symmetry

MDPI

Article

Stochastic Diffusion Model for Analysis of Dynamics and Forecasting Events in News Feeds

Dmitry Zhukov [1], Elena Andrianova [2] and Olga Trifonova [3,*]

[1] The Institute of Integrated Safety and Special Instrument Engineering, MIREA—Russian Technological University, 78, Vernadskogo pr., 119454 Moscow, Russia; zhukovdm@yandex.ru
[2] The Institute of Information Technologies, MIREA—Russian Technological University, 78, Vernadskogo pr., 119454 Moscow, Russia; dtghmflysq@gmail.com or andrianova@mirea.ru
[3] The Institute of Innovative Technologies and Public Administration, MIREA—Russian Technological University, 78, Vernadskogo pr., 119454 Moscow, Russia
* Correspondence: ol-novikova@bk.ru

Abstract: One of the problems of forecasting events in news feeds, is the development of models which allow for work with semi structured information space of text documents. This article describes a model for forecasting events in news feeds, which is based on the use of stochastic dynamics of changes in the structure of non-stationary time series in news clusters (states of the information space) on the basis of use of diffusion approximation. Forecasting events in a news feed is based on their text description, vectorization, and finding the cosine value of the angle between the given vector and the centroids of various information space semantic clusters. Changes over time in the cosine value of such angles between the above vector and centroids can be represented as a point wandering on the [0, 1] segment. This segment contains a trap at the event occurrence threshold point, which the wandering point may eventually fall into. When creating the model, we have considered probability patterns of transitions between states in the information space. On the basis of this approach, we have derived a nonlinear second-order differential equation; formulated and solved the boundary value problem of forecasting news events, which allowed obtaining theoretical time dependence on the probability density function of the parameter distribution of non-stationary time series, which describe the information space evolution. The results of simulating the events instance probability dependence on time (with sets of parameter values of the developed model, which have been experimentally determined for already occurred events) show that the model is consistent and adequate (all the news events which have been used for the model verification occur with high values of probability (within the order of 80%), or if these are fictitious events, they can only occur over the course of inadmissible long time).

Keywords: forecasting events in news feeds; information space; news cluster; news clustering; stochastic dynamics of changes in information system states; news event threshold

check for updates

Citation: Zhukov, D.; Andrianova, E.; Trifonova, O. Stochastic Diffusion Model for Analysis of Dynamics and Forecasting Events in News Feeds. *Symmetry* **2021**, *13*, 257. https://doi.org/10.3390/sym13020257

Academic Editor: Evgeny Nikulchev
Received: 12 January 2021
Accepted: 1 February 2021
Published: 3 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A huge role in the processing of large amounts of data belongs to Data Mining technologies, considered as a set of methods for discovering previously unknown, unusual, practically useful, formalized knowledge necessary for making decisions in data. The concept of "Data Mining" can be used both in a narrow sense for a set of data from a limited subject area, and in general for the analysis of the information space, representing the totality of the results of all semantic activity. The main components of semantic activity are information resources, means of information interaction, and information infrastructure.

Extraction and discovery of knowledge from texts in natural languages is one of the most important areas of Data Mining. In particular, it can solve the problem of searching for hidden patterns that allow for making forecasts about possible news events in the future

and creating models of proactive impact on various social and economic processes (for example, the emergence of economic, political, and social crises).

In our opinion, the development of new mathematical models for predicting news events based on the analysis of texts in natural languages is a further development of Data Mining and Knowledge discovery technologies. Since it can allow not only to extract the already existing hidden patterns and knowledge, but also to predict the occurrence of various news events in the future on their basis. Analysis of texts and extracting data from them, for use in a forecasting model, can be based on already known methods of computational linguistics or Text mining methods (a set of machine learning and natural language processing methods in order to obtain structured information from a corpus of text documents, which can be considered, as one of the directions of Data Mining), and to create forecasts based on them, the proposed model can be used.

The work presented by us considers the model for discovering and predicting the occurrence of events in news feeds, based on an analysis of the information space as a whole. In this sense, the model we have created is one of the possible directions for the development of Data Mining technologies.

The purpose of the study described in this article was to develop a model for predicting emerging news events based on the analysis of events dynamics that have already happened. We have solved a number of tasks to achieve this goal, which are described later in the article. First, we put together a collection of news text messages over a long period of time (100,000 text documents for 2016). Then, we used computational linguistics methods to process them (lemmatization and vectorization based on a dictionary of terms; creation of a TF-IDF (TF—term frequency, IDF—inverse document frequency) matrix: term-document; clustering by thematic groups with event dating by time) and divided them into thematic clusters. The model developed by us, the vector representation of the text description of predicted events is used as input data, which allows us to find the values of the angle cosine between them and the centroids of thematic clusters obtained from a collection of news texts. The change in the value of a given cosine over time is considered as the wandering of a point on the segment [0, 1], which contains a trap at the threshold point of the event realization, where the wandering point can fall over time. As a trap, we consider the minimum value of the allowed cosine metric of similarity of vectors (similar to how it is done when determining the relevance of text search requests). The probability schemes of transitions between different states in the information space were considered when creating the model. The model parameters were determined based on the analysis of changes in the structure of existing news thematic clusters through time. The position of the cluster centroid vector and the number of messages on this topic during the day can be considered as a non-stationary time series. The appearance over time in news feeds of descriptions of events of a certain type related to a given topic (for example, references to terrorist acts, the activities of certain political leaders, etc.) can over time be considered as the formation of a discrete time series (the parameter is the frequency of mentions of this event during the day). The analysis of the dynamics characteristics of a given series can be used to predict its evolution, as well as to calculate the probability of events occurring during a given time interval.

We experimentally tested the model of forecasting news events presented in this article based on the description of their dynamics. We evaluated the accuracy and reliability of forecasts for the implementation of future events obtained on the basis of the developed model as well. As a basis for determining the parameters of the model, a collection 100,000 news text documents collected for 2016 was used, and as predicted news, a text description of events that occurred in 2017. The model created by us allows us to describe the change in the probability of the predicted event realization over time and to evaluate the possible time of its realization.

2. Review of Research on Forecasting Events Based on Text Analysis

The use of the analysis of news texts for the possible prediction of events is still a poorly studied research area, and there is not a very large number of works on this subject.

For example, the possibility of detecting and studying the features of the occurrence of interrelated sensational news events based on text analysis was studied using natural language processing (NLP) in this work [1]. This article examines the occurrence of pairs of realized events in the news space and attempts to identify patterns in which the event can occur after the first of them is detected.

We can also mention the work [2], in which an attempt was made to extract causal chains between events from their textual descriptions in order to detect previously unknown and hidden connections between events. A method based on linguistic patterns was used to extract the causal chain. In work [3] The authors developed a model for detecting causal relationships between events in social networks, which is used to predict their tonality and the time passed between them. In the first step of realization, the model in [3] tweets are selected for a certain period of time, then keywords are extracted from them. In the second step, the key words are used to determine the tonality of the extracted words—positive, negative, or neutral. To determine the tonality of words, a classifier trained on the basis of the support vector machine (SVM) is used. In the third stage, the cause-and-effect relationships between keywords are determined, for this purpose, the association rule learning method is used, which extracts the "if-then" rules from the data. In the fourth step is the prediction of events using temporal analysis of tweets and calculated cause-and-effect relationships. The author presented a study on the use of tweets to spatial and temporal markers to predict crime in work [4]. The author uses linguistic analysis and statistical modeling for tweets to automatically identify topics discussed in major cities. The author [4] proposes to use a topic modeling to highlight the topics of tweets. Before thematic modeling, the text of tweets was tokenized using a special tokenizer and a partial speech tagger. In the proposed model, a special tokenizer recognizes emoticons as separate tokens. To model the topic of tweets, semantic content describing the user's emotional state is also analyzed.

The author [4] determined a one-month training window to predict the occurrence of a crime, then placed the marked points (latitude/longitude pairs) along the city boundaries. The points were taken from two sources for training of the binary classifier: from known crime scenes marked in the training window, and from a grid of evenly spaced points with an interval of 200 m that did not coincide with the points from the first set.

The problem of the impact of news headlines on the behavior of investors and the movement of financial markets was considered in work [5]. The model is based on weighted association rules, which are used to determine whether a news release is important enough for investors. During training on real data, the weighted association rules algorithm detects terms or keywords that simultaneously appear frequently in the news headlines. The keyword or term (p) appears in the news headline (s) on the j-th day, and n represents the total number of days in which the keyword (p) appears in the news headlines. The weight (wks) for an individual keyword or term (p) is determined as: $wks_p = \frac{\sum_j pf_{p,j}}{n}$, where $pf_{p,j}$ means the fluctuation of the closing price of the stock on the next trading day. These weights help you decide whether the keywords in the news headlines affect the trading result.

It was studied that the state of public mood, measured by time series OpinionFinder and GPOMS, it can predict changes in the closing index values DJIA (Dow Jones Industrial Average in study [6] analyzed the text content of daily Twitter feeds using two sentiment measurement tools: OpinionFinder, which measures positive and negative moods and a Google profile of mood states (GPOMS), which measures mood in terms of six dimensions (calm, anxious, confident, cheerful, kind, and happy). Using Granger causality analysis and a self-organizing fuzzy neural network, the resulting mood time series were cross-checked for their ability to detect public reaction to the 2008 Presidential election and Thanksgiving Day.

The results of the experiments show that the accuracy of predictions of the DJIA closing index values can be improved by including specific measurements of public sentiment.

An overview of works in the field of intellectual text analysis for securities market forecast is presented in [7].

Currently, most of the works on predicting the dynamics of various processes based on the analysis of text data are focused around user behavior in social networks, forums, and chats. For example, Gruhl et al. in their work [8] showed how online communication can predict book sales. Mishne and Rijke [9] used sentiment analysis in blog posts to predict movie sales. The article by Liu et al. [10] describes the application of the PLSA model of probabilistic hidden semantic analysis to assessing sentiment in blog posts to predict future sales. In [11], the authors showed that Google search queries are able to predict the epidemiological development of infectious diseases, as well as consumer preferences and costs. L. Zhao et al. [12] showed how space-time tweets can be used to predict crime. In this work, linguistic analysis of texts was applied to tweets, topic modeling (to highlight topics of tweets) to automatically determine topics that are then used in a crime forecasting model.

Despite the evidence that open-source data, including news, are surrogates for predicting various events (disease outbreaks [13], election results [14,15], and protests [16]), there are far less studies examining the possibility of predicting the occurrence of news events in the information space.

The authors of [17] have developed a method that solves the problem of identifying precursors and predicting future events. According to data from the collection of streaming news (news taken from several open sources of three Latin American countries), a nested approach was developed to predict significant public events and protests. The capability of consistent identification of news articles and harbingers of protests has been demonstrated. The strengths of the approach proposed in [17] are demonstrated by the empirical assessment, which consists in filtering potential precursors in accurately predicting the characteristics of civil unrest events and in predicting the occurrence of events with an advantage in execution time.

Paper [18] shows the model for fatal accidents and natural disasters forecasting. Its authors suggest analyzing historical data, and extracting event patterns related to the disasters, to use the resulting patterns in machine learning as training samples in order to predict upcoming disasters based on the current events.

To predict a new event with a given topic, it is necessary to create a model of its formation based on the description of its time series; then, to find the function of density of its parameters distribution probability.

When making forecasts, the main problem of analyzing and modeling the behavior of a time series of news feed events is that at any moment in time there is only one implementation of the process (one statistical sampling, one sample of a time series that has already been implemented), which you need to use to create a forecast for the following points in time.

Regardless of the tools used (statistical models, neural network models, fuzzy logic models, etc.), in the existing analysis methods, a non-stationary time series is divided into separate areas where it is quasi-stationary with its own selective (for each section of the time series) distribution function, and there is a part of the series, in which a transition process (disorder) takes place, between each of the areas. Duration of the transient process is determined both by the factors characterizing the change in regime (the actual disorder) and the sampling size used for statistical analysis [19]. The parameters of the sample distribution function can be established based on the analysis of data observed over the time interval of quasi-stationarity. In particular, nonparametric methods can be used to reconstruct the probability density from the observed values [20]. In practice, two tasks need to be solved. The first is to determine the time interval of quasi-stationarity. The second is in determining the onset of disorder during the transition period, and with a minimum delay.

The stationary time series is represented as the sum of the deterministic component (trend or periodic) and the remainder, the autocorrelation function, which is close to zero with sufficient accuracy and indicates the proximity of the remainder to the "white noise". After that, the problem is posed of finding the closest statistic (distribution function) that simulates behavior of the remainder.

When studying stationary random processes, according to Glivenko's theorem (on the convergence of empirical probability to a theoretical distribution) [21], the more observed values are taken into account, the more accurately theoretical characteristics of the distribution of a random variable from a certain interval can be obtained. For non-stationary random processes, this condition, due to their specificity, cannot be met, which makes it difficult to use the results of their analysis for further forecasting.

For non-stationary time series, indicators of its particular properties have their own specific form, which cannot be generalized to the series of another type. For example, a linear trend indicator is not particularly effective for series with quasi-periodic change, like an indicator of nonstationarity of variance for series with a quasi-linear trend. Moreover, indicators based on some average characteristics of series (for example, the first few moments) do not form a basic system by which it is possible to determine the tendency of change in a random process that is local in time.

The identification of the state of a non-stationary random process can be formulated as the problem of recognizing the selective distribution function (SDF) as belonging to a certain general population. However, if the distribution function is non-stationary, then training the recognition algorithm on past data often turns out to be inadequate. There is only one trajectory, which, due to nonstationarity, does not allow using a large sampling size for testing particular indicators of the local behavior of a time series.

Thus, it can be said that analysis of quasi-stationary sections of the observed time series and construction of selective distribution functions may be ineffective for predicting subsequent evolution.

Currently, diffusion equations, including nonlinear diffusion [22], Liouville equation [23], Fokker-Planck equation [23], and a number of others are most often used as approximations of distributions in practical models for analyzing and predicting the evolution of nonstationary time series.

The use of existing methods of time series analysis for modeling the dynamics of news feed events can lead to significant errors, this is due to the large variability of their characteristics, as well as nonlinearity and nonstationarity. Therefore, it is necessary to search for new methods for analyzing their dynamics and approximating distribution functions.

Some works describe a number of probabilistic-theoretical approaches to forecasting news events. For example, in [24], the authors describe a model for predicting future events by generalizing specific sets of sequences of events extracted from news over a period of 22 years: from 1986 to 2008. The authors are trying to build a model that takes into account the relationship between past historical events and predicts future events. The authors assume that events in the real world are generated by a probabilistic model, which also generates news messages about these events. The messages from news events are used to build a model in the form of determining the probability $P(ev_j(t+\Delta)|ev_i(t))$ of realization of some future event ev_j at time t+Δ and the event ev_i. that has passed at time t. This probability approximates the relationship between two real-world events that have occurred. The model shows that with a probability of 18%, the drought event ev_j occurs after the flood event ev_i.

The use of text data and machine learning methods for predicting fatal accidents and natural disasters is described in [18]. The authors collected text messages about disasters from the Google search engine by keywords. Then, the text documents obtained as a result of requests were processed by the methods of mathematical linguistics, and using a trained Bayesian classifier any false results were sifted out. After data collecting, semantic clustering of the data collected was carried out. A matrix of transitions was built from the keywords for which search queries were generated. An observation matrix was constructed

from the grouped events. Then, both matrices were fed to the input of the latent Markov model for forecasting. According to the authors, this approach allows predicting future events and locations of events happening.

In [25], to solve the problem of forecasting news events, the authors study time dependences in the streams of events and introduce piecewise-and-constant approximation of their intensity, having the Bayesian approach and Poisson distribution used for description of the future events importance sampling. This allows non-linear time dependencies to be built to predict future events using decision trees.

Over time, appearance of descriptions of a certain type events related to a given topic (for example, mention of terrorist acts, the activities of certain political leaders, etc.) in news feeds can be considered as the formation of a discrete time series (which parameter is the frequency of mentioning this event within 24 h). Analysis of dynamic characteristics of this series can be used to predict its evolution, and to determine the absence or presence of long-term dependencies in its behavior, as well as to calculate the probability of the events occurrence within specified time interval.

It should be noted that in order to form the time series of the events appearance in the news feed, it is necessary to solve an important auxiliary problem: it is necessary to select from the entire set of news feeds text messages (hundreds of thousands and millions) those particular and specifically related to the given topic (clustering of events by semantic groups) with high accuracy. Ensuring high accuracy of clustering guarantees that a significant part of information will not be lost during the formation of the time series, for example, according to the frequency of this event appearance, which will make it possible to achieve a more accurate determination of parameters of the considered time series and will not affect the forecast of its evolution.

The literature review of the works with topics close to our research shows that the development of various models for predicting news events is a very relevant topic and their further realization is required, which can significantly expand the capabilities of Data Mining.

In our view, one of the promising directions for creating models for predicting news events based on the analysis of text information is the use of probability-theoretic approaches based on the construction of approximating distribution functions that take into account the possibility of self-organization for events described by news feeds. In this case, the predicted event can be constructed from those that have already occurred using the obtained theoretical approximating distribution functions. The results obtained can be used for analytical and predictive purposes, for example, to determine the probability of an increase in terrorist activity in the future.

In our opinion, to create a probability-theoretic forecasting model, it is necessary to highlight the following main properties of news feed events.

1. There is an accident in the nature, time and place of the news event implementation. A realizable event is a manifestation of stochastic processes with an initially unknown law of probability distribution, and its statistical characteristics (mathematical expectation, dispersion, etc.). However, at the same time, it should be noted that there are possible causal relationships between different events, which creates prerequisites for predicting some events based on the realization of others.

2. As the analysis of the observed time series of news feed dynamics shows, they are nonstationary, and the processes taking place have the possibility to self-organize.

The basic idea of our model is that the predicted event can be described in the information space by a text document that can be attributed to a certain semantic group (cluster) that has its own characteristics. At any given time, there are many different information clusters that describe various ongoing processes (natural, social, economic, cultural, political, sports, military, and other news events), and display the main properties of the real world and the interrelationships of events. The described image of the predicted news event can be dynamically formed from a set of images of already realized events and clusters.

3. Materials and Methods

The results of this research were obtained using systems modeling, system theory and system analysis, methods of mathematical analysis, probability theory, differential calculus, operational calculus, methods of computer linguistics, theories of classification, and systematization.

The appearance in news feeds of descriptions of events of a certain type over time related to a given topic (for example, references to terrorist acts, the activities of certain political leaders, etc.) can be considered as the formation of a discrete time series (the parameter of which is the frequency of mentions of this event during the day). The analysis of the dynamics characteristics of a given series can be used to predict its evolution, as well as to calculate the probability of events occurring during a given time interval.

The core of the approach proposed in this article, which may be used to create a model for constructing a forecast of a future event from those that have already occurred using theoretical approximating distribution functions, is as follows:

(1) Let us take the collection (corpus) of **N** text documents describing news feed events for a certain period of time with references to the dates of their occurrence. Then, using lexical and semantic methods of computational linguistics (removal of punctuation marks, stop words, bringing words to normal forms, lemmatization, creating a glossary of terms, etc.) [26–29], by means of the glossary of terms (words, n-grams, or objects of associative-semantic classes) of **M** size, let us create a vector representation of a set of texts in information space with (which dimension will be $\mathbf{R^M}$). To improve the accuracy of text analysis and further clustering by semantic groups, you can use the approaches based on combining words that have a similar meaning in texts into associative-semantic classes, for example, using the word2vec algorithm.

Each document in this set is assigned to vector $X_i = \left\{x_{1,i}, x_{2,i}, \cdots, x_{k,i}, \cdots, x_{M,i}\right\}$, where i takes values from 1 to **N**, and each element of the vector $x_{k,i}$ describes the TF-IDF normalized frequency of the k-term (words, n-grams, or objects of associative-semantic classes) occurrence from the glossary into the i-document of the collection: $\mathrm{TFIDF} = \mathrm{TF} * \mathrm{IDF} = \frac{\mathrm{n_k}}{\sum_{\mathrm{k}} \mathrm{n_k}} * \log \frac{\mathrm{D}}{\mathrm{d}}$, where n_k is the number of occurrences of the k-term in a document; $\sum_k n_k$ is the total number of terms in the document; D is the total number of documents in the collection; and d is the number of documents where this term is found. Using TF-IDF reduces the weight of commonly used terms, which are logically relevant, and finally increases the text clustering accuracy. Vectors X_i form a matrix of **N** by **M** dimension: term—document:

$$\begin{vmatrix} x_{1,1} & x_{1,2} & \cdot & x_{1,i} & \cdot & x_{1,N} \\ x_{2,1} & x_{2,2} & \cdot & x_{2,i} & \cdot & x_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{j,1} & x_{i,2} & \cdot & x_{k,i} & \cdot & x_{i,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{M,1} & x_{M,2} & \cdot & x_{M,i} & \cdot & x_{M,N} \end{vmatrix}$$

(2) Next, using standard methods [26–29] and algorithms, we will cluster text documents (division into semantic groups) using their vector representation.

To perform clustering, we can use K-Means algorithm, which belongs to the class of hierarchical, distinct, integrative algorithms, has a great number of advantages (simplicity of implementation, high clustering quality, and execution speed), and is most widely used for these purposes. A certain disadvantage of this algorithm is the necessity to determine the number of clusters in advance. However, it has a high operating speed and compilation $O(j \cdot C \cdot D \cdot \sum_k n_k)$, where C is the number of clusters, $\sum_k n_k$ is the number of elements values in a vector, D is the number of documents, j is number of iterations. The goal of this algorithm is finding such cluster centers so that the distance between the cluster document vector and the cluster center (centroid) vector shall be minimal: $\underset{\mathrm{N_p}}{\operatorname{argmin}} \sum_{\mathrm{p}=1}^{\mathrm{k}} \sum_{\mathrm{y}_i \in \mathrm{N_p}} \mathrm{d}\left(\mathrm{y}_i, \mu_{\mathrm{p}}\right) = \underset{\mathrm{N_p}}{\operatorname{argmin}} \sum_{\mathrm{p}=1}^{\mathrm{k}} \sum_{\mathrm{y}_i \in \mathrm{N_p}} \parallel \mathrm{y}_i - \mu_{\mathrm{p}} \parallel^2$, where y_i is a document from the cluster; μ_p is the centroid of cluster p; $\mathrm{N_p}$ is the set of documents of cluster p. The

centroid, i.e., the arithmetic mean vector of all vectors in a cluster (or a subgroup thereof), can be calculated as follows: $\mu_p = \frac{\sum N_p}{D_p}$, where N_p is vector of a news item from cluster p; D_p is the number of news texts in the cluster. As the distance between vectors, we use the cosine metric (the cosine of angle between vectors): $d(y_i, z_i) = \frac{\sum_{i=1}^{n}(y_i * z_i)}{\sqrt{\sum_{i=1}^{n} y_i^2} * \sqrt{\sum_{i=1}^{n} z_i^2}}$, where y_i is the coordinate value of the first vector; z_i is the coordinate value of the second vector (the larger the cosine of the angle between vectors, the higher the similarity of documents). Given that all the elements of vectors are positive numbers: $0 \leq d(y_i, z_i) \leq 1$.

In addition to the K-Means algorithm, other well-proven clustering algorithms may be used. For example, DbScan, Affinity Propagation, Agglomerative Clustering, BIRCH, which have high clustering accuracy.

Due to the fact that news events can appear and disappear over time, the structure of news clusters and the position of vectors defining their centers (centroids) will vary. Therefore, we can create a time series describing a certain type of events in news feeds [30]. As an example, parameters of such series may be the frequencies of messages of this type of event appearance in a news feed or position of centroids of clusters, which include the texts describing these events.

To generate the time series describing the events of a certain type in the news feeds and to perform research, a collection of 100,000 of text documents for 2016 ("Vedomosti", "Kommersant", "RBC", "News. First Channel") were collected from four Russian news sites. The maximum number of words in one document was 10,404, the minimum 101. The vocabulary of terms used to create the "term-document" matrix included 2,570,724 words and 1,451,828 terms.

According to the results of the study conducted and comparison of clustering algorithms, the non-hierarchical clustering algorithms K-Means and Affinity Propagation showed the best quality and runtime. The K-Means algorithm selects clusters with more general topics, while the Affinity Propagation algorithm selects clusters with subtopics (one common topic is split into several clusters with narrower topics). As models of text presentation, the best results were shown by the models "document—term" with TF-IDF without using n-grams and "document—associative-semantic group".

As for the result of clustering, 300 clusters on various topics were obtained from the existing news corpus. Further, each of them was segmented into 365 subgroups of news per day (24 h) without summation for previous periods.

(3) We create a textual description of a news event text image for which we shall determine the probability of its occurrence over time (forecast). Then, we vectorize the text description of the predicted event (vector $\mathbf{X_{bs}}$). Then, we determine the values of cosines of angles between the centroid vectors and the predicted event vector for some point of time t. Then, we calculate their mean value. The mean value of the cosines at this point of time t will be the point on the numerical segment [0, 1], and in view of the change over time of the cluster's structure, this point will perform movements (wandering) on the segment. Eventually this point may reach the given cosine value, which will be considered as the threshold of the event occurrence (let us call it l). We refer to the current value of the mean value of cosines as the information system state at a time (denote it as x_0). Probability of reaching the event threshold l will depend on the time t (i.e., in fact, we consider virtually random wandering of the point on segment [0, 1], which contains a trap in l, where the wandering point can eventually fall.

The described approach allows us to derive a nonlinear differential equation of the second order, based on consideration of the schemes of probabilistic transitions between states. It also allows us to formulate and solve, regarding the prediction of news events, the boundary value problem of the dependence of the probability of reaching the predicted event on time and to consider its solution (to obtain a theoretical approximating distribution function).

4. Results

4.1. Deriving the Distribution Function for the Time Series Parameters, Which Describe Dynamics of the News Feed Content

4.1.1. Plotting of Difference Schemes of Probabilities of State Transitions in Information Space. Deriving the Main Equation of the Model

As a measure of meaning similarity between two text documents in computational linguistics, they frequently use a cosine metric. Proximity of the cosine to the figure of one indicates the grade of similarity between the texts meanings, and that of zero indicates their difference. Besides, the cosine value itself in this case will always be on the segment from 0 to 1 ([0, 1]). Let us denote the current value of the mean cosine of the angle between the forecast textual description vector and the centroids of text clusters from which this event may be supposedly formed, as x_i (the information system state).

Suppose that the time interval of the states changing process has the value τ (infinitely small). Suppose that for the time interval τ, the state of the system may be increased by a certain value ε (an increasing trend) or decreased by value ξ, (a decreasing trend). Let us denote the entire set of states on the forecasting axis as X. The state observed at time point t may be denoted as x_i ($x_i \in$ X). Ultimately, the system state x_i may be near the predicted event threshold equal to 1 (or set as the implementation of another cosine value of the angle between the cluster centroid and the vector of the predicted event text).

Let us record the value of the current time as $t = h\tau$, where h is the step number of transitions between states (the process of transition between states becomes quasi-continuous with an infinitely small-time interval τ), h = 0, 1, . . . , N. The current state x_i at step h, after transition at step $(h+1)$ may be increased by some value ε or decreased by value ξ, and, accordingly, be equal to $x_i - \varepsilon$ or $x_i + \xi$.

Les us introduce the concept of probability of finding the information space in some state. Suppose that, after a certain number of steps h, we can say about the system described that:

- $P(x - \varepsilon,\ h)$ is the probability that the system is in state $(x - \varepsilon)$;
- $P(x,\ h)$ is the probability that it is in state x;
- $P(x + \xi,\ h)$ is the probability that it is in state $(x + \xi)$.

After each step, state x_i (the index i for brevity can be omitted below), can change by value ε or ξ.

The probability $P(x, h+1)$—that, at the next step $(h+1)$, the system will be in found in state x—will be determined by several transitions (see Figure 1):

$$P(x, h+1) = P(x - \varepsilon, h) + P(x + \xi, h) - P(x, h). \quad (1)$$

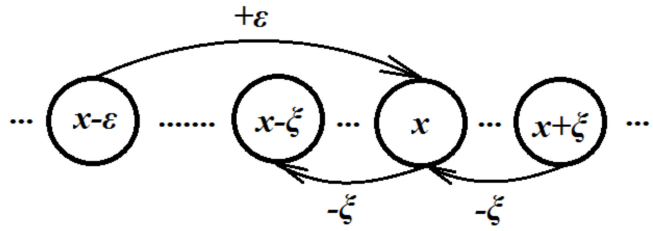

Figure 1. Scheme of possible transitions between states of the system at step $h + 1$.

Let us explain the expression (1) and the scheme shown in Figure 1. Probability of transition to state x at step $hP(x, h+1)$ is determined by the sum of the probabilities of transitions to this state from states $(x - \varepsilon)$: $P(x - \varepsilon, h)$ and $(x + \xi)$: $P(x + \xi, h)$, where the system was at step h, without considering the probability of transition $(P(x, h))$ of the system from state x (where it was at step h) to any other state at step $(h+1)$. In this case, we assume that the transitions themselves occur with probability equal to 1.

Let us expand the terms of Equation (1) in the Taylor series and, taking into account derivatives of no higher than second order for x and the first derivative for time t, we obtain the following differential equation to change the probability of finding an information process in a certain state x depending on the value of time:

$$\tau \frac{\partial P(x,t)}{\partial t} = \frac{1}{2}\left\{\varepsilon^2 + \xi^2\right\}\frac{\partial^2 P(x,t)}{\partial x^2} - [\varepsilon - \xi]\frac{\partial P(x,t)}{\partial x}. \quad (2)$$

Having differentiated Equation (2) for x, let us proceed to the dependence of the probability density of finding the information process in a certain state x depending on the value of time t:

$$\frac{\partial \rho(x,t)}{\partial t} = \frac{\varepsilon^2 + \xi^2}{2\tau}\cdot\frac{\partial^2 \rho(x,t)}{\partial x^2} - \frac{\varepsilon - \xi}{\tau}\cdot\frac{\partial \rho(x,t)}{\partial x}. \quad (3)$$

This equation may be considered as the equation of a simple diffusion model. In addition, Equation (3) takes into account the ordered movement (drift) between the states of the information process. Equation term $\frac{\partial \rho(x,t)}{\partial t}$ reflects the system change rate over time; equation term $\frac{\partial^2 \rho(x,t)}{\partial x^2}$ accounts for random transitions (diffusion wandering of the system state); equation term $\frac{\partial \rho(x,t)}{\partial x}$ describes ordered transitions (trend or drift), e.g., when the state value either increases ($\varepsilon > \xi$) or decreases ($\varepsilon < \xi$).

In terms of the model application scope, Equation (3) shall take into account the limitation imposed on factor $(\varepsilon^2 + \xi^2)/2\tau$ before the second derivative for x, which accounts for the probability of an accidental state change. The condition $(\varepsilon^2 + \xi^2) < (l - x_0)^2$ must be met, which is all about the transition from initial state x_0 across the event reaching threshold (l), which cannot occur faster than in one step τ. If $(\varepsilon^2 + \xi^2) \geq (l - x_0)^2$, the system crosses the event reaching threshold in one step.

The above approach to building a model to analyze the formation of events in news feeds was generally considered in papers [31–35].

4.1.2. Formulating and Solving a Boundary Value Problem When Predicting News Events in the Information Space for Systems with Memory Implementation and Self-Organization

Considering function $P(x,t)$ to be continuous, we can move from probability $P(x,t)$ (Equation (3) to probability density $\rho(x,t) = \partial P(x,t)/\partial\, x$ and formulate a boundary value problem, the solution of which we will describe the process of transition between states in the information space.

The first boundary condition: Let us choose the first boundary condition for state $x = 0$. The probability of finding such a condition over time may differ from 0; however, the probability density, describing the stream in state $x = 0$, shall be taken as equal to 0 (system states cannot fall in the area of negative values (the reflection condition is implemented; the value of angle cosine in this case cannot be negative according to the definition of the cosine metric for text vectors)), see Equation (4) i.e.,:

$$\rho(x,t)_{x=0} = 0. \quad (4)$$

The second boundary condition: Let us restrict the area of possible states of the information system to value L (the cosine metric cannot be greater than 1) and we choose the second boundary condition for the state $x = L = 1$. The probability of finding such a state over time will differ from 0. However, the probability density, which describes the stream in state $x = L = 1$, shall be taken as equal to 0 (the system states cannot fall into the range of values greater than the maximum possible value (the condition of reflection from the boundary is implemented)), see Equation (5) i.e.,:

$$\rho(x,t)_{x=L} = 0. \quad (5)$$

Since at time point $t = 0$ the system state may already be equal to some value x_0, then the initial condition can be set as follows see Equation (6) i.e.,:

$$\rho(x, t = 0) = \delta(x - x_0) = \begin{cases} \int \delta(x - 0)dx = 1, x = x_0 \\ 0, \qquad\qquad x \neq x_0 \end{cases}. \tag{6}$$

Using operational calculus methods for probability density $\rho_1(x, t)$ and $\rho_2(x, t)$ of finding the system state in one of the values on the segment from 0 to L, we can obtain the following system of Equations (7) and (8):

With $x \geq x_0$:

$$\rho_1(x, t) = -\frac{2}{L} e^{a_1 \frac{(x - x_0)}{2a}} e^{-\frac{a_1^2 t}{4a}} \sum_{n=1}^{M} \frac{sin\left(\pi\, n \frac{x_0}{L}\right) \sin\left(\pi\, n \frac{L-x}{L}\right)}{\cos(\pi\, n)} e^{-\frac{a \cdot \pi^2 n^2}{L^2} t}. \tag{7}$$

With $x < x_0$:

$$\rho_2(x, t) = -\frac{2}{L} e^{a_1 \frac{(x - x_0)}{2a}} e^{-\frac{a_1^2 t}{4a}} \sum_{n=1}^{M} \frac{sin\left(\pi\, n \frac{L-x_0}{L}\right) sin\left(\pi\, n \frac{x}{L}\right)}{\cos(\pi\, n)} e^{-\frac{a \cdot \pi^2 n^2}{L^2} t}. \tag{8}$$

where $a = \frac{\varepsilon^2 + \xi^2}{2\tau}$ and $a_1 = \frac{\varepsilon - \xi}{\tau}$.

If the predicted event occurrence is associated with an increase in the value of the system initial state x_0, then the integral $P(l, t)$ shown in Equation (9):

$$P(l, t) = \int_0^{x_0} \rho_2(x, t)dx + \int_{x_0}^{l} \rho_1(x, t)dx, \tag{9}$$

will set the probability that the system state at time point t is on the segment from 0 to l; i.e., the event threshold l will not be reached. As the occurrence threshold, we can use the set mean cosine value of the angle between the cluster centroids and predicted event text vector.

We consider the process of walking a point on the segment [0, 1] from the condition x_0. This is why, in Equation (9), the first integral is calculated from the lower limit of 0 to the upper limit x_0 by using $\rho_2(x, t)$, and the second integral of x_0 to l by using $\rho_1(x, t)$. So, the Equation (9) determines the time dependence of the probability of the "survival" of the wandering point (that it will not fall into the trap).

Accordingly, the probability that the event threshold l will be reached or surpassed by time point t may be defined as Equation (10):

$$Q(l, t) = 1 - P(l, t). \tag{10}$$

Our analysis shows that $\rho_1(x, t)$ and $\rho_2(x, t)$ with any values of t and x are not negative; for function $Q(l, t)$ with $t \to \infty$, the condition of $Q(l, t) \to 1$ ($P(l, t) \to 0$) is met.

4.2. Experimental Testing of the Suggested Model for Forecasting News Feed Events

4.2.1. Definition of the Parameters of the Event Forecasting Model Based on Changes in the Cluster Structure in the Information Space of News Feeds

To simulate a topical event in a news feed using the developed model, we need to determine its parameters (ξ and ε). The model for forecasting information events in news feeds by solving a boundary value problem for systems with memory implementation and self-organization is based on the use of parameters, which take into account the possible decrease in the current value of the system state (decreasing trend ξ) and its increase (increasing trend ε). These parameters are associated with the change dynamics in the news cluster structure and can be determined on its basis.

Experimental testing of the model is based on the fact that we can take a time series that has already been implemented for a certain time interval and describes the dynamics

of some type of event in the news feed. Next, by analyzing the initial part of this series, we shall determine the values of parameters ξ, and ε. As a predicted event, we can take the textual description of a news event in the subsequent part of the time series and calculate its dependence on the occurrence probability time. Then, when shall compare the data obtained with the observed implementation time.

(1) From news collected for 2016, we create vectors and divide them into **W** topical clusters (in our case, **W** = 300, i.e., we have 299 topical cluster + one cluster containing all texts that were not included in 299 thematic clusters). Next, each of **W** clusters is divided into 365 subgroups of text vectors by days of news publication. If there were no thematic news on a given day, the day subgroup of this cluster will contain an empty set of vectors. Thus, in each of cluster, news feed events for 2016 form time series that determine the model's parameters.

(2) To test the model and to determine its parameters, we shall use the text description of topical event that occurred on any of days in 2017. Taking a known news message from 2017 (a published text with the date of the event occurrence described in the news), we create its vector N_i.

(3) For each day in 2016, within each day subgroup of vectors in each cluster, we determine the coordinates of centroids: $C_j(t) = \left\{c(t)_{1,j}, c(t)_{2,j}, \cdots c(t)_{k,j}, \cdots c(t)_{M,j}\right\}$, where $c(t)_{k,j}$ is the arithmetic mean of the coordinates of vectors in the subgroup (for the given day, without accumulation for previous periods) at time point t, whereas j takes values from 1 to **W** (i.e., we get **W** centroids for each day). If there were no topical news on the given day, the day group of this cluster will contain an empty set of vectors, and the centroid forms an empty set.

(4) For each time t =1, 2, 3, . . . , 365 (of each day), within each of clusters **W**, we find the cosine values of the angles between the day centroid vectors $C_j(t)$ and the vector of news N_i of the textual description of the predicted event (these cosines are denoted as $S_j(t) = cos\{C_j(t); Ni\}$). If there are no news on the given day, the cosine metric will be equal to the empty set.

(5) We choose the cosine metric values and the corresponding days of year, which are different from an empty set. We take the first ($S_j(t_1)$) and second ($S_j(t_2)$) values and find the difference between the second and first ones ($\Delta S_j(t_2 - t_1) = S_j(t_2) - S_j(t_1)$); then, we divide it by the time interval ($t_2 - t_1$) in days between the first and second cosine metric values, which are different from the empty set. Thus, we find deviation ($\Delta_j(t_2 - t_1) = \frac{S_j(t_2) - S_j(t_1)}{t_2 - t_1}$ reduced to one day (τ = 1), which may be either positive or negative. Then, we take the third ($S_j(t_3)$) non-zero cosine metric value; subtract the second value ($S_j(t_2)$) from it, and divide the resulting difference ($\Delta S_j(t_3 - t_2) = S_j(t_3) - S_j(t_2)$),) by the time internal ($t_3 - t_2$) in days between the third and second cosine metric values. Thus, we find the deviation ($\Delta_j(t_3 - t_2) = \frac{S_j(t_3) - S_j(t_2)}{t_3 - t_2}$ reduced to one day (τ = 1), it may be either positive or negative). We repeat this procedure for all non-zero cosine metric values until the end of the year.

(6) We sort all deviations into two groups: $\Delta_j(\Delta t) < 0$ and $\Delta_j(\Delta t) > 0$ and find the mean values for each of them (the sum $\Delta_j(\Delta t)$ divided by their number). Let us adopt the mean value for the deviation cosine for group $\overline{\Delta_j(\Delta t)} < 0$ as the decreasing trend value ξ, and for group $\overline{\Delta_j(\Delta t)} > 0$, as the increasing trend value ε.

(7) The last mean value of the cosine metric at the end of the year (without taking into account the number of empty sets) is adopted as the system initial state x_0.

4.2.2. Evaluation of the Value of Cosine Measure of the Event Occurrence Threshold in the Information Space of News Feeds

To evaluate the value of cosine measure of the event occurrence threshold, let us consider a text example where two documents, S_1 and S_2, have very close semantic meanings:

S_1= "to buy a bookcase at a discount";

S_2= "to buy a bookcase cheap with free delivery".

Let us draw a table of normalized (lemmatized) words of these sentences (see Table 1):

Table 1. Normalized words of the sentences from the test example.

	Cheap	Buy	Book	Case	Free	Delivery	Discount
S_1	0	1	1	1	0	0	1
S_2	1	1	1	1	1	1	0

Calculating the cosine metric gives us a value of 0.61. In this case, we have considered short texts having a great semantic similarity. With increasing the text length, the cosine metric value will decrease significantly, although their semantic meaning will remain very close; therefore, we can assume that l = 0.5.

4.2.3. Modelling of the Predicted Event Occurrence Probability Dependence on Time. Analysis of Modelling Results

To test the model, five news (see Table 2), which describe events occurring in 2017, were randomly selected as a predicted event. Then, using the algorithm described in "4.3.1 Definition of the Parameters of the Event Forecasting Model Based on Changes in the Cluster Structure in the Information Space of News Feeds" and the created text clusters (W = 300), we have determined the values of the model parameters ξ, ε, and x_0 for each predicted news feed event (when finding ξ and ε, we used τ = 1 day), see Table 2.

The experimentally calculated model parameters, presented in Table 2, show that $\varepsilon = \xi$ in all cases considered. This leads to the fact that both $a_1 = \frac{\varepsilon - \xi}{\tau} = 0$ and Equations (7) and (8) are converted into Equations (11) and (12):

$$\rho_1(x,t) = -\frac{2}{L}\sum_{n=1}^{M} \frac{sin\left(\pi\, n\frac{x_0}{L}\right)\sin\left(\pi\, n\frac{L-x}{L}\right)}{\cos(\pi\, n)} e^{-\frac{a\cdot\pi^2 n^2}{L^2}t}, \text{ with } x \geq x_0, \tag{11}$$

$$\rho_2(x,t) = -\frac{2}{L}\sum_{n=1}^{M} \frac{sin\left(\pi\, n\frac{L-x_0}{L}\right)sin\left(\pi\, n\frac{x}{L}\right)}{\cos(\pi\, n)} e^{-\frac{a\cdot\pi^2 n^2}{L^2}t}, \text{ with } x < x_0. \tag{12}$$

If we compare semantic content of news texts No.1 and No.3 in Table 2, then one can note that they are highly close (both news describe criminal events). It is very important to note that the experimentally calculated values of the model parameters ξ, ε, and x_0 turn the same, which impliedly confirms correctness of the model used.

The results of simulating the time dependence probability of forecast implementation for the events described in Table 2 using Equations (9)–(12) and the set of model parameters determined using the set of 300 clusters (see also Table 2) are shown in Figure 2 (the number of the curve corresponds to the number of event in the Table).

The large black dots in Figure 2 corresponds to the time points of the actual occurrence of events. The results obtained show that the developed forecasting model of news feed events is adequate and consistent (depending on what memory depth is taken into account, all the described news events have occurred at high probability values (about 0.8), see Figure 2.

It seems interesting to test the developed model for the ability to predict fictitious news (something that cannot actually happen). As an example, we can take a small excerpt from a Russian folk tale about Roly-Poly Bun:

"Once upon a time there lived an old man and old woman. One day the old man says to his old woman: Hey you old woman, go and scrape our box, sweep our cornbin. Would you scrape some flour to bake me a bun?

Table 2. Some normalized news events in 2017 and model parameters for their forecasting.

No.	Normalized News Text	Date of Event	Value of Parameter ε	Value of Parameter ξ	Initial State of System x_0 31 December 2016
1.	{"id":"9dc7c737-0359-418f-a809-28a4aa23b3bb","date":1490774096000,"title":"The head of the Ministry of Internal Affairs was killed after he identified theft for 10 billion","content":"couple a week attempt on the life of Nikolai Volk write a statement dismissal own desire to refuse to sign inventory internal financial report information life killed the day before head of the Ministry of Internal Affairs of the Ministry of Internal Affairs Nikolay Volkov complained native department to steal an asset billion ruble force to sign blank document testimony native witness to check investigator IC Russia direct killer to look for authorised operative central directorate Criminal Investigation Department Ministry of Internal Affairs of Russia source editorial office report wolf to identify multi-billion dollar embezzlement assign a lot of internal check number of inventory suspicion to be confirmed establish an investigation person demand a high-ranking police officer signing an act verification of the Ministry of Internal Affairs RID allegedly no financial hole theft of the Ministry of Internal Affairs RID to be able to pay in time contractor owes many organizations previously the Ministry of Internal Affairs initiate a case the fact of fraud against the organization mariotrek responsible construction sanatorium ministry Olympics Sochi FSUE RID Ministry of Internal Affairs speak Customer service we are talking about fraud million rubles identification of the fact of involvement of the employee of the Ministry of Internal Affairs RID fraud the case is transferred to the Investigative Committee is known at the moment the Ministry of Internal Affairs should remain a Sochi builder at least one million rubles of the Ministry of Internal Affairs RID to be the defendant arbitration case lawsuit lawsuit Stroy Universal LLC debt million rubles Organization LLC Enterprise RTSPP RID owes a million to the Ministry of Internal Affairs Russia comment this situation refuse to remind the killer to pursue the goal of robbing the wolves take the portfolio money leave the place expensive phone cash money the killer is hiding car VAZ forget the place medical mask IC consider contract murder priority version death head of the Ministry of Internal Affairs RID "","url":" https://life.ru/991216 ","siteType":"LIFE"}	29 March 2017 (implementation term is 88 days)	0.016	0.016	0.046
2.	{"id":"3845f74e-c144-4ec3-9b8f-333e8e08b8ad","date":1490776169000,"title":"Tajikistan becomes the main foreign supplier of suicide bombers for ISIL ","content":"conclusion come author study war by suicide statistical analysis industry martyrdom Islamic state yoke publish international center fight terrorism Hague Netherlands period December year November year only suicide bomber yoke to control to load explosives Inghimashi machine fighter belt suicide bomber fight conventional weapons need to be blown up nearby enemy prima life live bomb house indicate foreign fighter mark author research general difficulty foreigner die quality suicide bomber to consider fifteen year mention Kuni accept Islamic tradition nickname associated place of origin prima life Al Muhajir similarly Al Ansari indicate foreigner indicate country of origin stay die quality drive car explosives originate country Tajikistan then go native Saudi Arabia Morocco Tunisia Russia further give the table indicate the exact figure suicide bomber yoke Tajikistan Saudi Arabia Morocco Tunisia Russia strange year numerous to immigrate the Salafi Tunisia to be a large foreign legion yoke to number about a thousand fighter go close thousand a native of the Wahhabi Kingdom of As Saud native to found follow the immigrant Jordan to rule the royal dynasty belong to the Hashemite clan to originate great-grandfather Prophet Muhammad it is possible therefore the list of the suicide bomber indicate the period only and Jordan Moroccan twelve month to go talk significantly Tajik perish Syria Iraq stroke attack to load explosives Ingimasi machine native foreign country celebrate representative International Center fighting terrorism number amazing consider soul population quantity of natives various country number of yoke prima life assume Tajik frequently direct to suicidal explosion minimum partly nationality Organization to prohibit Russia Supreme Court of the Russian Federation","url":" https://life.ru/991022 ","siteType":"LIFE"}	29 March 2017 (implementation term is 88 days)	0.021	0.021	0.083

Table 2. *Cont.*

No.	Normalized News Text	Date of Event	Value of Parameter ε	Value of Parameter ξ	Initial State of System x_0 31 December 2016
3.	{"id":"5fbf3918-22cc-4ef3-8ad0-20ae2654286c","date":1491441192000,"title":"In the area of the attack on the employees of the Russian Guard in Astrakhan a firefight is going on ","content":"inform life source law enforcement agency Leninsky district Astrakhan to start a firefight crime figure presumably a few hours earlier to attack a Rosguard officer preliminary data special operation pass the area railway station Astrakhan specify the source remind today night three Rosguards get a gunshot wound attack several criminal declare the regional directorate of the ID of RF attack fighter Rosguard involved crime figure April kill police officer Astrakhan","url":" https://life.ru/994664 ","siteType":"LIFE"}	6 April 2017 (implementation term is 96 days)	0.016	0.016	0.047
4.	{"id":"c7584973-348d-417a-90c3-2199a4040558","date":1491047117000,"title":"NATO Does Not Intend to Fight with Russia for Abkhazia and South Ossetia","content": "representative NATO South Caucasus William Lahue declare treaty organization fight Russia Abkhazia South Ossetia case joining Georgia North Atlantic Alliance Georgia must decide status territory clearly understand so far stay Russian army the fifth article Georgia use nobody want war Lahue report member alliance agree Georgia member NATO none term possible joining Georgia alliance call report Interfax slowly matter go forward future Georgia receive invitation know Lahue speech joining Georgia NATO depend parallel factor politics various country willingness Georgia","url":" http://www.vesti.ru/doc.html?id=2872818 ","siteType":"VESTI"}	1 April 2017 (implementation term is 91 days)	0.011	0.011	0.036
5.	{"id":"dacb1299-f6fa-4b25-a4cd-95795657cf4c","date":1490474466000,"title":"Syrian military liberated 195 settlements from IS * since January ","content":"number of settlement liberate January Syrian government army terrorist organization Islamic State yoke January reach report Saturday Russian center reconciliation feuding party Syria number of settlement liberate January year Syrian government troops armed formation international terrorist organization Islamic State increase be said bulletin publish web-site Ministry of Defense of the Russian Federation 24 h control government troops cross a square kilometer territory total difficulty liberate a square kilometer number of settlement join reconciliation process 24 h change message center reconciliation continue negotiations accession regime cessation of hostilities detachment armed opposition Aleppo province Damascus Ham Homs El Quneitr number of armed groups declare a cessation of hostilities compliance agreement armistice change terrorist organization forbid Russia","url":" https://ria.ru/syria/20170325/1490808936.html ","siteType":"RIA"}	25 March 2017 (implementation term is 84 days)	0.016	0.016	0.060

The old woman took a winglet, scraped the box, swept the cornbin and scraped about two handfuls of flour. She kneaded the flour with some sour cream, concocted the Bun, fried in some butter and put it on the windowsill to chill.

The Bun lied for a moment and suddenly rolled—from windowsill to bench, from the bench to the floor, on the floor to the door, jumped through the threshold and rolled to mudroom, from the mudroom to the yard, from the yard to the gates, further and further away.

When running along the road, the Bun met a hare:
Little Bun, little Bun, I want to eat you!, says the hare.
Don't eat me, I'll sing a song for you:
I am a Roly-Poly Bun, Roly-Poly Bun,
I have been scraped on a box,
I have been swept on a cornbin,
I have been kneaded on sour cream,
And yarned on some butter,
And chilled on a windowsill.
I've run away from Grandfather,

I've run away from Grandmother.
And from you hare I can run away all the more!
And he has run along the road—and the hare has lost sight of him!"

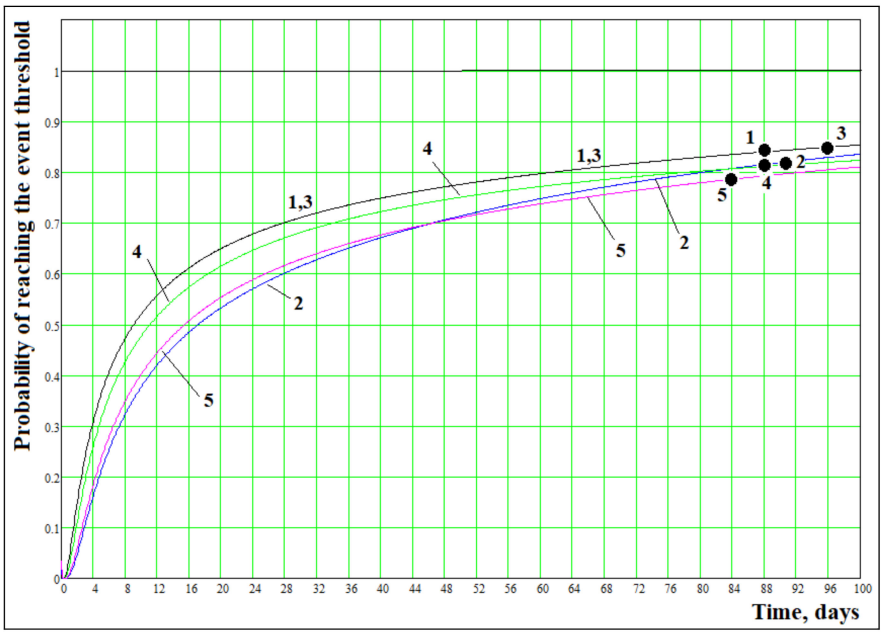

Figure 2. Results of modelling the event threshold crossing for five news described in Table 2 (l = 0.5) for a simple diffusion model.

Further, using the algorithm described in "Determination of the parameters of the event prediction model based on changes in the structure of clusters in the information space of news feeds" and the previously created text clusters of 2016 (W = 300) for this predicted event, we determine the values of the model parameters ξ, ε and x_0 (for finding ξ and ε, τ = 1 day was used), see Table 3.

Table 3. Normalized news about the Roly-Poly Bun and the parameters of the model for predicting it.

Normalized Text of News	Date of Event	Value of Parameter ε	Value of Parameter ξ	Initial State of System x_0 31 December 2016
{"id":"85e74845-70da-434c-a602-497efa002de6","date":1514753700000,"title":"Roly-Poly Bun","content":"grandmother of the gate speak a handful of two door grandfather the road to live to knead fry roll a winglet swept kneaded towards the window song all the more go to roll the floor half put a chimney sweeper sweep a threshold jump yarned scraped on to concoct sing chill chilled eat through take a distant yard porch bench butter scrape window lie scrape mudroom sour cream old man take a distant yard porch bench butter scrape up a window lie scrape mudroom sour cream old man hare box flour cornbin leave hare box flour cornbin leave old woman old woman old woman old woman Bun Bun Bun Bun Bun Bun bun","url":"http://null.ru/null","siteType":"Fictitious"}	implementation time is not known	0.0022	0.0022	0.0076

Using the results of modeling the implementation of real events in the news feed using a simple diffusion model, for an acceptable probability of the occurrence of events, we can take a value equal to 0.8 (in fact, this value is a calibration for the value of the probability of an event implementation, at which the occurrence of the event should already be considered). This allows us to estimate the time of realization of a given event (for a given probability). Modeling the dynamics of the probability of realization of news about the Roly-Poly Bun over time for the developed model gives an estimate of the time of its realization (with a probability value of 0.8 about 90,000 days ≈ 240 years), which is unlikely for the implementation of a news feed event. Thus, the example with fictitious news also shows that the developed model for predicting events in the news feed is adequate and consistent (all the news events used to test the model, depending on what memory depth is taken into account, may be realized at high probability values or if they are fictitious, then they can only be realized in an unacceptably long time).

4.2.4. Assessment of the Accuracy and Reliability of Forecasts of the Implementation of Events in the News Feed, Obtained on the Basis of the Developed Model of the Dynamics of the News Feeds Content

The main problem is the impossibility of carrying out multiple tests of the occurrence of an event realization in the news feed. For experimental verification of the forecast, there is only one observable realization of an event with a known time that has already occurred.

When predicting the values of physically measured values, the forecast accuracy is higher if the error value is lower. Error is the difference between the predicted and actual values of the value studied. In case of forecasting news events occurrence, we work with the dependence of the probability that the described event can occur on time. At each moment of time there is a value of the probability of the occurrence (or non-occurrence) of the predicted event. The experimentally observed physical quantity is only the time when the given event occurs. There is no parameter of measuring the event value (monetary units, kilograms, meters, etc.).

The theoretical distribution function obtained during the development of the model makes it possible to estimate the values of the probability of the event occurrence ($P_{p.}$), which correspond to a given time. Determining the accuracy and reliability of forecasting an event in a news feed based on a single implementation observed is an ambiguous task, un the sense that an event can occur even at a very small value of probability and may not yet occur at a probability close to the figure of one, but there is no possibility of conducting a series of tests. To assess the accuracy and reliability of the proposed forecasting methodology, an evaluative analysis and comparison of the probabilities of the predicted ($P_{p.}$) and random events (P_r) occurrence may be conducted.

Determination of the time dependence of the probability of the predicted event realization ($P_{p.}$) has already been described earlier.

Let us consider the methodology for determining the time dependence of the probability of occurrence of a random event (P_r). To determine P_p, the vector representation of the predicted event was used to determine P_r. It is also necessary to specify a certain vector, relative to which the change in the daytime centroids of the existing clusters will occur. Let us take as a basis that any event that occurs during the year can be largely random (their sum or superposition will also be random), then we can use the vector of the annual centroid of all the events of the year contained in the text corpus as the vector, with respect to which the cosine metric will be calculated and the model parameters ξ, ε, and x_0 will be determined.

The vector of the annual centroid itself is not a random event, but the averaged parameters of the change in the cosine metric of all information clusters can be taken as random variables. The annual centroid is the "average value of the news of the year" relative to which it is possible to determine the changes that in this case will characterize the "randomness" of the news. In this case, the values of the parameters ξ, ε, and x_0 averaged over all clusters found can be considered as quantities describing a random event. Their determination can be made using the previously described algorithm, but as the base

vector, it is necessary to not use the predicted news vector, but the vector of the annual centroid. Moreover, it is necessary to average the parameters ξ, ε, and x_0 over all news clusters of the text corpus.

When using this algorithm to predict random news on the basis of the existing corpus of news texts, the following values of the model parameters were obtained: ξ = 0.008, ε = 0.008, and x_0 = 0.12.

Taking into account that on the day of implementation, the probability of a predicted event is 1, then the forecast accuracy can be estimated by calculating the value of the relative error η% (using the values of the probabilities of realization at a given time point of the predicted (P_p) and random events (P_r): $\% = \frac{1-P_p}{1-P_r} \cdot 100\%$.

Accordingly, the relative accuracy will be Y% = (1 − η) × 100%. To assess the reliability, you can use the deviation value (square root of variance) from the average value of accuracy. The calculations of these parameters are shown in Table 4.

Table 4. Values of accuracy and reliability of their determination for news from Table 2.

News Number	Accuracy Y %	Deviation Square dσ^2 %
1.	79.5	16.0
2.	74.0	2.3
3.	79.2	13.7
4.	73.0	6.3
5.	72.0	12.3
Average value	$\overline{\%}$ = 75.5	$\overline{\sigma\%}$ = ±3.2

Given the properties and characteristics of the source data (poorly structured text information) that can be used to predict news events, the average accuracy of the realized forecasts of about 75% is a fairly good value (see Table 4). Firstly, attention should be paid to the very nature of the data used and their properties, which significantly affects the accuracy and reliability of the forecasts. When creating a model for predicting news events, a mathematical apparatus is needed that would allow formalizing the nature of the data and bringing them to a single scale of measurement. Obviously, it is impossible to perform computational operations in one model, for example, on linguistic estimates and metric scale values, without using mapping procedures to a formal dimensionless set. To solve this problem, we can use the methods of computational linguistics, which allow us to formalize the description of real-world processes using texts in natural languages to create their information images suitable for mathematical processing. It should be noted that already at this stage there is a question about the accuracy and reliability of the transformation of the object into a formal image, the question that some of the information was not lost and how exactly the image corresponds to the object. The representation of text documents in vector form, the elements of which are the values obtained from the use of TF-IDF representation of their semantic units (especially for short texts, such as news) does not always have a high accuracy of correspondence. From two identical sets of words, you can construct sentences that are completely different in meaning, and at the same time, you can get the same semantic constructions from different sets. In addition, there will always be some inaccuracy in the clustering of texts by semantic groups, caused both by errors in the vector representation of texts and by the properties of the clustering methods themselves.

Secondly, there is some error in the model itself. When discussing the value of the accuracy of the realized forecasts, it should be taken into account that we are dealing with a total error that consists of an error in the vector representation of documents, their clustering, and model errors. Because of additivity law, from the data presented in Table 4, we cannot estimate the error of the diffusion model itself, but we get an estimate as a whole for the forecasting method developed by us.

The data obtained allow us to make the assumption that the developed method may be used for forecasting, and the relative forecasting accuracy may be higher than 70%.

5. Discussion

One of the problems of predicting news feed events is the development of models and methods that allow working with a weakly structured information space of text documents.

Given that news events are generated by the action of the human factor when developing such mathematical models, on the one hand, it is necessary to take into account the uncertainty of the impact on the running processes, which creates stochasticity, and on the other hand, it also creates opportunities for self-organization in this system.

The purpose of the research described in this article was to develop a model for predicting news events based on the description of their dynamics and the possibility of using weakly structured text data. To achieve this goal, the following tasks were solved. First, we put together a collection of news text messages over a long period of time. Then we used computational linguistics methods to process them (lemmatization, vectorization based on a dictionary of terms, and the creation of a TF-IDF matrix: term-document, clustering by thematic groups with event dating by time). Next, we examined the change in the structure of news topic clusters over time. This allowed us to describe the changes taking place in the news space in the form of a time series. Then we developed and experimentally tested the model of forecasting news events presented in this article based on the description of their dynamics and the possibility of using weakly structured text data. In addition, we evaluated the accuracy and reliability of forecasts for the implementation of events in the news feed, obtained on the basis of the developed model.

The results of modeling the time dependence of the probability of events realization (with experimentally determined sets of parameter values of the developed model for already realized events) show that the model is consistent and adequate (all news events used to test the model are realized at high probability values (about 80%), or, if they are fictional, they can be realized only for an unacceptably long time).

The limited applicability of our model is the need to use large sets of textual news data collected over a long period of time and the use of a significant number of various tools of computer linguistics.

It should be pointed out that the accuracy and reliability of forecasting is largely defined by the size of the text sample used and the accuracy of their vectorization and clustering by semantic groups. It is necessary to select with high accuracy from the entire set of text messages of news feeds (hundreds of thousands and millions), exactly those that relate to this topic (clustering of events by semantic groups). Ensuring high accuracy of clustering ensures that when forming a time series, a significant part of the information is not lost, for example, on the frequencies of occurrence of this event, which will allow for a more accurate determination of the parameters of the time series under consideration, and will not affect the forecast of its evolution. However, the realization of clustering of a large set of text documents requires significant computing resources and time in technological terms.

In the future, our study will be aimed at developing faster clustering algorithms, but which could provide the necessary accuracy without losing the required quality. We also plan to study the influence of the presence of memory about previous states of the system and the processes of self-organization on the probability of the predicted events. Due to the influence of the human factor, they cannot only have a stochastic character, but also show the ability to self-organize, and in addition have a memory of previous states.

6. Conclusions

Forecasting of news feed events is carried out based on their textual description, vectorization, and finding the value of the cosine of the angle between this vector and the centroids of various semantic clusters of the information space. The change in the value of this cosine over time can be examined as the wandering of a point on the segment $[0, 1]$, which contains a trap at the threshold point of the event implementation, where the wandering point can go over time.

The probability schemes of transitions between states in the information space were considered during the creation of the model. The second-order nonlinear differential equation was derived based on this approach and a boundary value problem for predicting news events was formulated and solved, which allowed us to obtain a theoretical time dependence of the probability density function for the distribution of parameters of non-stationary time series describing the evolution of the information space.

The results of our research, which were described in the article, allow us to draw a number of very important conclusions.

The developed model for forecasting events in the news feed is adequate and consistent (all news events used to test the model occurred at a high probability (about 80%) or, if they are fictitious news, they can only occur in an unacceptably long time).

Analysis of the model for forecasting events in the news feed based on a simple diffusion model confirms the possibility of predicting news feed events based on their text description, vectorization, and finding the cosine value of the angle between this vector and the centroids of various information clusters. The change in this cosine over time can be considered as the point wandering on the segment [0, 1] that includes in l a trap where the wandering point can eventually fall. The result of simulating the time dependence of the event occurrence probability with experimentally determined sets of parameter values of the developed model are not inconsistent in terms of probability behavior (among other things, at large times, the probabilities asymptotically tend to the figure of one).

Estimates of the accuracy and reliability of news forecasting allow us to suggest that the developed model can be used for forecasting, and the relative accuracy of forecasting can be higher than 70%.

Author Contributions: D.Z.: conceptualization, formal analysis, writing-review & editing; E.A.: methodology, visualization; O.T.: data curation, writing-original draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ao, X.; Luo, P.; Li, C.R.; Zhuang, F.; He, Q. Discovering and learning sensational episodes of news events. *Inf. Syst.* **2018**, *78*, 68–80. [CrossRef]
2. Huminski, A.; Bin, N.Y. Automatic extraction of causal chains from text. *Libres* **2020**, *29*, 99–108.
3. Preethi, P.G.; Uma, V.; Kumar, A. Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction. In *Procedia Computer Science*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 48, pp. 84–89.
4. Gerber, M.S. Predicting crime using Twitter and kernel density estimation. In *Decision Support Systems*; Elsevier: Amsterdam, The Netherlands, 2014; Volume 61, pp. 115–125.
5. Huang, C.-J.; Liao, J.-J.; Yang, D.-X.; Chang, T.-Y.; Luo, Y.-C. Realization of a news dissemination agent based on weighted association rules and text mining techniques. *Expert Syst. Appl.* **2010**, *37*, 6409–6413. [CrossRef]
6. Bollen, J.; Huina, M.; Zeng, X.-J. Twitter mood predicts the stock market. *J. Comput. Sci.* **2010**, *2*. [CrossRef]
7. Novikova, O.A.; Andrianova, E.G. Rol metodov intellektual'nogo analiza teksta v avtomatizacii prognozirovaniya rynka cennyh bumag. (Role of the methods of intellectual analysis of text in automation of security market forecast). *Cloud Sci.* **2018**, *5*, 196–211.
8. Gruhl, D.; Guha, R.; Kumar, R.; Novak, J.; Tomkins, A. The predictive power of online chatter. In *KDD '05: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*; ACM Press: New York, NY, USA, 2005; pp. 78–87.
9. Mishne, G.; Rijke, M.D. Capturing global mood levels using blog posts. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*; Nicolov, N., Salvetti, F., Liberman, M., Martin, J.H., Eds.; The AAAI Press: Menlo Park, CA, USA; Stanford, CA, USA, 2006; pp. 145–152.

10. Liu, Y.; Huang, X.; An, A.; Yu, X. ARSA: A sentiment-aware model for predicting sales performance using blogs. In *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*; ACM: New York, NY, USA, 2007; pp. 607–614.
11. Choi, H.; Varian, H. Predicting the Present with Google Trends. *Tech. Rep.* **2009**. [CrossRef]
12. Zhao, L.; Sun, Q.; Ye, J.; Chen, F. Multi-task learning for spatio-temporal event forecasting. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 8–9 August 2015; pp. 1503–1512.
13. Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.-H.; Liu, B. Predicting flu trends using twitter data. In Proceedings of the IEEE Conference on Computer Communications Workshops, Shanghai, China, 10–15 April 2011; pp. 702–707.
14. O'Connor, B.; Balasubramanyan, R.; Routledge, B.R.; Smith, N.A. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the Fourth International Conference on Weblogs and Social Media, Washington, DC, USA, 23–26 May 2010; pp. 122–129.
15. Tumasjan, A.; Sprenger, T.; Sandner, P.; Welpe, I. Predicting elections with twitter: What 140 characters' reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC, USA, 23–26 May 2010; pp. 178–185.
16. Ramakrishnan, N.; Butler, P.; Muthiah, S. "Beating the News" with EMBERS: Forecasting Civil Unrest Using Open Source Indicators. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; KDD '14, New York, NY, USA, 24–27 August 2014; pp. 1799–1808. [CrossRef]
17. Ning, Y.; Muthiah, S.; Rangwala, H.; Ramakrishnan, N. Modeling Precursors for Event Forecasting via Nested Multi-Instance Learning. *Soc. Inf. Netw.* **2016**, 1095–1104. [CrossRef]
18. Chouhan, S.S.; Khatri, R. Data Mining based Technique for Natural Event Prediction and Disaster Management. *Int. J. Comput. Appl. Found. Comput. Sci.* **2016**, *139*, 34–39.
19. Orlov, Y.N.; Shagov, D.O. Indicative statistics for non-stationary time series. *Keldysh Inst. Prepr.* **2011**, *53*, 1–20. (In Russian)
20. Kryzhanovsky, A.D.; Pastushkov, A.A. Nonparametric method of reconstructing probability density according to the observations of a random variable. *Russ. Technol. J.* **2018**, *6*, 31–38. (In Russian)
21. Gnedenko, B.V. *Probability Theory Course*; Fizmatlit: Moscow, Russia, 1961; 406p.
22. Fuentes, M. Non-Linear Diffusion and Power Law Properties of Heterogeneous Systems: Application to Financial Time Series. *Entropy* **2018**, *20*, 649. [CrossRef] [PubMed]
23. Orlov, Y.N.; Fyodorov, S.L. Generation of unsteady trajectories of time series based on Fokker—Plank equation. *Pap. MFTI* **2016**, *8*, 126–133.
24. Radinsky, K.; Horvitz, E. Mining the Web to Predict Future Events. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining, Rome, Italy, 4–8 February 2013; ACM: New York, NY, USA, 2013; pp. 255–264. [CrossRef]
25. Gunawardana, A.; Meek, C.; Xu, P. A Model for Temporal Dependencies in Event Streams. In Proceedings of the Advances in neural information processing systems, Granada, Spain, 12–15 December 2011; Volume 4, pp. 1962–1970.
26. Christopher, D. Manning, Prabhakar Raghavan, Hinrich Schütze. In *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008; 544p.
27. Tan, P.-N.; Steinbach, M.; Vipin, K. *Introduction to Data Mining*; Pearson Addison-Wesley: Boston, MA, USA, 2006; 169p.
28. Andrews, N.O.; Fox, E.A. *Recent Developments in Document Clustering*; Department of Computer Science, Virginia Tech: Blacksburg, VA, USA, 2007; 25p.
29. Feldman, R.; Sanger, J. *The Text Mining Handbook*; Cambridge University Press: Cambridge, UK, 2009; 410p.
30. Lesko, S.A.; Zhukov, D.O. Trends, self-similarity, and forecasting of news events in the information domain, its structure and director. In Proceedings of the 2015 International Conference on Big Data Intelligence and Computing, Chengdu, China, 19–21 December 2015; pp. 870–873.
31. Sigov, A.S.; Zhukov, D.O.; Khvatova, T.Y.; Andrianova, E.G. A Model of Forecasting of Information Events on the Basis of the Solution of a Boundary Value Problem for Systems with Memory and Self-Organization. *J. Commun. Technol. Electron.* **2018**, *63*, 1478–1485. [CrossRef]
32. Lesko, S.A.; Zhukov, D.O. Stochastic self-organisation of poorly structured data and memory realisation in an information domain when designing news events forecasting models. In Proceedings of the 2016 IEEE 2nd International Conference on Big Data Intelligence and Computing, DataCom 2016, Auckland, New Zealand, 8–12 August 2016; pp. 890–893.
33. Zhukov, D.; Khvatova, T.; Otradnov, K. Forecasting News Events Using the Theory of Self-similarity by Analysing the Spectra of Information Processes Derived from the Vector Representation of Text Documents. *Commun. Comput. Inf. Sci.* **2020**, *1140*, 54–69.
34. Sigov, A.; Zhukov, D.; Novikova, O. Modelling of memory realization processes and the implementation of information self-organization in forecasting the new' s events using arrays of natural language texts. In Proceedings of the 1st International Scientific Conference Convergent Cognitive Information Technologies, Convergent 2016, Moscow, Russia, 25 November 2016; CEUR Workshop Proceedings. Volume 1763, pp. 42–55.
35. Zhukov, D.O.; Zamyshlyaev, A.M.; Novikova, O.A. Model of Forecasting the Social News Events on the Basis of Stochastic Dynamics Methods. In Proceedings of the ITM Web of Conferences, Moscow, Russia, 14–15 February 2017.