades' growth. During the first step, three algorithms, Decision Tree, Random Forest, and Gradient Boosting, were fit to the non-normalized training data, each of the algorithms predicting three cascade's metrics separately. After that, the same was reproduced on the normalized training dataset, with penalized regressions added. Finally, a total of 27 models, with 9 of them predicting each of three target values, were run on the testing sets and compared in prediction accuracy based on the R-squared metric. This metric was chosen as the only one that can be applied to both scaled and non-scaled datasets, disregarding discrepancy in variables' value range. In situations when R-squared values were about the same within one of the cascades' metrics prediction, the RMSE measure, calculated as the square root of a mean squared difference between predicted and observed outcomes, was additionally used.

3.3 Interpretation

After the best-fitting models predicting depth, width, and mass of a cascade were chosen, several methods were applied for the purpose of interpretation. Some of the utilized algorithms, i.e., linear regressions, are easily interpretable on their own, producing coefficients for all of the involved independent variables needed to specify how inputs interact with each other to generate the output.

However, other algorithms – sometimes called "black box" models – are more complex, which hampers the interpretability of results. Considering this fact, for interpretation of the Gradient Boosting models, the relative influence of explanatory features was used, which is the number of times a feature was selected for

Table 3. List of tuned hyperparameters

Algorithm	**Parameter name**	**Parameter Description**	**Range**
Decision Tree	*mincriterion*	A value of test statistics or (1-p) that should be exceeded to perform a split	0 (*default*)
Random Forest	*mtry*	The number of features randomly sampled as candidates at each split	50
	ntree	the total number of trees to grow in one run	500
Gradient Boosting Machine	*interaction.depth*	number of splits an algorithm has to perform on a single tree	6 (*Salford default setting* [22])
	n.trees	The total number of trees to grow in one run	500, 800, 1000, 2000
	shrinkage	A learning rate, stands for the amount of penalty that will be applied to reduce the effect of each additional tree built	0.001, 0.005, 0.01, 0.05, 0.1, 0.5
	n.minobsinnode	The minimum number of observations in terminal nodes of each tree	10, 15, 20, 25
Lasso Regression	*lambda*	Determines the amount of shrinkage, the penalty term, to be applied to regularize the effect of predicting features	$10^{seq(-3,3,length=100)}$
	alpha	Determines what type of penalized regression model is fit	1
Ridge Regression	*lambda*	Determines the amount of shrinkage, the penalty term, to be applied to regularize the effect of predicting features	$10^{seq(-3,3,length=100)}$
	alpha	Determines what type of penalized regression model is fit	0
Elastic Net Regression	*lambda*	Determines one of the amounts of shrinkage, the first penalty term, to be applied to regularize the effect of predicting features	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
	alpha	Determines one of the amounts of shrinkage, the second penalty term, to be applied to regularize the effect of predicting features	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1

each tree split balanced by the improvement of the SSE (sum of squared errors) value and averaged over all trees [23].

In the case of Random Forest, the variable importance score was used. The scores are computed using MSE (mean square error) first obtained on the subsample of data for each tree that was not used during model construction, and once again with variables reshuffled. The differences are then averaged and divided by the standard error. Lastly, to infer the significance of independent variables of penalized linear regression, the coefficients as the estimators of features' importance, controlled by a penalty term, were obtained. After the order of importance of predicting features in each considered model was acquired, the individual conditional expectation plots (ICE plots) were constructed to decide the directionality of the relationship between the outcome and the most important features. The partial dependence plot displays the marginal effect of one or more features on the predicted outcome, and it was used to assess the direction of the relation between an outcome and a feature [24].

Finally, the whole procedure was repeated without variable with the number of reposts at the first level of resharing to investigate its relative importance for reasons discussed below.

4 Results

Table 4 shows the performance of all nine models predicting the mass of a cascade on both non-normalized and normalized data. Due to differently preprocessed sets of the data, values of RMSE for the first three and the remaining six models fall into different number ranges. However, it is still possible to compare R-squared values among nine models, and RMSE values within models built on the same data. The highest performance on cascade's mass prediction is attributed to the penalized linear regression – the percentage of explained variance in the outcome goes beyond 97%. Among these three models, the most accurate appears to be the Lasso regression, having the smallest root mean square error.

R-squared and RMSE metrics of nine models predicting cascade's width and nine models for cascade's depth are laid out in the same table, respectively, with the very same technical peculiarities mentioned for the mass predicting models' performance. The best models for prediction of cascade's width are Gradient Boosting Machine on non-normalized data and Lasso and Elastic Net regressions on the normalized dataset, achieving the highest R-squared values of approximately 0.99, and the lowest RMSE values. Finally, in the case of depth of a cascade, the best predictive performance is shown by the Gradient Boosting algorithm on both datasets, with 57% explained variance in the outcome variable. Judging by lower RMSE value, this model built on the non-normalized data is slightly more accurate.

Additionaly, it should be noted that width and depth of information cascade are predicted more accurately, with models accounting for almost 100% of the variance in the target variable while cascade's depth was predicted poorly, com-

Table 4. Models' performance figures.

Cascade's Metric	**Model**	**R-squared Value**	**R-squared Value** (with one feature out)	**RMSE**	**RMSE** (with one feature out)
Mass	Decision Tree (*non-scaled data*)	0.962	0.247	20.226	89.598
	Random Forest (*non-scaled data*)	0.966	0.252	18.540	90.018
	Gradient Boosting (*non-scaled data*)	0.97	0.242	17.670	88.503
	Lasso Regression (*scaled data*)	0.972	0.230	0.166	0.888
	Ridge Regression (*scaled data*)	0.970	0.178	0.196	0.906
	Elastic Net Regression (*scaled data*)	0.972	0.232	0.168	0.888
	Decision Tree (*scaled data*)	0.960	0.024	0.208	1.861
	Random Forest (*scaled data*)	0.923	0.084	0.320	2.039
	Gradient Boosting (*scaled data*)	0.966	0.059	0.182	2.156
Depth	Decision Tree (*non-scaled data*)	0.539	0.089	0.673	0.957
	Random Forest (*non-scaled data*)	0.550	0.117	0.662	0.949
	Gradient Boosting (*non-scaled data*)	0.571	0.136	0.648	0.921
	Lasso Regression (*scaled data*)	0.316	0.115	0.828	0.950
	Ridge Regression (*scaled data*)	0.334	0.120	0.815	0.971
	Elastic Net Regression (*scaled data*)	0.305	0.103	0.836	0.951
	Decision Tree (*scaled data*)	0.511	0.034	0.70	1.059
	Random Forest (*scaled data*)	0.560	0.113	0.666	0.981
	Gradient Boosting (*scaled data*)	0.572	0.112	0.665	0.956
Width	Decision Tree (*non-scaled data*)	0.969	0.312	15.260	68.407
	Random Forest (*non-scaled data*)	0.980	0.319	11.817	70.180
	Gradient Boosting (*non-scaled data*)	0.984	0.320	11.064	67.911
	Lasso Regression (*scaled data*)	0.984	0.266	0.122	0.867
	Ridge Regression (*scaled data*)	0.983	0.201	0.155	0.893
	Elastic Net Regression (*scaled data*)	0.984	0.270	0.122	0.867
	Decision Tree (*scaled data*)	0.964	0.058	0.196	2.528
	Random Forest (*scaled data*)	0.924	0.107	0.333	2.121
	Gradient Boosting (*scaled data*)	0.981	0.070	0.137	2.405

pared to other metrics of cascade's size, reaching maximum of 57% of explained variance.

The following models were considered for interpretation: Lasso regression for cascade's mass prediction and Gradient Boosting built on the non-normalized dataset predicting both width and depth of a cascade. The most prominent feature associated with cascade's mass was the number of reshares on the first level of a cascade. On the relative scale of overall variable importance, assessed on the base of absolute values of regression coefficients, ranging from 0 to 100, this feature scored the maximum. Figure 1 (see in Appendix) depicts the mentioned variation. Further conclusions about features' importance were drawn on the base of the same method, partial dependence plots. The same was observed for other metrics – their increase was strongly associated with the number of reposts on the first level of a cascade. In the model predicting cascade's width, this feature reached 97 out of 100 points on the normalized scale of features' relative influence. In predicting the depth, it scored 70 out of 100, compared to other features that barely exceeded 3 points.

Thus, most of the variation in cascade's size can be explained by the change in the number of reshares on the first level of the information propagation tree. This relation itself appeared to be positive, meaning that the increase in the number of reposts on first-level is attributed to the growth of a cascade in depth, width, and mass (see Fig. 3 and Fig. 2 in Appendix).

Among other features, relatively significant appeared to be content features: probability distributions of several text topics, all of the used channel's audience features, one of the root poster features, and channel features. For prediction of cascade's mass, the number of reposts on the first level of a cascade is followed by prevailing sex of channel's followers, the number of channel's followers, the cumulative number of channel's audience' followers, the number of comments to channel's posts, the number of root post's likes, and average age of channel's audience. According to the coefficients (see Fig. 2 in Appendix), the prevalence of male users among channel's followers and the number of channel's followers are negatively associated with the mass of a cascade, while the rest of the mentioned features – positively.

As for the cascade's width prediction, the number of reposts on the first level is followed by the probability distributions of the following topics: "West-Russia relations", "The Voice Russia", "TV shows: comedy and dancing", and "Weather". They appeared to be positively correlated with the cascade's width, as the marginal effect of each topic exceeds zero. Finally, for the depth prediction the number of comments to channel's posts, the number of channel's audience' followers, and the number of likes to channel's posts are relatively important. Interestingly, only the number of reposts at the first level of resharing, among all mentioned features, is positively associated with the depth of a cascade.

As the number of reposts at the first level of propagation cascade accounted for the overwhelming part of the variance in all three metrics modeled, we propose that the growth of a cascade can be attributed solely to this feature. To investigate this proposition, we repeated the whole procedure without this variable. As a result, the set of best-fit algorithms has changed (see Table 4). Now, the best model for predicting the mass of a cascade is Random Forest trained and assessed on non-normalized data. As for cascade's width and depth, Gradient Boosting Machine on the non-normalized dataset emerged as most accurate. R-squared values of all nine models for each metric substantially decreased, not reaching even half of the value of their counterparts in a situation when the removed feature was present.

Further, scaling of the dataset for prediction of all three metrics gives a substantial fall in R-squared values compared to those resulting from the non-scaled data. As for the most crucial features, the cumulative number of channel's audience followers, content characteristics, and the number of comments to channel's posts explain most of the variation. Surprisingly, the increase in the latter shows a negative association with cascade's depth, while others are positively associated with cascade's metrics.

Additionally, we considered marginal effects of age and sex of a channel's audience on the cascade's size in the models without the number of reposts on the first level. The prevailing sex of a channel's audience indicated by users as "male" is negatively correlated with the size of a cascade. The marginal effect of the average age of a channel's audience on cascade's mass equals zero for most of the predictor values except for the value of 33.84, where it falls below zero. Its marginal effect on the width of a cascade is negative, while on the cascade's depth, it is the exact opposite, indicating positive association. However interesting, we abstain from an in-depth analysis of factors associated with cascade growth and their contribution and limit our interpretations as we aim to investigate the methodological value of the findings.

5 Conclusions & Discussion

In this work, we examine the issue of information spread in online social networks by attempting to find the best-fit model that can be used to predict the growth of propagation cascades. Although there are existing studies that have addressed similar objectives, we contribute to the research by approaching the problem with the data from Russian-speaking social networking service. After obtaining the best-fit models, we look at the features that contributed the most in order to determine the possibly noteworthy constituents of the models able to predict variation in the outreach of an information cascade.

The results showed that non-linear algorithms perform better when predicting the growth of information cascade – which is consistent with [10] and [9] works. Except for the prediction of mass of a cascade with a full set of features, which is most accurately predicted with regularized linear regression. Furthermore, it should be noted that R-squared values for the models with all features included are objectively quite high, indicating that models can explain up to 98% of the variation in cascade's size – in contrast to Martin and his colleagues' study [15] which achieved a maximum of 48% of explained variation.

The interpretation of the best-fit models indicated that features of all categories either way appeared as one of the most contributing to the prediction, similarly to [14]. However, not all features have the same degree of importance for the accuracy of prediction, which is more in line with [9]. The growth of an information cascade in width, depth, and its gain in mass can be attributed entirely to the number of reposts on the first level of a cascade – formally, to the number of edges directly connected to the root of a cascade, while other features have a relatively small effect on the outcome when included in a model altogether. Although logically evident, this finding raises questions whether this connection can be explained by the propagation mechanics and the data specificity, or is it an artifact of the platform where the study was conducted. It should be noted that on VK, there is no way a user can see the number of likes, reshares, or views of the original post at the point when this post was reposted, i.e., from the account of a user who did the repost. In addition to that, when a repost is done, a reposter is able to place a caption to the reshared post – fairly modifying resharing information.

Nevertheless, a sharp decrease in R-squared value by three times for all models, after removing the number of reposts on the first level of resharing from the predicting features, suggests that the eventual outreach of content diffusion can be predicted with high accuracy by this feature solely. Hence, we can conclude that the observed structural variation in depth and width of a propagation tree, and its eventual magnitude highly depends on the activity of the source' audience that serves as an intermediary in letting the information leak beyond where it was initially posted. Note that such an audience includes not only followers of a channel but also users who are not bounded by the followership relations to a channel.

Additionally, the results of [11] study were also partly supported – features related to an author of a post (in our case, channel features) combined with content features benefited the predictive power of the models, yet only when the number of shares on the first level was excluded. This finding lets us suggest that if we assume that there are specific scripting strategies that each channel uses to differentiate itself and the content it posts from other news channels, the eventual reach of information propagation depends on the source's (channel's) specificity. Furthermore, users that make the propagation possible – in line with the selective exposure theory [25] – intentionally choose from what source and what news information to consume depending on the attributes of both source and content.

Content features on their own were consistently important for predicting cascade' s growth with models using all features and models without the number of shares on the first level, compared to other features. This validates conclusions of work by Hong and colleagues [12]. However, predicting models without the number of reposts on the cascade's first level let us conclude that, contrary to Elsharkawy and colleagues' work [13], audience features are more relevant for prediction of cascade's growth than channel's features. Further, if cascade's growth can be relatively successfully predicted by audience features, a question about users' similarity or dissimilarity as a possible driver of information diffusion can be considered.

Finally, channel features and one of its audience features, prevailing sex of channel's audience indicated as male, that showed negative association when reaching a certain value with the depth of a cascade, raise suspicions. In the case of channel features, it appears that the number of followers of the author of the content (a channel) – simply put, channel's popularity – has a negative influence on information propagation depth. We can speculate that such a phenomenon can be linked to the reputation of Russian news channels' representation in the chosen social network. The online channels are assumed to be perceived by users of this social network as not trustworthy enough to 'participate' in spreading its content. Besides, the negative association between the number of channel's followers and the depth of a diffusion tree can be attributed to specifics of the publication sorting mechanism of the VKontakte platform. Still, further research is needed for the explicit connection and rationale of such a pattern to be established. As for male audiences having a negative impact on the cascade's growth, it can be allegedly explained by male users' tendency to less frequently become intermediate recipients (brokers) of information in a network of information diffusion, compared to female users. Yet, to claim it as a fact, more research on the topic should be conducted.

Limitations

It is important to note that the reposting data used in the study is news posts published on official communities of Russian television channels, making the results valid only for such or a similar sample. This study's findings cannot be generalized on reposting patterns of any other content on social media platforms except for news data. Another limitation of our research is connected to metrics used to assess models. In the case of "black box" models, it is debatable whether the R-squared value can be applied for assessment of their predictive power.

Acknowledgements

This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

References

1. Thorson, K., Wells, C.: Curated flows: A framework for mapping media exposure in the digital age. Communication Theory 26(3), 309–328 (2015)
2. Boyd, D.M., Ellison, N.B.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication 13(1), 210–230 (2007)

3. Sun, E., Rosenn, I., Marlow, C.A., Lento, T.M.: Gesundheit! modeling contagion through facebook news feed. In: Third international AAAI conference on weblogs and social media (2009)
4. González-Bailón, S., Borge-Holthoefer, J., Moreno, Y.: Online networks and the diffusion of protest. Analytical Sociology p. 261–278 (2014)
5. Liben-Nowell, D., Kleinberg, J.: Tracing information flow on a global scale using internet chain-letter data. Proceedings of the National Academy of Sciences 105(12), 4633–4638 (2008)
6. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. ACM Transactions on Knowledge Discovery from Data 5(4), 1–37 (Jan 2012)
7. Myers, S.A., Zhu, C., Leskovec, J.: Information diffusion and external influence in networks. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 12 (2012)
8. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyones an influencer. Proceedings of the fourth ACM international conference on Web search and data mining - WSDM 11 (2011)
9. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted? Proceedings of the 23rd international conference on World wide web - WWW 14 (2014)
10. Cao, Q., Shen, H., Cen, K., Ouyang, W., Cheng, X.: Deephawkes: Bridging the gap between prediction and understanding of information cascades. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 1149–1158 (2017)
11. Petrovic, S., Osborne, M., Lavrenko, V.: Rt to win! predicting message propagation in twitter. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
12. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. Proceedings of the 20th international conference companion on World wide web - WWW 11 (2011)
13. Elsharkawy, S., Hassan, G., Nabhan, T., Roushdy, M.: Towards feature selection for cascade growth prediction on twitter. Proceedings of the 10th International Conference on Informatics and Systems - INFOS 16 (2016)
14. Tsur, O., Rappoport, A.: What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In: Proceedings of the fifth ACM international conference on Web search and data mining. pp. 643–652 (2012)
15. Martin, T., Hofman, J.M., Sharma, A., Anderson, A., Watts, D.J.: Exploring limits to prediction in complex social systems. In: Proceedings of the 25th International Conference on World Wide Web. pp. 683–694 (2016)
16. Leskovec, J., Mcglohon, M., Faloutsos, C., Glance, N., Hurst, M.: Patterns of cascading behavior in large blog graphs. Proceedings of the 2007 SIAM International Conference on Data Mining (2007)
17. Vicario, M.D., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the National Academy of Sciences 113(3), 554–559 (Apr 2016)
18. Mail.ru Group Limited Annual Report for FY 2019 andunaudited IFRS results for Q1 2020 (Apr 2020), https://corp.imgsmail.ru/media/files/engq1-2020-results.pdf
19. Koltsov, S., Pashakhin, S., Dokuka, S.: A full-cycle methodology for news topic modeling and user feedback research. In: International Conference on Social Informatics. pp. 308–321. Springer (2018)

20. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning, vol. 112. Springer (2013)
21. Becker, R.A., Chambers, J.M., Wilks, A.R.: The new s language (Apr 2018)
22. TreeNet stochastic gradient boosting: An implementation of the MART methodology, `http://docs.salford-systems.com/TreeNetManual_v1.pdf`
23. Quan, Z., Valdez, E.A.: Predictive analytics of insurance claims using multivariate decision trees. SSRN Electronic Journal (2018)
24. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
25. Sullivan, L.E.: Selective exposure. The SAGE Glossary of the Social and Behavioral Sciences p. 465 (2009)

A Appendix

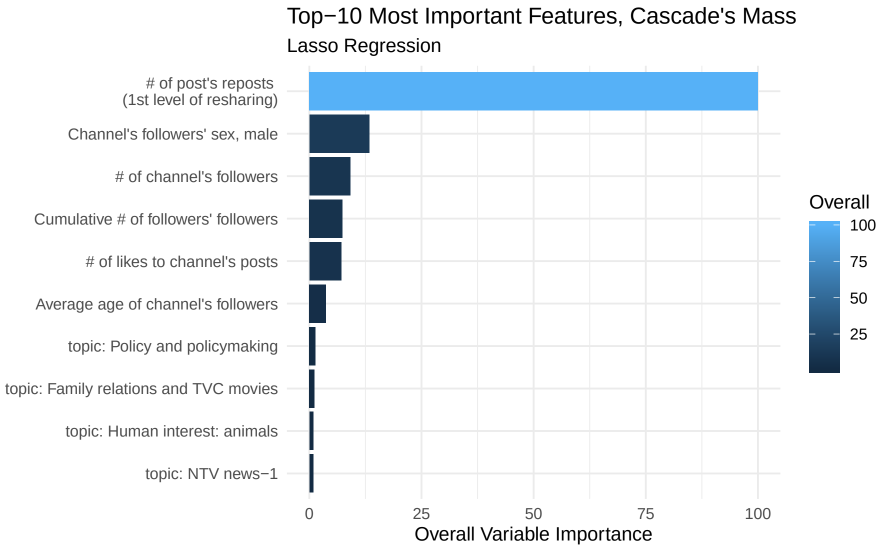

Fig. 1. A graph showing top-10 of the most tangible predicting features used as input by the Lasso Regression algorithm for cascade's mass prediction.

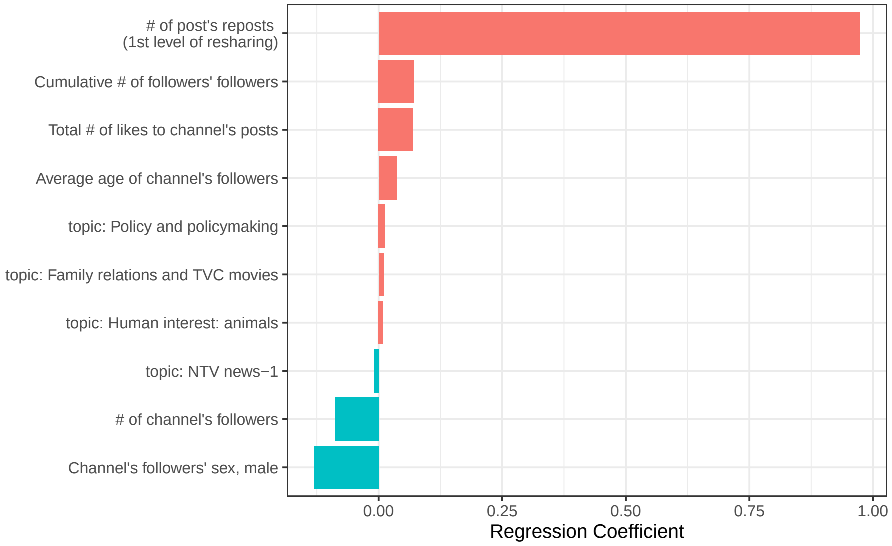

Fig. 2. A plot displaying negative to positive values proportion of Lasso Regression variables' coefficients.

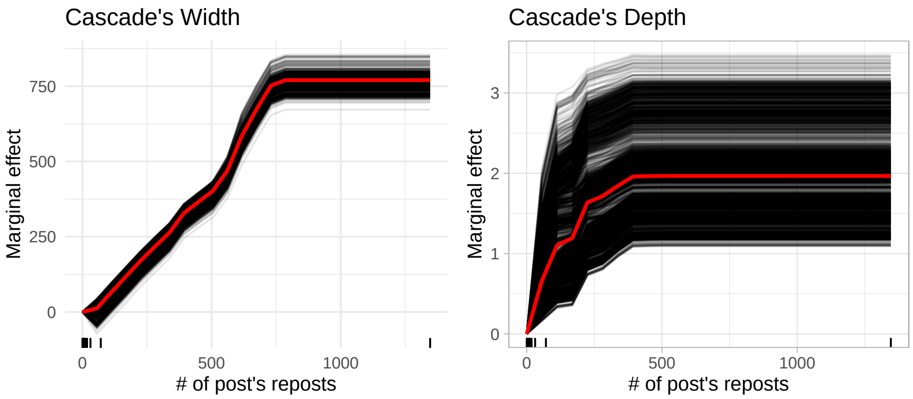

Fig. 3. Plots showing observation-level effects of the number of posts first-level reposts on cascade's depth (on the left) and width (on the right). The multiple black lines are individual conditional expectation (ICE) curves, while the red line stands for the averaged values across all predictions [24] by Gradient Boosting algorithm.