Ukrainian Catholic University

Master Thesis

Detecting patterns of coordinated news article dissemination

Author:
Petro Bodnar

Supervisor:
Dmytro Karamshuk

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2021

Declaration of Authorship

I, Petro BODNAR, declare that this thesis titled, “Detecting patterns of coordinated news article dissemination” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Detecting patterns of coordinated news article dissemination

by Petro BODNAR

Abstract

This study aims at devising methods for detecting coordination among content spreaders at scale. We focus on methods which uncover the latent structure of the content dissemination networks from the time-series of publications. We identify the advantages of generative models, especially self-exiting stochastic processes, for modeling information cascades and detecting structural patterns in groups of events. We validate the most popular of these models – the Multivariate Hawkes processes – on a large dataset of news websites and achieve an improvement in comparison to simpler baselines, e.g., cosine similarity between time-series of publications.

Acknowledgements

First, I want to thank my supervisor Dmytro Karamshuk for his invaluable advice, patience, and ability to inspire. Also, I want to thank my former college Nadiia Romanenko for her mentoring and thoughtful comments. I am grateful to the team of texty.org.ua for the possibility of using their data.

Also, I am thankful to Oleksii Molchanovskyi and Ukrainian Catholic University for the outstanding experience of studying at the Data Science Master's program.

Many thanks to my groupmates for the two years we spent learning both in classes and from each other. Finally, I am grateful to my friends and family for their unconditional support and understanding.

Contents

List of Figures

List of Tables

List of Abbreviations

MHP	**M**ultivariate **H**awkes **P**rocesses
LRHP	**L**ow **R**ank **H**awkes **P**rocesses
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency
EM	**E**xpectation **M**aximization
NPHC	**N**on **P**arametric **H**awkes **C**umulant
RNN	**R**ecurrent **N**eural **N**etwork
NMF	**N**on-negative **M**atrix **F**actorization
ROC	**R**eceiver **O**perating **C**haracteristic
AUC	**A**rea **U**nder **C**urve

List of Symbols

ϕ	memory kernel for Hawkes process
α	parameter of exponential kernel, impact of history on Hawkes process
μ	background intensity of Hawkes process

Chapter 1

Introduction

1.1 Motivation

Big social media platforms have to moderate content daily. This requires analysis of unprecedented volumes of information. As a result, Facebook plans to increase the role of the automatic pre-moderation and rating of questionable content to prioritize moderation of most damaging posts, Vincent, 2020: "Now, Facebook says it wants to make sure the most important posts are seen first and is using machine learning to help. In the future, an amalgam of various machine learning algorithms will be used to sort this queue, prioritizing posts based on three criteria: their virality, their severity, and the likelihood they are breaking the rules".

This study will consider two of the components mentioned above: virality and the likelihood that the content will break the rules. Modeling information cascades is a standard method to estimate content virality according to Zhou et al., 2020. The Facebook research team has extensively surveyed the methods for identifying doubtful content based on its diffusion topology and temporal features in Halevy et al., 2020. Our goal is to fuse these different methodologies for predicting the coordinated spread of misinformation.

After an initial exploration of a large corpus of news articles, we observed the trace of coordination between information spreaders. As has been discussed in Pacheco et al., 2020, this can often be a sign of misinformation or other sorts of unreliable content. However, we also faced a challenge that, for many datasets like ours, a network structure is not readily available or not fully reliable. Therefore, the methods which require knowledge about the network topology of information diffusion are unpractical. We aim to obtain a model able to work only with temporal data without knowing the latent structure of the spreading network. During the work on this master's thesis, we will research the tools and methods for detecting coordinated actions of information spread in various environments.

1.2 Goals of the master thesis

The current thesis aims to assess the feasibility of inferring coordination between news publishers based purely on the time series of publications they produce without knowing the actual content. To end this, we aim to:

- Check the feasibility of uncovering network structure based on timestamps of publications. Consider and develop a procedure for estimating true network structure from the content of news articles. Evaluate how well does prediction based on temporal data corresponds to the network structure suggested by content.

- Validate the proposed methods on a new dataset of news articles from Ukrainian and Russian websites collected in 2020 by media studying misinformation in Ukraine.
- Develop evaluation method that would enable to compare similarity matrices suggested by the use of textual and temporal information. Find a simpler baseline approach that could be compared with the selected methods. Compare metrics for the selected method and baseline approach.
- Interpret the model outputs and suggest the direction for further research and improvement of the method.

1.3 Structure of the thesis

Chapter 2 reviews the prior work on detection coordination and misinformation. In Chapter 3 we present the problem formulation, outline the validation method and methodology used for experiments. In Chapter 4, we describe the dataset focusing on textual and temporal signals. Chapter 5 presents the results of the experiments. Finally, in Chapter 6 we sum up all the results and contributions of the current research.

Chapter 2

Background and Related Work

2.1 Research on detection of misinformation

The first aspect of this research deals with the literature about harmful content and misinformation. A survey by Facebook researchers Halevy et al., 2020 summarizes efforts by Facebook to preserve the integrity of the social network and remove harmful content. It also touches the topic of misinformation as one of the types of this content. Researches point out the different direction to take to detect misinformation based on varied data available:

- based on text understanding with deep neural networks (Devlin et al., 2018), using big pretrained model and fine-tuning on task-specific data. Big models may use language modeling objective as (Peters et al., 2018, Radford and Sutskever, 2018) or language masking objective as (Devlin et al., 2019).
- as a subset of text-based approaches, there is research on the text's tone as a source of signal for detecting misinformation.
- coordinated action uncovering described by Pacheco et al., 2020 in the framework for misinformation detection. Another work (Farajtabar et al., 2017) researches use of point processes and Reinforcement learning for the mitigation of "fake" news.
- network properties are actively used for misinformation detection (Wu and Liu, 2018) as well as information cascades and other types of content (mainly image and video)[1];

In the Pacheco et al., 2020 authors described several cases of coordination discovered using their framework. The method is based on finding similar traces of activity in social media that speak about the lack of independence between accounts involved. Coordination between accounts spreading false information is an important hypothesis for our research. If it exists, we would be able to systematically detect it from information appearing both in social networks and other types of networks that involve dissemination of content (for example, news websites posting similar articles). One of the promising directions for the study of coordination is the use of information cascades.

Informational cascades are well-studied tools for describing social phenomenons that are spreading over time on networks. We will define them as follows: "The trajectories and structures of information diffusion, as well as the adopters/participants in information spreading "Zhou et al., 2020. This definition is broad enough to describe content popularity on social media and other types of information spread on

[1] https://ai.facebook.com/blog/community-standards-report/

networks. Prediction of information cascades is a way to estimate how influential individual messages are and how many people they would affect.

According to Zhou et al., 2020 prediction of information, cascades refers to different actions. For example, they predict the popularity of content in social media (likes, shared, hashtag use), number of citations for scientific papers, number of comments under articles, number of similar news articles, and rating for movies. As there are different prediction tasks, we will use the taxonomy from Fig. 2.1, describing the most significant prediction models used to fulfill them.

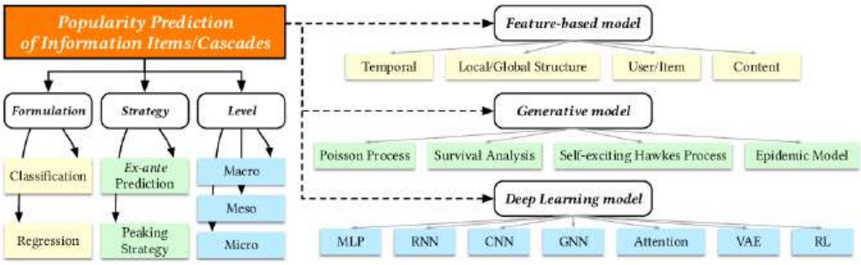


FIGURE 2.1: Cascade prediction taxonomy Zhou et al., 2020

The taxonomy proposed by a survey mentioned previously breaks the problem into three dimensions. First, consider what type of problem we are solving regression - finding the size and numeric properties of a cascade or classification - finding if given content would become popular or the property reach certain threshold. Also, the figure describes method families that could be used to model cascades. These are feature-based models (using varied properties on cascades with classic machine learning models), Generative models (focusing on most important temporal features), and deep learning models (trying to use graph embedding to avoid feature engineering and combine network topology with temporal features and content).

It would be useful to consider if cascades are actually predictable. In Cheng et al., 2014 authors discovered that temporal features and structure are key predictors for cascade growth. Another study Martin et al., 2016 provided interesting results about the same issue. It suggested that there is a theoretical bound for the predictability of cascades. Work by Shulman, Sharma, and Cosley, 2016 investigates peeking strategy prediction as a classification task (looking for those items that would become popular).

As mentioned earlier, there are three groups of methods for modeling cascades. Those include:

1. traditional machine learning models trained on hand-crafted features
2. generative models such as epidemic models, point processes, and survival analysis
3. deep learning, RNN for modeling sequence of events, and graph embedding to capture the cascade structure.

In this research, we decided to focus on the coordination aspect of misinformation and discover methods to spot it effectively. We work not with a single cascade or multiple independent cascades but rather several interacting cascades in real life. Therefore, it is important to find a method that enables tracking how cascades affect each other and the spread of content. The strong interaction between cascades can be a sign of coordination of event sources.

We do not discuss the use of images and other media for detecting misinformation and utilization of network structure of information cascades appearing with the spread of content because this information is not available in our dataset. In contrast, we have plenty of textual and temporal information that has to be considered.

2.2 Point Processes and Time Series in misinformation detection

Features-based methods and Deep Learning-based methods leverage the network topology and information about spreaders and content in addition to temporal information. However, as we mentioned previously, temporal information is already essential in prediction, and also it is easily obtained and requires less prepossessing than other features of the cascade. Therefore, from the point of view of scalability and limited access to data, it would be fruitful to focus on methods that leverage information about the occurrence of events in cascade.

Generative models are common tools for popularity prediction and, more generally, modeling events occurrence in time. We are interested in a specific subset of these models self-exciting the Hawkes process, and more specifically, it is Multivariate form. There is research on how Farajtabar et al., 2017 Multivariate Hawkes Processes could have been used for mitigation of fake news. We plan to develop a similar approach that would show coordination among several nodes of a diffusion network with an unknown structure.

As it was stated in Farajtabar et al., 2017 and Dutta et al., 2020 point process, especially Hawkes process is commonly used in social media and other network environments to predict the popularity of the content and its share speed. As we mentioned earlier, tackling misinformation requires both the ability to estimate the probability for the message to spread and the probability that the message contains misinformation. Hawkes process allows estimating how much certain messages could spread in the network and see the coordination pattern between the actors spreading information. We believe that this task can be formulation through the lens of information cascade modeling.

There is also research (Haimovich et al., 2020) about the scalability of Hawkes processes for information cascade prediction and its ability to predict the size of a cascade over the arbitrary time horizon. This is another useful property of Hawkes process for the task of misinformation detection as it allows to understand the dynamics of cascade growth.

By using the Multivariate Hawkes Process (MHP) that extends the abilities of the single thread Hawkes process, we can reason about the causal connection between different types of events as it was described in Achab et al., 2017. Standard MHP is limited by the amount of data used for prediction. However, there are interesting works in extending and scaling it by using approximation techniques. For example, Nickel and Le, 2020 used partial information about the content of propagating events (state policies, for example) to speed up MHP. They demonstrated that the method could be used to study policy diffusion in the United States. Similarly, the interaction between "memes" in online publication and news was studied in (Leskovec and Krevl, 2014). The Fig. 2.2 shows how different states influence each other based on the policy adoption history.

Separate group of methods that can work with temporal data is time series methods. In Yang and Leskovec, 2011 authors research patterns of the sharing of content in social media and identify typical "life" of different pieces of content. For example,

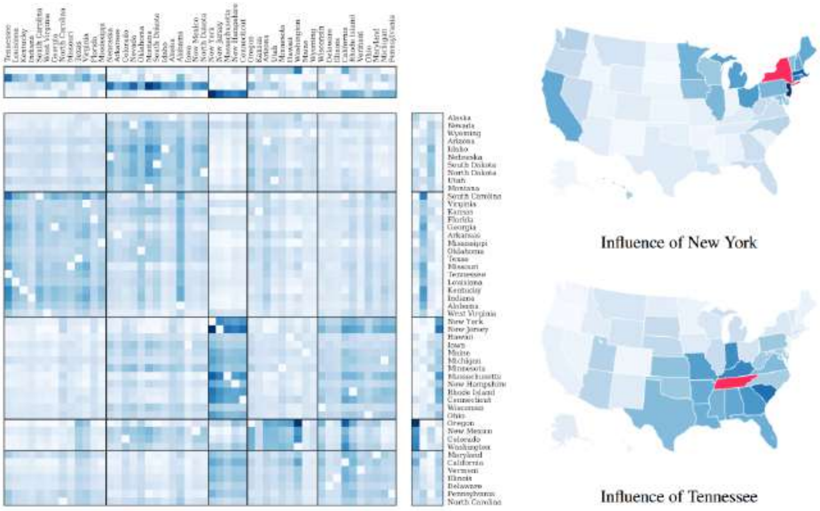


FIGURE 2.2: Policy diffusion from Nickel and Le, 2020

they found that news gets a lot of attention in a short time, but soon their sharing diminishes. They used a novel K-Means-inspired method. Also, there is some research on the use of anomaly detection in the time series. Shipmon et al., 2017 describe several unsupervised methods for anomaly detection from temporal data.

Another concern for working with social media and the popularity of content is privacy. In Halevy et al., 2020 report, authors write that maintaining privacy while fighting misinformation is a rising concern. Messaging apps are getting encrypted, and messages' content is often not available even for companies operating those systems. However, the temporal data is still available and is not considered private. We can process temporal data and get insights about the popularity of a certain piece of content and how it is affected by other messages. It is possible to do this in a scalable unsupervised manner.

2.3 NLP for misinformation detection

The most common approach for content moderation and misinformation detection, according to Halevy et al., 2020 is based on text analysis. The recent breakthrough in natural language modeling with big language models as Radford and Sutskever, 2018, Devlin et al., 2019 and Lample and Conneau, 2019 allow creating target-specific classifiers on the top of pretrained models. There is also some successful work with the detecting of hate speech in multilanguage setting[2].

However, text is still very nuanced data, and we need to consider different additional signals to understand that certain text is misinformation or distributed in a coordinated manner. Some work has been done in tone detection. Authors in Zhang et al., 2021 review how fake messages arouse "social emotions" and how those can be used for fake detection.

Also, there is research on multitask model use for misinformation detection Rajamanickam et al., 2020 while training model on both abuse and emotion detection. There is also research on specific "Hyperpartisan" language in Potthast et al., 2017, that gives some clue for limiting the scope of messages eligible for misinformation

[2] https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/

detection. Another aspect of this topic is the phenomenon of "clickbait" catchy headlines with often misleading content.

To sum up, the advantages of text-based misinformation detection are the high accuracy, interpretability, and ability to update the model with new content and share information between languages. However, the disadvantages are the need for lots of labeled data; models still are not language-agnostic. Also, people can adapt their language to make it less understandable by the model (it is easier than changing their own temporal posting patterns, for example).

Chapter 3

Methodology

3.1 Problem formulation

As we described previously, one of the important features of misinformation is coordination between content disseminators. To understand that coordination occurs, we need to inspect the message and find features shared by different actors. Many researchers working with misinformation concentrate their efforts on message: narratives used, tone, citation, and other features.

However, there is also a temporal dimension of this problem. Parties simultaneously spreading the same message are apparently connected. If we know that a certain message includes misinformation, we can trace its path through the latent network of content dissemination. Therefore, temporal features of messages may be an important asset in tracing misinformation. Moreover, by concentrating on the time of publication rather than the text, we avoid many privacy concerns and legal issues. Being able to detect coordination with only temporal information could also be further scaled to work with social media content and other domains that could be represented as a set of events on a timescale.

In this thesis, we overview the task of finding coordination in news article publishing by different web sources, using a temporal signal only. We formulate this task as an unsupervised task that looks for the similarity between sources and validates this similarity by the internals of publications. We use a simple "naive approach" based on the comparison of temporal vectors of different websites and a more advanced method based on the Hawkes process for modeling.

3.2 Validation

We will evaluate our method by comparing model results with our prior knowledge about news websites' closeness. We calculated the initial resemblance approximation by comparing headlines of those news articles (using cosine similarity of vectorized text). The strength of connection in our diffusion network is defined as the percentage of similar content in websites.

We would consider the model successful if it would detect the connection between news websites closely connected according to the study of their content. The model may also notice coordination between other websites whose relationship is not straightforward from the content itself.

As a further extension of the research, we would investigate other closeness metrics based on the similarity of content between various news providers from our dataset. We plan to calculate the similarity scores for the text of news articles. However, as the dataset is extensive, we would consider only the most performant approaches.

Metrics As a metric for evaluation of the connection between content and temporal dimension, we use three correlation coefficients: Spearman, Pearson, Kendall Tau. Rank correlation coefficients (Spearman and Kendall Tau) are standard tools to deal with the ranking of events. They are commonly used in varied settings; for example, in this study Amornbunchornvej et al., 2018 they were used for detecting events indicating coordination in groups of animals. Spearman and Kendall Tau allow detecting the nonlinear relationship between data. We also use plain Pearson correlation to find out if there is a noticeable linear relation between the model output and approximated content similarity.

The validation procedure includes the following steps:

- calculates initial resemblance approximation by comparing headlines of news articles (using cosine similarity of vectorized text) - results in a square matrix, each value represents the strength of the connection between two websites
- calculate the coordination between websites based on temporal data with the selected method - results in a square matrix, each value represents the strength of the connection between two websites
- flatten both matrices obtained
- calculate correlation scores between two vectors of pairwise connection scores (temporal and based on the text)

Some details about the correlation scores used are following. Pearson correlation can be calculated by this formula (for sample):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3.1}$$

where **n** is a size of sample, **x** and **y** are samples compared and $\overline{x}, \overline{y}$ are means of samples. However, as it is a measure of linear relation, it is not directly applicable for measuring ranks. Therefore, we use two other correlation coefficients that can describe the nonlinear nature of data.

Spearman correlation is formulated (for integer ranks) as following from Hald, 1952:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3.2}$$

where **n** is number of ranks and **d** is distance between two ranks. Lastly Kendall tau Kendall, 1945 is the following:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{3.3}$$

3.3 Multivariate Hawkes Processes

Hawkes process (Hawkes, 1971) is a point process that describes sequance of events where each new event increases the probability of getting another event. Intensity of this sequance, or the rate at which new event occur in the infinitelly small timeframes could be described by the following formula:

$$\lambda(t|H) = \lambda_0(t) + \sum_{i:t>T_i} \phi(t - T_i) \tag{3.4}$$

where T_i is a list of times when events have occurred up the specific point t. We are calculating the intensity of the process at point t, based on the previous history. H - is a history of previous events, ϕ is the memory kernel function. The kernel function is commonly exponential because it is scalable and it allows calculating many properties of the process (number of the future event during a certain time or till the end of time horizon) analytically, described in Rizoiu et al., 2017.

However, this is only one dimensional Hawkes process that considers a single sequence of events. It could be extended to consider multiple sequences of events of different type, as described in Yang et al., 2017. Multivariate Hawkes processes (MHP) allows to measure the effect of both history and neighbor events on each other. For p-dimentional MHP the formula on i-th event take a form:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} \int_0^t f_{i,j}(t-\tau) dN_j \tag{3.5}$$

here μ_i represent base intensity of the i-th dimention (by time scale), $N_j(\tau)$ is a count of events from each source within [0, t], and $f_{i,j}(t)$ is a kernel function that approximates the relationship in data. This function could be exponential as previously, but also can take any other form (other distribution, neural networks etc...). We are going to stick with exponential function that is most common and has many useful properties described earlier. Its formula is:

$$f_{i,j}(t) = \alpha_{i,j} exp\{-\beta_{i,j} t\} \mathbb{1}\{t > 0\} \tag{3.6}$$

there are three parameters that has to be optimized for MHP μ, α, β. In case of MHP α represents an element of adjacency matrix of all event sources.

In the paper Nickel and Le, 2020, researchers developed a method to scale MHP to a large number of events and topics. This is especially important as most real-life applications include a huge number of actors that create sparse networks of events. MHP allows detecting coordination between different types of events and approximates the latent structure of diffusion network that may not be available in many cases. Afterward, we can use the Hawkes process to predict the future intensity of the information cascade.

We plan to use the multivariate Hawkes process to model relationships between actors (news providers) in our data and detect the diffusion network's structure. The model would be trained on event sequences from the news publications on different websites from texty.org.ua[1] database. After the model training, we will study similarity scores between different news websites. A similar approach was used in Nickel and Le, 2020.

We plan to use the same technique with our dataset. However, in contrast with the policy research, we have a baseline approximation of the diffusion network based on a comparison of the news article content. Another difference is that in Nickel and Le, 2020 authors use some signal from the policy document's content (the fact that certain "meme" occurred in this policy). We do not want to use the content of the messages due to privacy concerns and the quick growth of the amount of data considered. In contrast, we rely solely on temporal data.

Our main goal is to quantify how well does MHP allow estimating network structure suggested by the content. As a secondary, goal we are willing to check whether MHP reveals connections that are not apparent from news articles' content.

[1] https://topic-radar.texty.org//

To calculate the intensity and use it for prediction or clustering event sources, we need to approximate the parameters of MHP. There are a variety of methods to tackle this problem. In the following section, we are going to give an overview of some those methods.

3.4 Scaling MHP

There is a closed form of log likelihood for MHP it is give in the following formula:

$$l(\Theta) = \sum_{n=1}^{n} log(\lambda(t_i)) - \int_0^T \lambda(t)dt \tag{3.7}$$

here λ is an intensity of Hawkes process and t, is a point on a timescale. Theoretically log likelihood could be used for direct optimization for parameter estimation. But this method have not prooved to be effective because of the curvature of optimization space near local optimum as stated by Veen and Schoenberg, 2008. Therefore, some regularization is required for effective parameters estimation.

The most basic method to estimate parameters for MHP is Expectation Maximization (EM) developed by Ozaki, 1979. There are also a variant of this approach with Bayesian formulation proposed by Morse, 2017 and ML approach by Veen and Schoenberg, 2008. As described in Lewis and Mohler, 2011 EM algorithm for estimating MHP parameters runs as follows:

- Start with the guess for parameters Θ^0
- **Expectation step**: Calculate expected value of p_{ij} (probability that event j triggers event i)
- **Maximization step**: Calculate new approximation of parameters

Formulas for calculating each of steps are following:
Expectation:

$$p_{ij}^k = \frac{\alpha^k \omega^k e^{-\omega^k(t_i - t_j)}}{\mu^k + \sum_{j=1}^{i-1} \alpha^k \omega^k e^{-\omega^k(t_i - t_j)}} \tag{3.8}$$

$$p_{ii}^k = \frac{\mu^k}{\mu^k + \sum_{j=1}^{i-1} \alpha^k \omega^k e^{-\omega^k(t_i - t_j)}} \tag{3.9}$$

Maximization:

$$\mu^{k+1} = \frac{\sum_{i=1}^{n} p_{ii}^k}{T} \tag{3.10}$$

$$\alpha^{k+1} = \frac{\sum_{i>j}^{n} p_{ii}^k}{n} \tag{3.11}$$

$$\omega^{k+1} = \frac{\sum_{i>j}^{n} p_{ii}^k}{\sum_{i>j}(t_i - t_j) p_{ij}^k} \tag{3.12}$$

This method provided a quality approximation of parameters of MHP in a reasonable time, but with the growth of data, it can get too slow. Another method that provides better speed of approximation was developed in Achab et al., 2017. The method called Non Parametric Hawkes Cumulant (NPHC) tries to estimate the matrix of α directly instead of estimating the exponential kernel of MHP.

$$\alpha^{ij} = \int_0^{+\infty} \phi^{ij}(u)du \geq 0 for 1 \leq i,j \leq d \tag{3.13}$$

The NPHC method relies on matching cumulants. It contains following steps:

- Compute estimation $\widehat{\mathbf{M}}$ of some moments $\mathrm{M}(\mathbf{G})$ that are defined by $\mathbf{G}$;
- Look for $\widehat{\mathbf{G}}$ that minimalized L^2 error in $||M(\widehat{\mathbf{G}}) - \widehat{\mathbf{M}}||^2$

This method is robust to kernel shape, but the resulting matrix is symmetrical; therefore it uncovers only the mutual influence between to event sources, not the direction of influence.

Final approach we are using is Low Rank Hawkes Process (LRHP) from Türkmen, Çapan, and Cemgil, 2020. It is similar to NPHC and also results in symmetric matrix. Methods works through direct optimization of parameter matrix using momentum to find approximation of matrix of α. It it based on two assumptions. First, that there is unique matrix Φ satisfying:

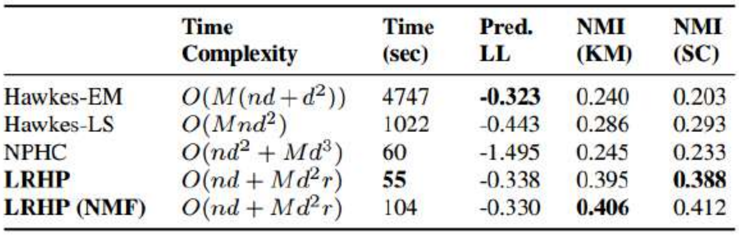
$$C = (I - \Phi)^{-1} D (I - \Phi^T)^{-1} \tag{3.14}$$

C, D are Hawkes cumulants calculated from data and $\Phi = \Phi^T$. Second assumption is Φ can be factorized with non-negative low rank factorization. This method is influenced by NMF matrix factorization tools that is widely used for clustering and community detection. One last point is that authors do not estimate Φ directly, but rather use truncated Neumann series approximation truncated Neumann series approximation $\widetilde{\Phi}^{(p)} = I + \Phi + \Phi^2 + ... + \Phi^p = I + (W^T W) + (W^T W)^2 + ... + (W^T W)^p$.

The method is combined of two steps:

- calcuting estimations of cumulants **C** and **D** from data
- fitting **W** via gradient descent to $L(W) = ||C - \Psi(W) D \Psi(W)||$

As we can see from Fig. 3.1 LRHP model provides good performance and much better speed and complexity compared to similar models. In our experiments, we used the implementation of the method provided by authors[2]. We modified the code for logging purposes and for the optimization of parameters.



	Time Complexity	**Time (sec)**	**Pred. LL**	**NMI (KM)**	**NMI (SC)**
Hawkes-EM	$O(M(nd+d^2))$	4747	**-0.323**	0.240	0.203
Hawkes-LS	$O(Mnd^2)$	1022	-0.443	0.286	0.293
NPHC	$O(nd^2 + Md^3)$	60	-1.495	0.245	0.233
LRHP	$O(nd + Md^2 r)$	**55**	-0.338	0.395	**0.388**
LRHP (NMF)	$O(nd + Md^2 r)$	104	-0.330	**0.406**	0.412

FIGURE 3.1: LRHP performance compared to similar models from Türkmen, Çapan, and Cemgil, 2020

This model was used by Lemonnier, Scaman, and Kalogeratos, 2016 for different tasks of cascade prediction in simulated and real-life datasets and demonstrated that the model is highly scalable but still capable of producing state-of-the-art results.

[2] https://github.com/canerturkmen/lowrankhawkes

Chapter 4

Dataset

We obtained the dataset from texty.org.ua[1] project. The data includes a period from January 6 to September 30, 2020, with 3.5 million news articles in Russian. Each article has the following properties headline, publication time, website name, website type based on emotional manipulation score (according to texty.org.ua internal model), and website country of origin (by domain). Those articles have been collected from 920 news websites. Main properties of data in Table. 4.1.

Number of events	Number of websites	Start date	End date	Language
920	3 523 095	2020-01-06	2020-09-30	Russian

TABLE 4.1: News dataset properties

The dataset was created by scraping RSS feeds of the selected websites. The list of websites to scrape was manually created by media experts, it included both big respectable media and websites which were proven to share misinformation. It also included plenty of small and medium-sized websites without substantial reputation, both positive or negative. Therefore, even though the dataset is mainly targeted at websites sharing disinformation, it also represents the overall media landscape.

We have used the news dataset to extract headline pairs with similar content. Those pairs were calculated in 30 hour time frames to save computation time, starting at 00:00 each day and ending at 06:00 on the next day. We used the pre-trained navec[2] model to create embeddings for headings and then calculated cosine similarity between all the headlines in the timeframe. We used a 0.9 threshold to filter the most similar headlines (re-posting the same article in practical terms).

Afterward, we have conducted exploratory data analysis to understand the data structure and transformed data into a graph. It showed that a significant fraction of news re-posts belongs to a small set of websites. We also find out that most of the headlines do not spread much, and only 1% of articles have eight re-posts or more. This can be observed in Fig. 4.1.

4.1 Textual data

To solve the problem, we were looking for a way to define and measure the similarity of news articles as a sign of coordination. We decided to use TF-IDF for the vectorization of textual data because of the speed and simplicity of the approach. To get a vector representing the entire content of websites, we concatenated all the textual information from each website into separate "corpora" for each website.

[1] https://topic-radar.texty.org/

[2] https://github.com/natasha/navec

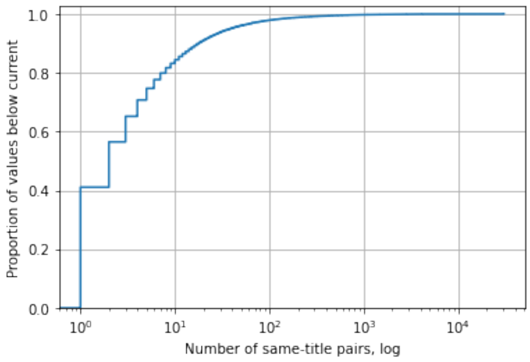


FIGURE 4.1: Empirical cumulative distributions of the number of news pairs in the dataset

TF-IDF is one of the simplest and most common tools for text vectoring available. It was presented in Ramos, 1999 and the author described its purpose as following: "TF-IDF works by determining the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire document corpus. Intuitively, this calculation determines how relevant a given word is in a particular document". The formula to calculate TF-IDF is:

$$w_d = f_{w,d} * log(|D| f_{w,D}) \tag{4.1}$$

where w is an individual word, D - document collection, and d - individual document. f_w, d equals the number of times w appears in d, |D| is the size of the corpus, and $f_{w,D}$, equals the number of documents in which w appears in D. This procedure allows to transform a collection of texts into vectors of numbers representing each text. However, vectorization is not enough to understand the relations between websites. Therefore, we applied cosine similarity (Manning, Raghavan, and Schütze, 2008) a simple formula that allows getting the relative closeness between different vectors. The pairwise similarity is calculated using sklearn[3] library with the following formula:

$$k(x, y) = \frac{xy^T}{||x||||y||} \tag{4.2}$$

where $k(x, y)$ is a measure of similarity between two vectors, x,y are vectors we compare.

We applied TF-IDF on each "corpora" and used cosine distance to measure similarity between website "corpora." The resulting textual similarity matrix represented the similarity of content between different websites. It can be viewed as an adjacency matrix approximating the latent structure of the dissemination network.

However, there is a scalability problem involved. Due to lengthy texts of news articles, the corpora for a single website can get big quickly. Working with a huge amount of data is both dependent on computational power, storage, and time-consuming. Therefore, we decided to check if titles could serve as an effective proxy for the content of news articles.

To prove that title is a reliable proxy for the content of an article, we decided to do the following experiment. We sampled random 50000 articles from our database. We concatenated all the titles for every website in the sample and calculated the textual similarity matrix. Repeated same procedure with article texts. Finally, flattened both

[3] https://scikit-learn.org/stable/

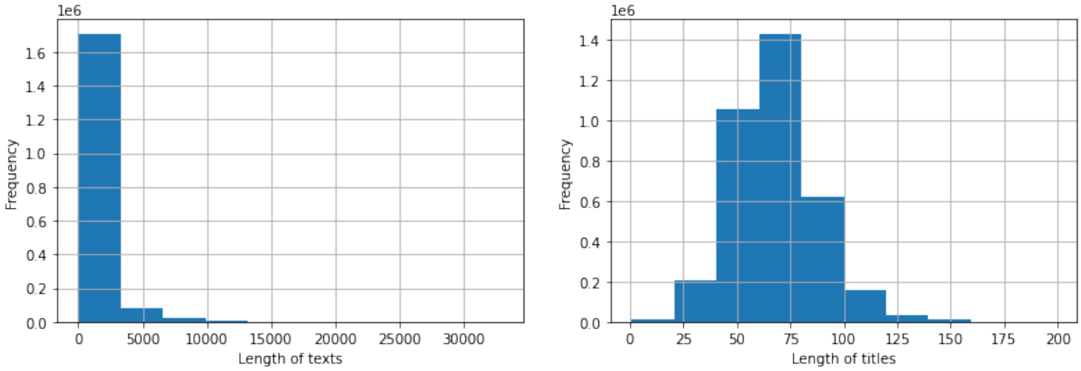


FIGURE 4.2: Distribution of title length and article length in the dataset.

matrices and calculated the correlation score. Results are: **Spearman correlation** - 0.804, **Pearson correlation** - 0.762, **Kendall Tau correlation** - 0.610.

The correlation between matrices is substantial; moreover, if we will check the scatter plot on Fig. 4.3 we will see that correlation is even more significant for larger values of "closeness." Therefore, if there is a substantial connection between the website's content, it is noticeable both from the titles and content of articles. However, working with titles is much quicker and requires much less memory.

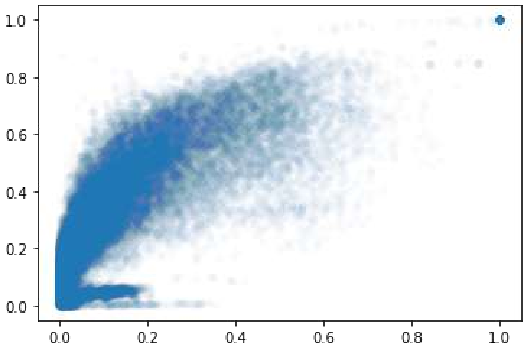


FIGURE 4.3: Correlation between connection score of websites based on text and title

4.2 Temporal data

Our temporal data consists of the time points of each publication. We used simple preprocessing to transform the date in string format into the numeric form suitable for further use of the model. We normalized the dataset with the earliest date, rounded values and converted them to integers.

We also have done some initial visual exploration of data. For example, Fig. 4.4 displays a total number of publications from the dataset per day. It can be seen that there was an increase in the daily publication in March 2020. Most of the time, websites included in the dataset produced less than 100 publications per day. While sometimes this number could increase to 500 publications, Fig. 4.5.

Also, we checked if the publications are evenly distributed along the timeline for all the websites (Fig. 4.4). There was a suspicious increase in the total number of

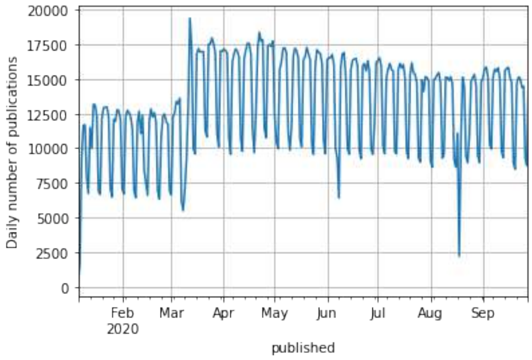


FIGURE 4.4: Daily number of publications in dataset

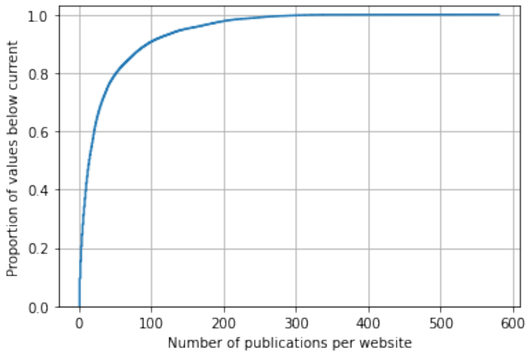


FIGURE 4.5: Distribution of number of publication per day for each website

publications in March 2020. First, we interpreted it as an increase in news coverage due to the Covid19 epidemic. However, after some further research, we found out that the increase has different nature. The dataset underwent an update in March 2020, and several websites were added to it. It can be easily observer from Fig. 4.6. We tried to estimate the Hawkes process with a full dataset, but due to the nature of this model, it resulted in worse performance because some websites "started posting" only around the middle of the dataset. Therefore, we removed websites that were added later, proceeding further with 260 websites.

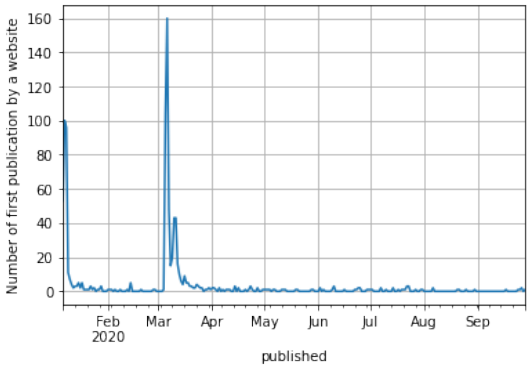


FIGURE 4.6: Number of websites that started publishing by per each week in the dataset.

Additionally, we checked what is the distribution of publication in the week or single day. The results are in Fig. 5.2. It can be viewed that the majority of news publications happen on workdays and work hours. Also, the number of publications is more or less stable both during the day and week. The reason may be that news websites always have a planned amount of content, but they react by creating additional coverage if anything unexpected happened. This makes it harder for a model to detect coordination because of the regularities in data. However, we are confident that MHP would be able to detect coordination in this data despite those complications.

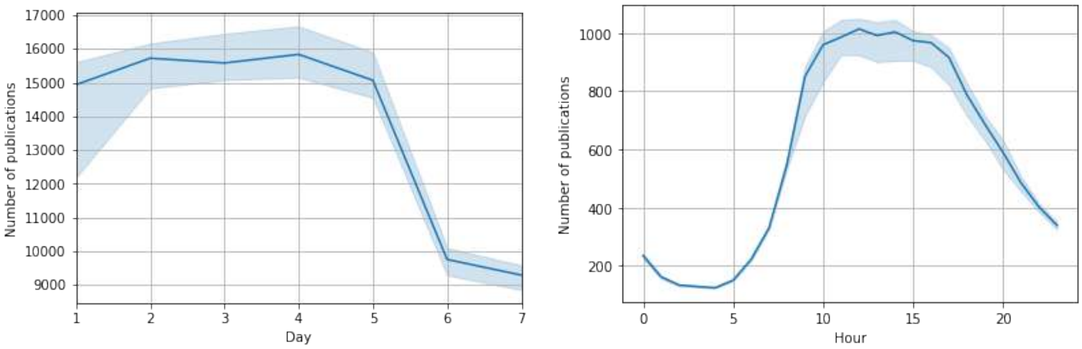


FIGURE 4.7: Median an 95% confidence interval for number of publication per hour and week in dataset.

Chapter 5

Experiments

We performed our experiments on the LRHM model described in Section 3.4. The baseline "naive" approach is to use the cosine similarity of publication vectors. We compare it with our model output. Also, we compared the speed and accuracy of our model with two other approaches for estimating Multivariate Hawkes Processes and the estimated number of data needed for training. Finally, we've done some work for interpreting the model output and understanding its results.

5.1 Accuracy

We used three approaches described in Section 3 for estimating parameters of Multivariate Hawkes Process: Expectation Maximization (EM), Non Parametric Hawkes Cumulant (NPHC), and Low-Rank Hawkes Process (LRHP). We compared them with a content similarity matrix for 192 websites that were categorized by Texty. The result could be found in Table 5.1.

	Spearman correlation	**Pearon correlation**	**Kendall Tau correlation**
LRHP	0.184	0.163	0.129
EM	0.168	-0.03	0.117
NPHC	0.2	0.066	0.147

TABLE 5.1: Correlation for all models on 192 websites

However, due to big disparity in the number of publication for different websites (min - 1, 25% - 30, 50% - 423, 75% - 3651) we decided to run a same experiment on comparable websites. We took a subsample of the top 20 websites with a similar number of publications and ran scores one more time. Results in Table 5.2.

	Spearman correlation	**Pearon correlation**	**Kendall Tau correlation**
LRHP	0.111	0.168	0.071
EM	0.262	**0.358**	0.179
NPHC	0.242	0.015	0.149

TABLE 5.2: Correlation for all models on top 20 websites

Result of the experiment show that with the smaller sample of comparable websites, EM method resulted in better performance. For a bigger set of data (especially a bigger number of event sources), EM gets worse performance than LRHP and NPHC. Also, the time of convergence grew quickly with the increase in the number of data sources. Therefore, we decided to evaluate scalability of all three models with the increasing number of data points.

5.2 Scalability

The amount of data in the dataset researched and other potential applications is very big. Models that outperform their counterparts in speed have a big competitive advantage for real-life applications. To research the potential for scaling of different models, we formulated the following experiment. We train our models on incrementally bigger subsets of data, measuring run time for each model. The results are displayed in the Fig.5.3.

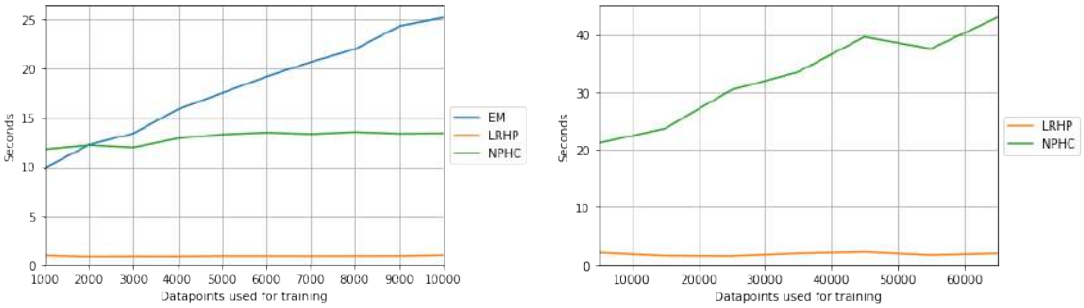


FIGURE 5.1: Scalability experiment, with three models (left chart) and with two quicker models (right). Low Rank Hawkes Process - orange, Non Parametric Hawkes Cumulant - green and Expectation Maximization methods - blue.

LRHP method is the quickest and requires only a single path over data; its complexity is $O(d^2r + nd)$ according to Türkmen, Çapan, and Cemgil, 2020. EM method is much slower, and it is getting slower with each additional subset of data. NPHC initially demonstrated stability with the increasing data, but after processing 10 000 events, it also slowed. We will further research the LRHP method for estimation Multivariate Hawkes Process because of its reasonable accuracy and ability to sustain high data load.

5.3 Hyper-parameters tuning

We used Bayesian optimization framework[1] to find optimal parameters of the LRHP model that would allow us to achieve better performance than our baseline. We optimized internal parameters of the model (integral support, non-negativity, approximation degree, and rank) as well as parameters of the training process (learning rate, improvement rate, and a number of epochs). We used Spearman correlation as a score guiding optimization process.

We ran the optimization process for 25 iterations and used a hold-out dataset to test the performance of the optimized model. The hold-out dataset constituted of $\frac{1}{3}$ of data spit by the time dimension. Two thirds were used to optimize parameters and one third for evaluation. For this task, we used 260 websites from the entire dataset, excluding those that were added to the dataset in March. Finally, we calculated the correlation between estimated coordination and title similarity of websites. The results are in the Table 5.3. LRHP managed to reach better performance compared to the baseline "naive" approach.

[1] https://github.com/fmfn/BayesianOptimization

	Spearman correlation	Pearon correlation	Kendall Tau correlation
LRHP	**0.728**	0.666	**0.531**
Baseline	0.678	**0.655**	0.487

TABLE 5.3: Correlation optimized model compared to baseline model

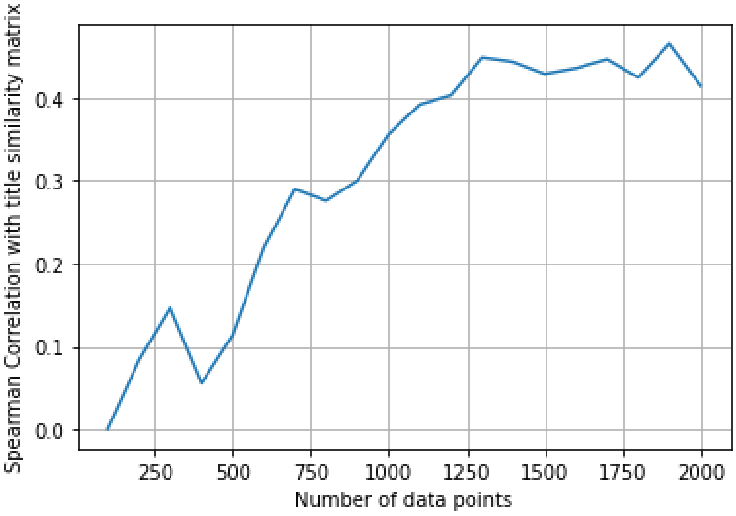


FIGURE 5.2: Minimal amount of data needed to train LRHP effectively

5.4 Data utilization

After choosing the method, we are interested in finding out what amount of data is required for LRHP to learn effectively and uncover hidden relationships in data. For this aim, we designed the following experiment. We trained LRHP on the increasing subset of events starting from 100 up to 2100 events. At each stage, we estimated parameters of MHP, calculated title similarity matrix for websites involved, and calculated correlation Spearman correlation score. As you can see from Fig.5.2 the model gets limited insights from data when the number of events is less than 1000. However, after 1000 events, the accuracy it starts increase.

5.5 Interpretation of the model

Because of the nature of MHP, the scores resulting from training contain information both about the process intensity and closeness of the connection between event sources. Therefore, we decided to do some further research and binarize the outcomes of the news title comparison described earlier.

We used a 0.5 threshold, treating websites with titles similarity scores more than 0.5 as connected and those less than 0.5 as not connected. Fig. 5.3 shows a Receiver operating characteristic (ROC) curves for LRHP model with optimal parameters and for baseline approach. The curves are similar, but LRHP is more prone to True Positive outcomes. Also, the Area Under the ROC Curve (AOC) score is bigger for LRHP - 0.812 compared to 0.769 for baseline.

ROC curve also allows us to define the best threshold for a binary classification using the output of our unsupervised model. We found the best threshold by subtracting the vector of true positive rates and false positive rates. The threshold that

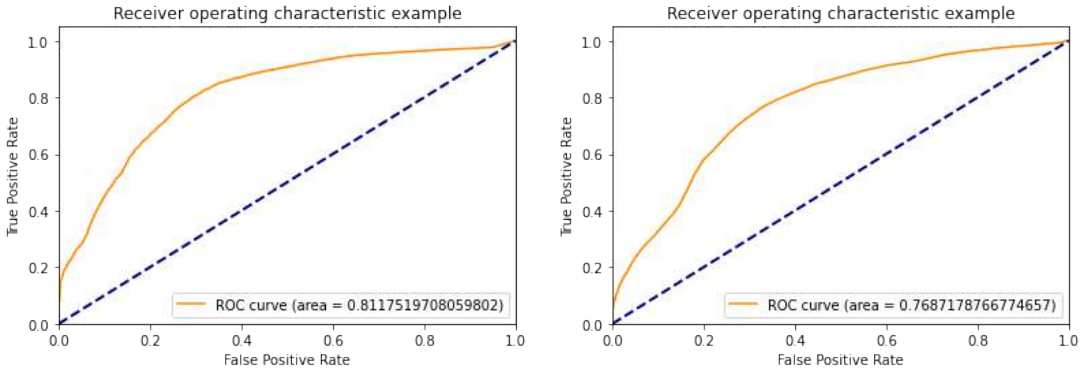


FIGURE 5.3: ROC curve for LRHP output (left) and baseline approach (right).

corresponded to the biggest rate difference is the best choice for classification. For our given model with the optimal parameters, the best threshold is 0.005229.

We used this threshold to binarize the LRHP cooperation matrix. We considered pairs with the value of temporal interaction higher than 0.005229 as "cooperating" and those lower as not cooperating. After the binarization, we were able to calculate common classification metrics presented in Fig. 5.4. Those metrics suggest that our approach can be used for classification and gives better results than baseline.

TABLE 5.4: Classification results table

	value	precision	recall	f1-score
Low Rank Hawkes Process	No cooperation	**0.81**	**0.71**	**0.76**
	Cooperation	**0.69**	**0.80**	**0.74**
Baseline	No cooperation	0.78	0.67	0.72
	Cooperation	0.65	0.77	0.71

Finally, we also researched false positive results of the LRHP model output binarization. We found that 8.98% of connected belong to this category. In fact, this may not be a mistake of the model as it considers different information than text-based similarity score. The model may have detected new signal pointing out on cooperation. This could be one of the directions of further research.

Chapter 6

Conclusions

6.1 Summary of contribution

During the course of this research, we have worked on the task of unsupervised detection of coordination based on temporal data.

- We applied Multivariate Hawkes Process to a dataset of news publication which was not previously covered in the literature. This method is scalable, privacy-friendly, and effective in spotting coordination in information dissemination.
- The work is based on a new dataset created by media covering the spread of misinformation in Ukraine. We have conducted an exploratory analysis of this dataset, identified potential problems, and outlined major trends.
- We developed a method to approximate content similarity between websites without knowing the underlying network structure.
- We thoughtfully tested our method, compared it to simpler baselines, and provided an interpretation of the connections between websites uncovered by our proposed method.

To summarise, we created an approach for coordination detection in an environment with limited data (except temporal information). This approach is unsupervised, scalable, and effective. We obtained a 5% improvement in comparison to a simpler baseline as measured by Spearman's correlation coefficient. In addition, we extended our model for binary classification of the website pairs; the model got a 3% better F-1 score for detecting cooperation than the baseline. Another important aspect of this work was ensuring scalability of our method, comparing it with Expectation Maximization and Non-Parametric Hawkes Cumulant approach for estimation of Multivariate Hawkes Process parameters. The selected Low-Rank Hawkes Process approach proved to be the most scalable and able to perform on increasingly big datasets.

6.2 Directions for further research

From this point we can see several directions for further research. The first one is devising a more complex estimation of website's cooperation by their content. It can be solved by, for example, the use of topic modeling to find the shared topics between different websites and then calculating website text similarity within each topic. Such a granular approach would allow understanding better when the news producers are cooperating with each other and why.

Another direction is to work with the intensity α obtained from the general model we outlined to predict future bursts of activity among "suspect " websites. This may help to create an "early warning" system for massive misinformation campaigns.

Finally, this method may be extended into other types of media (Telegram social messaging app or social networks, or any other environment with temporal information available). We believe it would be beneficial to assess this method with the data from ordinary users (for example in social media). They do not have such regular posting patterns and the coordination may be even more noticeable. Also, it would be useful to work with data that was not created with an aim of studying spread of misinformation. As it may have an inherited bias in its network structure.

Bibliography

Achab, Massil et al. (2017). “Uncovering Causality from Multivariate Hawkes Integrated Cumulants”. In: Proceedings of Machine Learning Research 70. Ed. by Doina Precup and Yee Whye Teh, pp. 1–10. URL: http://proceedings.mlr.press/v70/achab17a.html.

Amornbunchornvej, Chainarong et al. (2018). “Coordination Event Detection and Initiator Identification in Time Series Data”. In: *ACM Transactions on Knowledge Discovery From Data* 12.5, p. 53.

Cheng, Justin et al. (2014). “Can cascades be predicted”. In: *Proceedings of the 23rd international conference on World wide web*, pp. 925–936.

Devlin, Jacob et al. (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Devlin, Jacob et al. (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Dutta, Hridoy Sankar et al. (2020). “HawkesEye: Detecting Fake Retweeters Using Hawkes Process and Topic Modeling”. In: *IEEE Transactions on Information Forensics and Security* 15, pp. 2667–2678.

Farajtabar, Mehrdad et al. (2017). “Fake News Mitigation via Point Process Based Intervention”. In: *International Conference on Machine Learning*, pp. 1097–1106.

Haimovich, Daniel et al. (2020). “Scalable Prediction of Information Cascades over Arbitrary Time Horizons.” In: *arXiv preprint arXiv:2009.02092*.

Hald, Anders (1952). *Statistical tables and formulas*. eng. Wiley publications in statistics : applied statistics. New York: Wiley. ISBN: 0471340235.

Halevy, Alon Y. et al. (2020). “Preserving Integrity in Online Social Networks.” In: *arXiv preprint arXiv:2009.10311*.

Hawkes, Alan G. (1971). “Spectra of Some Self-Exciting and Mutually Exciting Point Processes”. In: *Biometrika* 58.1, pp. 83–90. ISSN: 00063444. URL: http://www.jstor.org/stable/2334319.

Kendall, M. (1945). “THE TREATMENT OF TIES IN RANKING PROBLEMS”. In: *Biometrika* 33.3, pp. 239–251.

Lample, Guillaume and Alexis Conneau (2019). “Cross-lingual Language Model Pretraining”. In: *CoRR* abs/1901.07291. arXiv: 1901.07291. URL: http://arxiv.org/abs/1901.07291.

Lemonnier, Rémi, Kevin Scaman, and Argyris Kalogeratos (2016). “Multivariate Hawkes Processes for Large-scale Inference”. In: *AAAI*, pp. 2168–2174.

Leskovec, Jure and Andrej Krevl (June 2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data.

Lewis, Erik and George Mohler (2011). “RESEARCH ARTICLE A Nonparametric EM algorithm for Multiscale Hawkes Processes”. In:

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. Web publication at informationretrieval.org. Cambridge University Press.

Martin, Travis et al. (2016). "Exploring Limits to Prediction in Complex Social Systems". In: *WWW '16 Proceedings of the 25th International Conference on World Wide Web*, pp. 683–694.

Morse, Steven T (2017). "Persistent cascades and the structure of influence in a communication network". In:

Nickel, Maximilian and Matthew Le (2020). *Learning Multivariate Hawkes Processes at Scale*. arXiv: 2002.12501 [cs.LG].

Ozaki, T. (Dec. 1979). "Maximum likelihood estimation of Hawkes' self-exciting point processes". In: *Annals of the Institute of Statistical Mathematics* 31.1, pp. 145–155. DOI: 10.1007/bf02480272. URL: https://doi.org/10.1007/bf02480272.

Pacheco, Diogo et al. (2020). *Uncovering Coordinated Networks on Social Media*. arXiv: 2001.05658 [cs.SI].

Peters, Matthew E. et al. (2018). *Deep contextualized word representations*. arXiv: 1802.05365 [cs.CL].

Potthast, Martin et al. (2017). *A Stylometric Inquiry into Hyperpartisan and Fake News*. arXiv: 1702.05638 [cs.CL].

Radford, Alec and Ilya Sutskever (2018). "Improving Language Understanding by Generative Pre-Training". In:

Rajamanickam, Santhosh et al. (2020). "Joint Modelling of Emotion and Abusive Language Detection". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4270–4279.

Ramos, Juan (1999). *Using TF-IDF to Determine Word Relevance in Document Queries*.

Rizoiu, Marian-Andrei et al. (2017). *A Tutorial on Hawkes Processes for Events in Social Media*. arXiv: 1708.06401 [stat.ML].

Shipmon, Dominique T. et al. (2017). "Time Series Anomaly Detection: Detection of Anomalous Drops with Limited Features and Sparse Examples in Noisy Periodic Data". In: *arXiv preprint arXiv:1708.03665*.

Shulman, Benjamin, Amit Sharma, and Dan Cosley (2016). "Predictability of Popularity: Gaps between Prediction and Understanding". In: *Tenth International AAAI Conference on Web and Social Media*, pp. 348–357.

Türkmen, Ali Caner, Gökhan Çapan, and Ali Taylan Cemgil (2020). "Clustering Event Streams With Low Rank Hawkes Processes". In: *IEEE Signal Processing Letters* 27, pp. 1575–1579. DOI: 10.1109/LSP.2020.3019964.

Veen, Alejandro and Frederick Paik Schoenberg (2008). "Estimation of Space–Time Branching Process Models in Seismology Using an EM–Type Algorithm". In: *Journal of the American Statistical Association* 103.482, pp. 614–624.

Vincent, James (2020). *Facebook is now using AI to sort content for quicker moderation*. URL: https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation.

Wu, Liang and Huan Liu (2018). "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 637–645.

Yang, Jaewon and Jure Leskovec (2011). "Patterns of temporal variation in online media". In: *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 177–186.

Yang, Yingxiang et al. (2017). "Online Learning for Multivariate Hawkes Processes". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30.

Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/92a0e7a415d64ebafcb16a8ca817cde4-Paper.pdf.

Zhang, Xueyao et al. (2021). “Mining Dual Emotion for Fake News Detection”. In: *WWW 2021 : The Web Conference*.

Zhou, Fan et al. (2020). *A Survey of Information Cascade Analysis: Models, Predictions and Recent Advances*. arXiv: 2005.11041 [cs.SI].